Mathematik für Ingenieure III

(für Informatiker)

Hans Grabmüller
Institut für Angewandte Mathematik

Vorlesung im Wintersemester 2000/2001

Friedrich–Alexander–Universität Erlangen–Nürnberg

Inhaltsverzeichnis

Kapitel 12

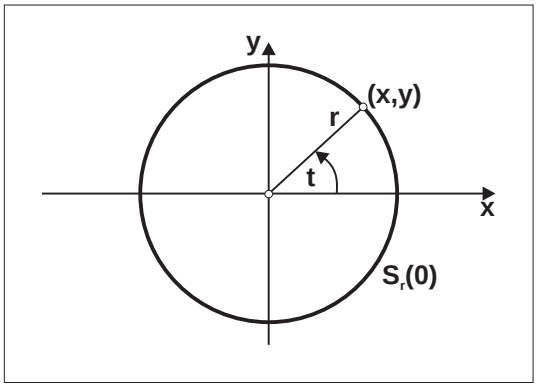
Differentialgeometrie von Kurven $\vec{x} : \mathbf{R} \to \mathbf{K}^n$

12.1 Parameterdarstellung und Parametertransformation

Parameterdarstellungen ebener und räumlicher Kurven hatten wir bereits in Abschnitt 6.5 diskutiert. Wir wollen hier nochmals an diese Diskussion anknüpfen und dazu einige Erweiterungen und Ergänzungen bringen. Wir erinnern an die Tatsache, dass es für *ebene* Kurven verschiedene Formen der Darstellung gibt:

- die explizite Darstellung
- die implizite Darstellung
- die Parameterdarstellung

BSP. (12.1.1) Im Vektorraum $\mathbf{R}^2$ seien kartesische Koordinaten (x, y) eingeführt. Wir betrachten eine Kreislinie $S_r(0)$ vom Radius $r > 0$ mit Mittelpunkt im Koordinatenursprung O.



Kreislinie $S_r(0)$ vom Radius $r > 0$

(I) *Explizite Darstellung:*

$$\boxed{S_r(0) = \{(x, y) \in \mathbf{R}^2 \ : \ y(x) = \pm\sqrt{r^2 - x^2}, \quad x \in [-r, r]\}}$$

(II) *Implizite Darstellung:*

$$\boxed{S_r(0) = \{(x, y) \in \mathbf{R}^2 \ : \ f(x, y) := x^2 + y^2 - r^2 = 0\}}$$

(III) *Parameterdarstellung:*

$$\boxed{S_r(0) = \{(x,y) \in \mathbf{R}^2 \ : \ x = x(t) = r\cos t, \ \ y = y(t) = r\sin t, \quad t \in [0, 2\pi]\}}$$

Die Darstellung (I) hat den Nachteil, dass $S_r(0)$ nicht durch eine einzige Funktionsbeziehung $y = f(x)$ beschrieben werden kann. Als eine der Konsequenzen ergibt sich, dass eine Ableitung $y'(x)$ in den Punkten mit vertikaler Tangente im eigentlichen Sinn nicht existiert. In der Darstellung (II) sind die Variablen x und y gleichberechtigt. Als Nachteil gilt die Tatsache, dass Ableitungen nicht direkt berechnet werden können. Ausgeräumt werden die hier geschilderten Nachteile im allgemeinen durch die Parameterdarstellung (III). Diese ist immer dann von Vorteil, wenn eine geschlossene Kurve vorliegt. Wie wir außerdem in Abschnitt 7.3 gesehen haben, wird der Parameter t häufig als Zeitvariable interpretiert. In diesem Fall entspricht die Darstellung $x = x(t)$, $y = y(t)$ der Beschreibung der (ebenen) Bahn eines Massepunktes in Abhängigkeit von der Zeit. Als weiteren Vorteil der Parameterdarstellung wertet man ihre Übertragbarkeit auf Kurven im Vektorraum $\mathbf{K}^n$.

Wir gehen nachfolgend davon aus, dass der Vektorraum $\mathbf{K}^n$ die Standardbasis $\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n$ und das Standardskalarprodukt $\langle \vec{x}, \vec{y} \rangle$ trägt sowie die durch das Skalarprodukt induzierte Norm $\|\vec{x}\| := \sqrt{\langle \vec{x}, \vec{x} \rangle}$. Es sei ferner $\emptyset \neq I \subset \mathbf{R}$ stets ein Intervall.

Definition 12.1 *Im Vektorraum* $\mathbf{K}^n$ *heiße eine Punktmenge* $\Gamma := \{\vec{x} \in \mathbf{K}^n \ : \ \vec{x} = \vec{x}(t), t \in I\}$ *eine* **Parameterkurve**, *wenn es Funktionen*

$$x_1 = x_1(t),\ x_2 = x_2(t), \ldots, x_n = x_n(t), \quad x_i : I \to \mathbf{K}, \tag{1.1}$$

gibt, so dass gilt:

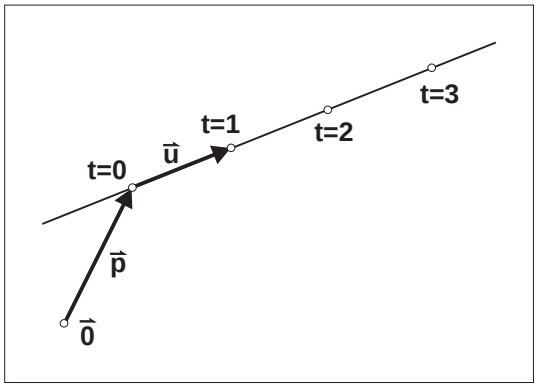
$$\boxed{\Gamma \ni \vec{x}(t) = \Big(x_1(t), x_2(t), \ldots, x_n(t)\Big)^T \ \ \forall\ t \in I.} \tag{1.2}$$

Die Darstellung (1.2) *heiße* **Parameterdarstellung** *von* Γ. *Die Kurve* Γ *heiße* **stetige** (*stetig differenzierbare, ...*) *Parameterkurve, wenn* $\vec{x} \in C(I; \mathbf{K}^n)$ ($\vec{x} \in C^1(I; \mathbf{K}^n)$, ...) *gilt. Ist* $I = [a, b]$ *ein abgeschlossenes Intervall, so heiße* $\vec{x}(a)$ **Anfangspunkt** *und* $\vec{x}(b)$ **Endpunkt** *der Parameterkurve* Γ. *Eine stetige Parameterkurve* Γ *heiße* **geschlossen**, *wenn* $\vec{x}(a) = \vec{x}(b)$ *gilt. Durch die Festlegung von Anfangs- und Endpunkt erhält* Γ *einen Durchlaufungssinn: Parameterkurven mit Anfangs- und Endpunkt heißen* **orientiert**.

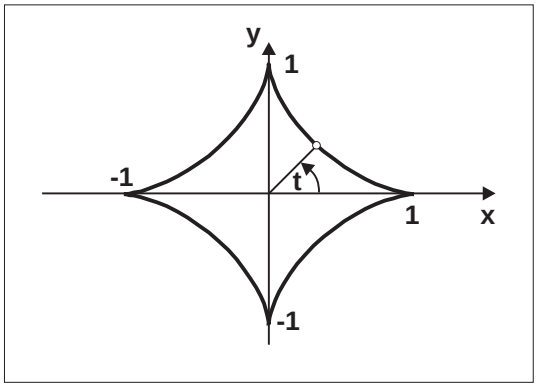
Für ebene Kurven $\Gamma \subset \mathbf{R}^2$ erweist sich häufig die **Polardarstellung**

$$\boxed{\vec{x}(t) = \Big(r(t)\cos t, r(t)\sin t\Big)^T, \quad r = r(t) \geq 0, \quad t \in I,}$$

als vorteilhaft. Wir verweisen auf Abschnitt 6.5, in welchem auch geometrische Beispiele ebener Parameterkurven gebracht wurden (*Kreis, Ellipse, Zykloide, Kardioide, archimedische Spirale*). Nachfolgend bringen wir weitere geometrische Beispiele von Parameterkurven.



Die Gerade



Die Astroide ($b = 1$)

BSP. (12.1.2)

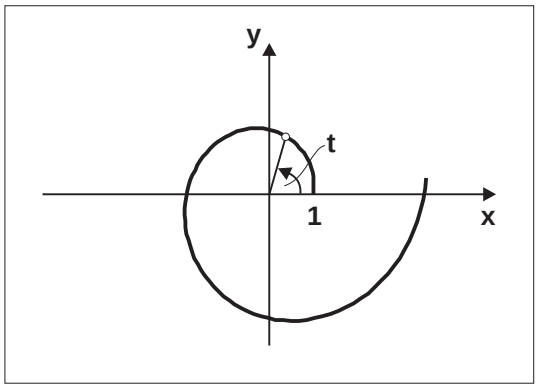
- Die **Gerade**

$$\boxed{\vec{x}(t) = \vec{p} + t\vec{u} \in \mathbf{K}^n, \quad t \in \mathbf{R}, \quad (\vec{u} \neq \vec{0})}$$

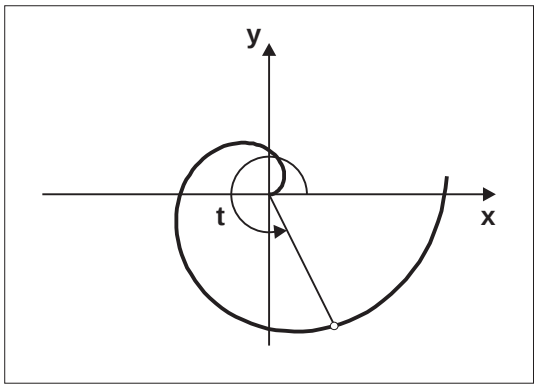
- Die **Astroide**

$$\boxed{\vec{x}(t) = (b\cos^3 t, b\sin^3 t)^T, \quad b > 0, \quad t \in [0, 2\pi]}$$

Die Astroide ist ein Sonderfall der *Hypozykloiden*: Auf der Innenseite eines Kreises vom Radius b rollt ein zweiter Kreis vom Radius $a < b$. Für $m := \frac{b}{a} = 4$ beschreibt ein Punkt auf Peripherie des Innenkreises eine Astroide.



Die logarithmische Spirale



Die archimedische Spirale

- Die **logarithmische Spirale**

$$\boxed{\vec{x}(t) = \Big(r(t)\cos t, r(t)\sin t\Big)^T, \quad r(t) := ae^{bt}, \quad a > 0, \quad t \in [0, 2\pi]}$$

Die Parameterkurve kann auf ganz $\mathbf{R}$ fortgesetzt werden. Der *Sonderfall* $b = 0$ liefert wieder die Kreislinie.

- Die **archimedische Spirale**

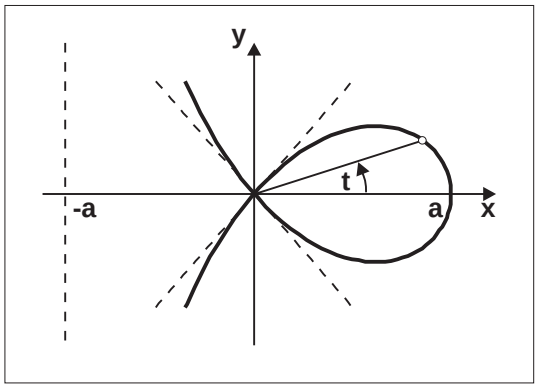
$$\boxed{\vec{x}(t) = \Big(r(t)\cos t, r(t)\sin t\Big)^T, \quad r(t) := at, \quad a > 0, \quad t \in [0, 2\pi]}$$

Die Parameterkurve kann auf das Intervall $[0, +\infty)$ fortgesetzt werden.

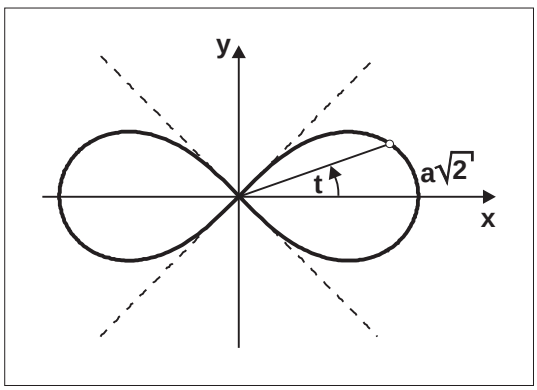
- Die **hyperbolische Spirale**

$$\boxed{\vec{x}(t) = \Big(r(t)\cos t, r(t)\sin t\Big)^T, \quad r(t) := \tfrac{a}{t}, \quad a > 0, \quad t \in (0, 2\pi]}$$

Die Parameterkurve kann auf das Intervall $(0, +\infty)$ fortgesetzt werden.



Die Strophoide



Die Lemniskate

- Die **Strophoide**

$$\boxed{\vec{x}(t) = \Big(r(t)\cos t, r(t)\sin t\Big)^T, \quad r(t) := \tfrac{a\cos 2t}{\cos t}, \quad a > 0, \;\; t \in [-\tfrac{\pi}{4}, \tfrac{\pi}{4}] \cup (\tfrac{\pi}{2}, \tfrac{3\pi}{4}] \cup [\tfrac{5\pi}{4}, \tfrac{3\pi}{2})}$$

- Die **Lemniskate**

$$\boxed{\vec{x}(t) = \Big(r(t)\cos t, r(t)\sin t\Big)^T, \quad r(t) := a\sqrt{2\cos 2t}, \quad a > 0, \;\; t \in [-\tfrac{\pi}{4}, \tfrac{\pi}{4}] \cup [\tfrac{3\pi}{4}, \tfrac{5\pi}{4}]}$$

- Die **Kettenlinie**

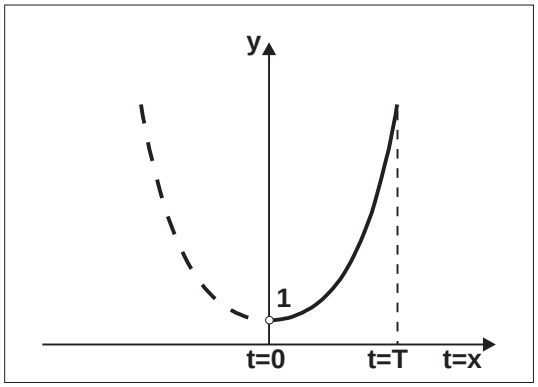
$$\boxed{\vec{x}(t) = (t, \cosh t)^T, \quad t \in [0, T]}$$

  Ein schwerer, nicht dehnbarer aber biegsamer Faden, der in zwei Punkten aufgehängt ist, nimmt die Form einer Kettenlinie an.

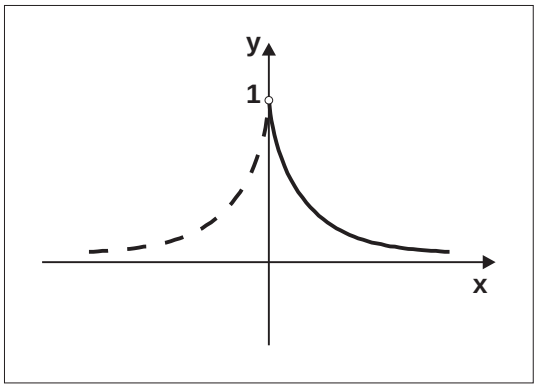
- Die **Traktrix** (Schleppkurve)

$$\boxed{\vec{x}(t) = \Big(t - \tanh t, \frac{1}{\cosh t}\Big)^T, \quad t \geq 0}$$

  Ein Massepunkt an einem nicht dehnbaren Faden fester Länge durchläuft eine Traktrix, wenn das andere Ende des Fadens längs der x–Achse gezogen wird und wenn die Bewegung aus geeigneter Ausgangslage beginnt.



Die Kettenlinie



Die Traktrix oder Schleppkurve

Für die Diskussion weiterer ebener Kurven verweisen wir auf die Standardliteratur, man vgl. auch I.N. Bronstein/K.A. Semendjajew, Taschenbuch der Mathematik. Als Beispiel einer räumlichen Parameterkurve erwähnen wir schließlich:

- Die **Schraubenlinie**

$$\boxed{\vec{x}(t) = \Big(a\cos t, a\sin t, \frac{ht}{2\pi}\Big)^T, \quad a > 0,\, h > 0, \;\; t \in \mathbf{R}}$$

Ist Γ eine (stetig) differenzierbare Parameterkurve, so ist in jedem Punkt $t \in I$ die Ableitung (also der Tangentenvektor) erklärt:

$$\boxed{\frac{d}{dt}\,\vec{x}(t) =: \dot{\vec{x}}(t) = \Big(\dot{x}_1(t), \dot{x}_2(t), \ldots, \dot{x}_n(t)\Big)^T, \quad t \in I,}$$

zum Beispiel für die Astroide ($b = 1$):

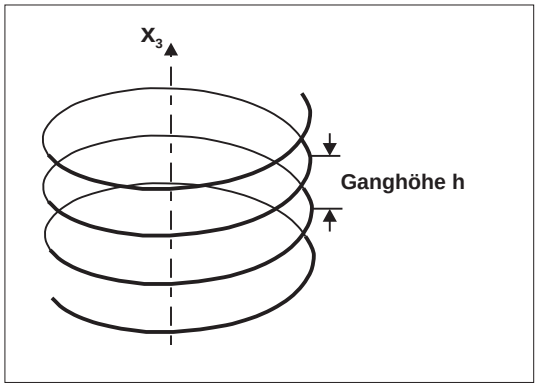
$$\vec{x}(t) = \begin{bmatrix} \cos^3 t \\ \sin^3 t \end{bmatrix}, \qquad \dot{\vec{x}}(t) = \begin{bmatrix} -3\cos^2 t \sin t \\ 3\sin^2 t \cos t \end{bmatrix}, \quad t \in [0, 2\pi].$$

Bemerkung 12.1 Anders als bei skalarwertigen Funktionen kann bei Parameterkurven **nicht** aus der stetigen Differenzierbarkeit auf die "Knickfreiheit" des Graphen geschlossen werden. Die Astroide hat Knicke, obwohl die Ableitung $\dot{\vec{x}}(t)$ stetig ist. Gemäß Definition 7.4 ist in jedem Punkt $t \in I$ mit $\dot{\vec{x}}(t) \neq \vec{0}$ eine Tangente

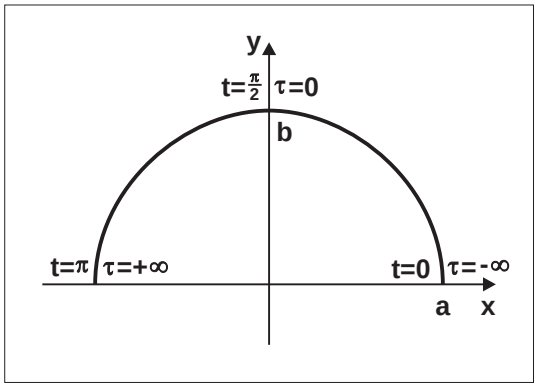
$$\boxed{T := \{\vec{y} \in \mathbf{K}^n \ : \ \vec{y} = \vec{x}(t) + \lambda \dot{\vec{x}}(t), \quad \lambda \in \mathbf{K}\}}$$

an die Parameterkurve Γ definiert. Also können Knicke im Falle von $\vec{x} \in C^1(I; \mathbf{K}^n)$ höchstens in Punkten $t \in I$ mit $\dot{\vec{x}}(t) = \vec{0}$ auftreten. □

Definition 12.2 *Die Parameterdarstellung* (1.2) *einer Kurve* $\Gamma \subset \mathbf{K}^n$ *heiße* **regulär** (*oder die Parameterkurve* Γ *heiße* **glatt** *oder* **knickfrei**), *wenn* $\vec{x} \in C^1(I; \mathbf{K}^n)$ *und* $\|\dot{\vec{x}}(t)\| \neq 0$ *für alle* $t \in I$ *gelten.*



Die Schraubenlinie



Ellipsenhalbbogen mit verschiedenen Parameterdarstellungen

Wir betrachten für $a > 0, b > 0$ die beiden Parameterkurven

$$\vec{x}(t) := \begin{bmatrix} a\cos t \\ b\sin t \end{bmatrix}, \quad t \in [0, \pi], \quad \vec{y}(\tau) := \begin{bmatrix} -a\sin(\operatorname{arc\,tan}_H \tau) \\ b\cos(\operatorname{arc\,tan}_H \tau) \end{bmatrix}, \quad -\infty \leq \tau \leq +\infty.$$

Beide Parameterkurven haben denselben Graphen – den oberen Ellipsenhalbbogen– und dieselbe Orientierung, denn es gilt $\vec{y}(\tau) = \vec{x}(\frac{\pi}{2} + \operatorname{arc\,tan}_H \tau)$. Wir sind also motiviert für die folgende

Definition 12.3 *Eine Abbildung* $t(\tau)$ *mit* $t : I' \to I$ *heiße eine* **zulässige Parametertransformation**, *wenn* $t \in C^1(I')$ *sowie* $t(I') = I$ *und* $(d/d\tau)t(\tau) > 0$ *für alle* $\tau \in I'$ *gelten. In diesem Falle heißen* $\vec{x}(t)$ *und* $\vec{y}(\tau) := \vec{x}(t(\tau))$ **Parameterkurven** *derselben* (*orientierten*) *Kurve* Γ.

Bemerkung 12.2 (a) Aus der obigen Definition folgt offenbar, dass eine Kurve Γ als Äquivalenzklasse von Parameterkurven interpretiert werden kann.

(b) Die Bedingungen $t \in C^1(I')$ und $(d/d\tau)t(\tau) > 0$ stellen sicher, dass die Abbildung $t : I' \to I$ bijektiv und orientierungserhaltend ist. Die Umkehrabbildung $\tau = \tau(t) : I \to I'$ existiert, und diese ist wieder zulässig. □

Ist $\vec{x}(t)$ eine differenzierbare Parameterkurve, so kann man unter allen zulässigen Parametertransformationen $t = t(s)$ diejenigen auszeichnen, die die Bedingung $\|\frac{d}{ds}\vec{x}(s)\| = 1$ erfüllen.

Definition 12.4 *Eine Parameterkurve $\vec{x}(s)$ heiße* **natürliche Parametrisierung** *einer differenzierbaren Kurve Γ, wenn die Bedingung $\|\frac{d}{ds}\vec{x}(s)\| = 1$ für alle $s \in I$ erfüllt ist. Existiert eine solche Parametrisierung, so heiße s* **natürlicher** *oder* **Bogenlängen–Parameter**. *Es ist üblich, die Bezeichnung*

$$\boxed{\frac{d}{ds} =: ' \, '}$$

zu verwenden; also gilt für einen natürlichen Parameter s stets $\|\frac{d}{ds}\vec{x}(s)\| = \|\vec{x}\,'(s)\| = 1$.

BSP. (12.1.3) Die Parameterkurve

$$\vec{x}(s) := \Big(r\cos\frac{s}{r}, r\sin\frac{s}{r}\Big)^T, \quad \vec{x}\,'(s) := \Big(-\sin\frac{s}{r}, \cos\frac{s}{r}\Big)^T, \quad s \in [0, 2\pi r],$$

ist eine natürliche Parametrisierung der Kreislinie $S_r(0)$. In der Tat, es gilt die Bedingung $\|\vec{x}\,'(s)\| = 1$.

Die Frage, welche Kurven eine natürliche Parametrisierung besitzen, beantworten wir in dem folgenden Satz.

Satz 12.1 *Gegeben sei eine reguläre Parameterkurve $\vec{x} = \vec{x}(t)$, $t \in I := [a, b]$. Dann gilt:*

(a) *Durch die Vorschrift*

$$\boxed{s(t) := \int_a^t \|\dot{\vec{x}}(\xi)\|\, d\xi} \tag{1.3}$$

wird ein natürlicher Parameter s zulässig eingeführt.

(b) *Sind s und $\tilde{s}$ zulässige natürliche Parameter der Parameterkurve $\vec{x}$, so folgt stets $s - \tilde{s} = const.$*

Begründungen: (a) Da $\|\dot{\vec{x}}(\cdot)\| : I \to \mathbf{R}$ eine stetige Abbildung ist, erhalten wir $s \in C^1(I)$ für den in (1.3) definierten Parameter $s(t)$. Es ist ferner $\frac{ds}{dt} = \|\dot{\vec{x}}(t)\| > 0$, so dass $s(t)$ und die Umkehrabbildung $t = t(s)$ beide zulässig sind. Darüber hinaus gilt:

$$\|\vec{x}\,'(s)\| = \Big\|\frac{d}{dt}\vec{x}(t)\cdot\frac{dt}{ds}\Big\| = \|\dot{\vec{x}}(t)\|\cdot\frac{1}{\|\dot{\vec{x}}(t)\|} = 1.$$

(b) Es gilt

$$1 = \|\vec{x}\,'(\tilde{s})\| = \Big\|\frac{d}{ds}\vec{x}(s)\cdot\frac{ds}{d\tilde{s}}\Big\| = \Big|\frac{ds}{d\tilde{s}}\Big|,$$

und somit $s = \pm\tilde{s} + const.$ Da aber nur $ds/d\tilde{s} > 0$ zulässig ist, erhalten wir die behauptete Relation $s - \tilde{s} = const.$ □

BSP. (12.1.4) Wir betrachten die **Kettenlinie** $\vec{x}(t) := (t, \cosh t)^T$, $t \geq 0$. Wegen

$$\dot{\vec{x}}(t) = (1, \sinh t)^T, \quad \|\dot{\vec{x}}(t)\| = \sqrt{1 + \sinh^2 t} = \cosh t > 0,$$

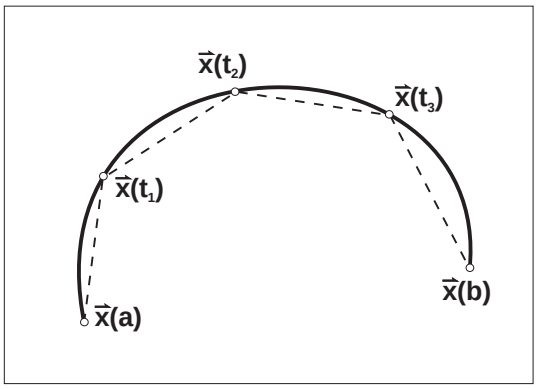
haben wir eine reguläre Parameterkurve vorliegen. Gemäß Satz 12.1 haben alle natürlichen Parameter die Form $\tilde{s} + const$ mit

$$\tilde{s}(t) := \int_0^t \cosh\xi\, d\xi = \sinh t.$$

Es ergibt sich $t = \text{Ar sinh}(s-c)$, womit jede natürliche Parametrisierung der Kettenlinie die folgende Form haben muss:

$$\vec{x}(s) = \begin{bmatrix} \text{Ar sinh}(s-c) \\ \\ \sqrt{1+(s-c)^2} \end{bmatrix}, \quad c = const.$$

An dem vorletzten Beispiel (12.1.3) der Kreislinie $S_r(0)$ ersehen wir, dass der natürliche Parameter s genau die Länge des Kreisbogens angibt. Dieser Zusammenhang besteht allgemein für Parameterkurven mit natürlichem Parameter, wodurch sich die Bezeichnung "Bogenlängen–Parameter" für s erklärt. Zur Erläuterung dieses Zusammenhangs sei eine orientierte stetige Parameterkurve $\vec{x} = \vec{x}(t) \in \mathbf{K}^n$ für $t \in [a,b]$ vorgegeben. Sei Z_m eine endliche Zerlegung des Intervalls $[a,b]$ mit den Stützstellen $a =: t_0 < t_1 < t_2 < \cdots < t_m := b$, und bezeichne $|Z_m| = \max\limits_{1\le j\le m} |t_j - t_{j-1}|$ das Feinheitsmaß dieser Zerlegung. Die Menge aller endlichen Zerlegungen Z_m von $[a,b]$ bezeichnen wir mit Z.



Zum Begriff der Bogenlänge

Werden die Kurvenpunkte $\vec{x}(a), \vec{x}(t_1), \ldots, \vec{x}(b)$ durch einen *Polygonzug* verbunden, so ist dessen Länge die Zahl

$$L(Z_m) := \sum_{j=1}^{m} \|\vec{x}(t_j) - \vec{x}(t_{j-1})\|.$$

Diese Zahl ist sicher eine Näherung für die Länge des Kurvenbogens zwischen $\vec{x}(a)$ und $\vec{x}(b)$. Offenbar wächst $L(Z_m)$ monoton mit m. Ist $L(Z_m)$ nach oben beschränkt, so existiert ein Supremum, welches gleichzeitig Limes der Folge $(L(Z_m))$ ist.

Definition 12.5 *Es sei eine stetige Parameterkurve $\vec{x}(t)$, $t \in [a,b]$, gegeben. Es bezeichne $L(Z_m)$ die Länge des Polygonzugs mit den Ecken $\vec{x}(a), \vec{x}(t_1), \ldots, \vec{x}(b)$ bezüglich einer endlichen Zerlegung Z_m des Intervalls $[a,b]$. Existiert die Zahl*

$$\boxed{L := \sup_{Z_m \in Z} L(Z_m),}$$

so heiße L die **Bogenlänge** *der Parameterkurve $\vec{x}(t)$, und $\vec{x}(t)$ heiße* **rektifizierbar**.

Um eine explizite Berechnungsvorschrift für die Bogenlänge herzuleiten, setzen wir $h_j := t_j - t_{j-1}$ und schreiben $L(Z_m)$ in der Form

$$L(Z_m) = \sum_{j=1}^{m} \left\| \frac{\vec{x}(t_{j-1}+h_j) - \vec{x}(t_{j-1})}{h_j} \right\| \cdot h_j.$$

Ist $\vec{x}(t)$ auf $[a,b]$ fast überall stetig differenzierbar und ist $\|\dot{\vec{x}}(t)\|$ beschränkt, so folgt für $|Z_m| \to 0$ fast überall die Konvergenz

$$\left\| \frac{\vec{x}(t_{j-1}+h_j) - \vec{x}(t_{j-1})}{h_j} \right\| \to \|\dot{\vec{x}}(t_{j-1})\|,$$

und die Folge der RIEMANN–Summen $L(Z_m)$ konvergiert nach dem Integrabilitätskriterium von LEBESGUE (Satz 8.17) gegen das RIEMANN-Integral

$$L = \sup_{Z_m \in Z} L(Z_m) = \int_a^b \|\dot{\vec{x}}(t)\|\, dt.$$

Zusammenfassend erhalten wir:

Satz 12.2 *Die Parameterkurve $\vec{x} : [a,b] \to \mathbf{K}^n$ sei fast überall stetig differenzierbar, und es gelte $\|\dot{\vec{x}}(t)\| \le M$ für alle $t \in [a,b]$. Dann ist $\vec{x}(t)$ rektifizierbar mit der Bogenlänge*

$$\boxed{L = \int_a^b \|\dot{\vec{x}}(t)\|\, dt.}$$

Die Bogenlänge L ist invariant gegenüber zulässigen Parametertransformationen.

Begründung: Um die behauptete Invarianz zu zeigen, betrachten wir eine zulässige Parametertransformation $t = t(\tau) : [\alpha,\beta] \to [a,b]$. Es sei $\vec{y}(\tau) = \vec{x}(t(\tau))$ die zugeordnete Parameterkurve. Diese hat die Bogenlänge

$$\tilde{L} = \int_\alpha^\beta \Big\| \frac{d}{d\tau}\vec{y}(\tau) \Big\|\, d\tau = \int_\alpha^\beta \Big\| \frac{d}{dt}\vec{x}(t(\tau)) \underbrace{\frac{dt}{d\tau}}_{>0} \Big\|\, d\tau \overset{t=t(\tau)}{=} \int_a^b \|\dot{\vec{x}}(t)\|\, dt = L.$$

Bemerkung 12.3 (a) Wegen $s(t) = \int_a^t \|\dot{\vec{x}}(\xi)\|\, d\xi$ haben wir $s(b) = L$. Dies rechtfertigt nun endgültig die Bezeichnung "Bogenlängen–Parameter" für s. In der Tat misst $s(t)$, beginnend beim Anfangspunkt $\vec{x}(a)$ einer Parameterkurve $\vec{x}(t)$, die Länge des auf der Kurve zurückgelegten Weges, wenn man mit $t > a$ voranschreitet.

(b) Die Größe

$$\boxed{ds := \|\dot{\vec{x}}(t)\|\, dt = \Big(\sum_{j=1}^n |\dot{x}_j(t)|^2 \Big)^{1/2} dt}$$

wird häufig das **Bogenelement** oder das **Bogendifferential** genannt. □

Folgerung 12.1 *Reeller Fall* $\mathbf{K} := \mathbf{R}$*:*

(a) *Für* **ebene** *Parameterkurven $\vec{x}(t) := (x(t), y(t))^T$, $t \in [a,b]$, gilt*

$$\boxed{ds = \sqrt{\dot{x}^2(t) + \dot{y}^2(t)}\, dt, \quad L = \int_a^b ds = \int_a^b \sqrt{\dot{x}^2(t) + \dot{y}^2(t)}\, dt.} \tag{1.4}$$

(b) *Bei* **expliziter Darstellung** $y = y(x)$ *kann* $t := x$ *gesetzt werden:*

$$\boxed{ds = \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx, \quad L = \int_{x_a}^{x_b} ds = \int_{x_a}^{x_b} \sqrt{1 + y'^2(x)}\, dx.} \tag{1.5}$$

(c) *Liegt die ebene Parameterkurve* $\vec{x}(t)$ *speziell in* **Polarkoordinaten** $x = r(t)\cos t, y = r(t)\sin t$ *vor, so gelten* $\dot{x}(t) = \dot{r}(t)\cos t - r(t)\sin t$, $\dot{y}(t) = \dot{r}(t)\sin t + r(t)\cos t$, *und somit*

$$\boxed{ds = \sqrt{r^2(t) + \left(\frac{dr}{dt}\right)^2}\, dt, \quad L = \int_a^b ds = \int_a^b \sqrt{r^2(t) + \dot{r}^2(t)}\, dt.} \tag{1.6}$$

BSP. (12.1.5) Bogenlänge einer **Schraubenlinie** der Ganghöhe $h > 0$ für einen Schraubengang. Wir haben

$$\vec{x}(t) = \begin{bmatrix} a\cos t \\ a\sin t \\ ht/2\pi \end{bmatrix}, \quad \dot{\vec{x}}(t) = \begin{bmatrix} -a\sin t \\ a\cos t \\ h/2\pi \end{bmatrix}, \quad ds = \|\dot{\vec{x}}(t)\|dt = \sqrt{a^2 + \left(\frac{h}{2\pi}\right)^2}\, dt, \quad t \in [0, 2\pi].$$

Somit folgt

$$\boxed{L = \int_0^{2\pi} ds = \sqrt{(2\pi a)^2 + h^2}.}$$

Dies ist aber nach dem Satz von PYTHAGORAS auch elementargeometrisch klar.

BSP. (12.1.6) Bogenlänge der **Kettenlinie**. Wir haben

$$y(x) = \cosh x, \quad ds = \sqrt{1 + y'^2(x)}dx = \sqrt{1 + \sinh^2 x}dx = \cosh x\, dx, \quad x \in [0, T].$$

Somit folgt

$$\boxed{L = \int_0^T ds = \int_0^T \cosh x\, dx = \sinh T.}$$

BSP. (12.1.7) Bogenlänge der **Kardioide**. Wir haben

$$r(t) = a(1 + \cos t), \quad \dot{r}(t) = -a\sin t, \quad ds = \sqrt{r^2(t) + \dot{r}^2(t)}\, dt = 2a|\cos\tfrac{t}{2}|\, dt, \quad t \in [0, 2\pi].$$

Somit folgt

$$\boxed{L = 2\int_0^{\pi} ds = 4a\int_0^{\pi} \cos\tfrac{t}{2}\, dt = 8a.}$$

BSP. (12.1.8) Bogenlänge der **Ellipse**. Wir haben

$$\vec{x}(t) = \begin{bmatrix} a\cos t \\ b\sin t \end{bmatrix}, \quad \dot{\vec{x}}(t) = \begin{bmatrix} -a\sin t \\ b\cos t \end{bmatrix},$$

und somit

$$ds = \|\dot{\vec{x}}(t)\|\, dt = \sqrt{a^2\sin^2 t + b^2\cos^2 t}\, dt = a\sqrt{1 - \epsilon^2\cos^2 t}\, dt, \quad t \in [0, 2\pi].$$

Hier ist $\epsilon = \frac{\sqrt{a^2-b^2}}{a}$ die numerische Exzentrizität, wobei $a > b$ vorausgesetzt wird. Wegen der Symmetrie der Ellipse braucht man nur die Bogenlänge der Viertelellipse $(\pi/2) \le t \le \pi$ zu berechnen. Die Gesamtlänge beträgt dann:

$$\boxed{L = 4a \int_{\pi/2}^{\pi} \sqrt{1 - \epsilon^2 \cos^2 t}\, dt \overset{\xi = t - \pi/2}{=} 4a \int_0^{\pi/2} \sqrt{1 - \epsilon^2 \sin^2 \xi}\, d\xi =: 4a\mathbf{E}(\epsilon).}$$

Das Integral

$$\mathbf{E}(k) =: \int_0^{\pi/2} \sqrt{1 - k^2 \sin^2 \xi}\, d\xi, \quad 0 < k < 1,$$

ist nicht mehr elementar darstellbar. Es heißt **vollständiges elliptisches Integtral 2.Gattung**. Tabellen der Funktionswerte von $\mathbf{E}(k)$ findet man zum Beispiel in M. Abramowitz/I. Stegun, Handbook of Mathematical Functions. In BSP. (9.2.15) hatten wir für $0 < \epsilon \ll 1$ – das heißt für "runde" Ellipsen $a \approx b$ – die folgende Näherungsformel begründet, die in der Geodäsie Verwendung findet:

$$L \approx \pi\Big(3\frac{a+b}{2} - \sqrt{ab}\Big) + O(\epsilon^8).$$

BSP. (12.1.9) Bogenlänge eines **Zykloidenbogens**. Wir haben

$$\vec{x}(t) = \begin{bmatrix} t - \sin t \\ 1 - \cos t \end{bmatrix}, \quad \dot{\vec{x}}(t) = \begin{bmatrix} 1 - \cos t \\ \sin t \end{bmatrix}, \quad ds = \|\dot{\vec{x}}(t)\|\, dt = 2|\sin \tfrac{t}{2}|\, dt, \quad t \in [0, 2\pi].$$

Somit folgt

$$\boxed{L = 2\int_0^{\pi} ds = 4\int_0^{\pi} \sin \tfrac{t}{2}\, dt = 8.}$$

BSP. (12.1.10) Bogenlänge einer Windung der **logarithmischen Spirale**. Wir haben

$$r(t) = ae^{-bt}, \;\; \dot{r}(t) = -abe^{-bt} = -br(t), \;\; ds = \sqrt{r^2(t) + \dot{r}^2(t)}\, dt = \sqrt{1+b^2}\, r(t)\, dt, \quad a > 0, b > 0.$$

Somit folgt

$$s(t) = \int_0^t ds = a\sqrt{1+b^2} \int_0^t e^{-b\xi}\, d\xi = \frac{a}{b}\sqrt{1+b^2}\,(1 - e^{-bt})$$

sowie

$$\boxed{L_{2\pi} := s(2\pi) = \frac{a}{b}\sqrt{1+b^2}\,(1 - e^{-2\pi b}).}$$

Die gesamte "Innenspirale" $0 \le t < \infty$ ist rektifizierbar:

$$L_\infty = \lim_{t \to +\infty} s(t) = \frac{a}{b}\sqrt{1+b^2}.$$

Wir erhalten ferner für $b \to 0+$ den Grenzfall des Kreises vom Radius a:

$$L_{\mathrm{Kr}} = \lim_{b \to 0+} L_{2\pi} = \lim_{b \to 0+} \frac{a}{b}\sqrt{1+b^2}\,(1 - e^{-2\pi b}) \overset{\mathrm{L'Hosp}}{=} 2\pi a \lim_{b \to 0+} e^{-2\pi b} = 2\pi a.$$

Das Beispiel der Ellipse zeigt, dass die Berechnung des Bogenlängen–Parameters s im allgemeinen nicht *explizit* durchführbar ist. Somit scheint auch die natürliche Parametrisierung einer

Kurve in Frage gestellt zu sein. Man kann jedoch Differentiation nach dem Bogenlängen–Parameter s auch *ohne explizite Kenntnis* von s durchführen. Wegen $ds/dt = \|\dot{\vec{x}}(t)\|$ gilt nämlich $dt/ds = \|\dot{\vec{x}}(t)\|^{-1}$, und somit

$$\boxed{\frac{d}{ds} = \frac{1}{\|\dot{\vec{x}}(t)\|}\,\frac{d}{dt}.} \tag{1.7}$$

BSP. (12.1.11) Wir betrachten die **logarithmische Spirale** $r(t) := ae^{-bt}$ mit $a, b > 0$, vgl. das vorangegangene BSP. (12.1.10). Wir haben $\|\dot{\vec{x}}(t)\| = \sqrt{1+b^2}\,r(t)$, und somit für $s = s(t)$:

$$\vec{x}(s) = ae^{-bt}\begin{bmatrix}\cos t\\ \sin t\end{bmatrix},\quad \vec{x}\,'(s) = \frac{1}{\sqrt{1+b^2}}\begin{bmatrix}-\sin t - b\cos t\\ \cos t - b\sin t\end{bmatrix},\quad \vec{x}\,''(s) = \frac{e^{bt}}{a(1+b^2)}\begin{bmatrix}-\cos t + b\sin t\\ -\sin t - b\cos t\end{bmatrix}.$$

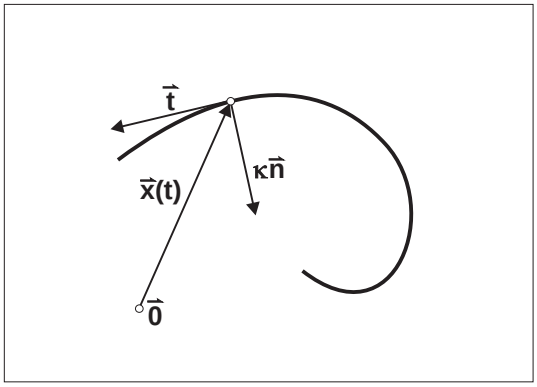
Man berechnet das Skalarprodukt $\langle \vec{x}\,'(s), \vec{x}\,''(s)\rangle = 0$; das heißt $\vec{x}\,'(s) \perp \vec{x}\,''(s)$. Dies ist kein Zufall, denn weil s ein natürlicher Parameter ist, gilt ja $\|\vec{x}\,'(s)\| = 1$. Aus der Ableitungsregel (3.1(d)) in Abschnitt 7.3 folgt somit stets

$$0 = \frac{d}{ds}\,\|\vec{x}\,'(s)\|^2 = 2\mathrm{Re}\,\langle \vec{x}\,'(s), \vec{x}\,''(s)\rangle.$$

Eine genauere Untersuchung dieses Zusammenhangs werden wir im nächsten Abschnitt vornehmen.

12.2 Tangente, Krümmungsvektor

Die folgenden Ausführungen sind nur im Reellen von praktischem Belang. Wir verlassen deshalb den allgemeinen Fall und setzen stets $\mathbf{K} := \mathbf{R}$ voraus.



Zur Krümmung einer Parameterkurve

Gegeben sei eine Kurve $\Gamma \subset \mathbf{R}^n$ mit regulärer Parameterdarstellung $\vec{x} = \vec{x}(t)$ für $t \in I$. Dann gilt $\dot{\vec{x}}(t) \neq \vec{0}$ für alle $t \in I$, und in dieser Situation ist gemäß Abschnitt 7.3 die **Tangente**

$$T = \{\vec{y} \in \mathbf{R}^n \;:\; \vec{y} = \vec{x}(t) + \lambda\dot{\vec{x}}(t), \quad \lambda \in \mathbf{R}\}$$

an die Kurve Γ in jedem Kurvenpunkt $\vec{x}(t)$ erklärt. Der Vektor $\dot{\vec{x}}(t)$ heißt der **Tangentenvektor** im Punkte $\vec{x}(t)$ an Γ. Betrachten wir den **Tangenteneinheitsvektor** $\vec{t} := \dot{\vec{x}}(t)/\|\dot{\vec{x}}(t)\|$, so finden wir wegen (1.7):

$$\vec{t} = \frac{\dot{\vec{x}}(t)}{\|\dot{\vec{x}}(t)\|} = \frac{1}{\|\dot{\vec{x}}(t)\|}\,\frac{d}{dt}\,\vec{x}(t) = \frac{d}{ds}\,\vec{x}(s) = \vec{x}\,'(s). \tag{2.1}$$

Definition 12.6 (a) *Sei Γ eine reguläre Parameterkurve und sei $\vec{x}(s)$ ihre natürliche Parametrisierung. Dann heiße*

$$\boxed{\vec{t} := \vec{x}\,'(s), \quad \|\vec{t}\| = 1,} \tag{2.2}$$

der **Tangenteneinheitsvektor** *im Punkte $\vec{x}(s)$ an Γ oder die* **Tangente** *schlechthin.*

(b) *Für $\vec{x} \in C^2(I;\mathbf{R}^n)$ existiert die Ableitung $\vec{t}\,' = \vec{x}\,''$. Wegen $0 = \frac{d}{ds}\|\vec{t}\|^2 = 2\langle \vec{t}, \vec{t}\,'\rangle$ folgt stets $\vec{t}\,' \perp \vec{t}$. Der Vektor $\vec{t}\,' = \vec{x}\,''(s)$ heiße* **Krümmungsvektor** *im Punkte $\vec{x}(s)$ an Γ. In allen Punkten $\vec{x}(s)$ mit $\vec{t}\,' \neq \vec{0}$ wird durch die Beziehung*

$$\boxed{\vec{t}\,' =: \kappa\vec{n}, \quad \|\vec{n}\| = 1,} \tag{2.3}$$

die **Krümmung** *κ und die* **Normale** *$\vec{n}$ der Kurve Γ bis auf einen Faktor ± 1 eindeutig festgelegt. Für $\vec{t}\,' = \vec{0}$ setzt man $\kappa = 0$.*

BSP. (12.2.1) Wir betrachten die **logarithmische Spirale** $r(t) := ae^{-bt}$ mit $a, b > 0$, vgl. BSP. (12.1.11). Wie wir dort bereits gezeigt haben, gilt $\|\dot{\vec{x}}(t)\| = \sqrt{1+b^2}\, r(t)$ und somit für $s = s(t)$:

$$\vec{x}(s) = ae^{-bt}\begin{bmatrix}\cos t\\ \sin t\end{bmatrix}, \quad \vec{t} = \vec{x}\,'(s) = \frac{1}{\sqrt{1+b^2}}\begin{bmatrix}-\sin t - b\cos t\\ \cos t - b\sin t\end{bmatrix}, \quad \|\vec{t}\| = 1.$$

Es gelten ferner

$$\vec{t}\,' = \vec{x}\,''(s) = \frac{e^{bt}}{a(1+b^2)}\begin{bmatrix}-\cos t + b\sin t\\ -\sin t - b\cos t\end{bmatrix}, \quad \|\vec{t}\,'\| = \frac{e^{bt}}{a(1+b^2)^{1/2}},$$

und hieraus resultieren

$$\boxed{\kappa = \frac{\pm e^{bt}}{a(1+b^2)^{1/2}}, \quad \vec{n} = \frac{\pm 1}{\sqrt{1+b^2}}\begin{bmatrix}-\cos t + b\sin t\\ -\sin t - b\cos t\end{bmatrix}.}$$

BSP. (12.2.2) Der Fall $\kappa = 0$ liegt genau für $\vec{t} = \vec{t}_0 = const = \vec{x}\,'(s)$ vor. Integration führt auf die **Gerade** $\vec{x}(s) = \vec{x}_0 + s\vec{t}_0$, $s \in \mathbf{R}$.

BSP. (12.2.3) Ist t der physikalische Zeitparameter, und ist durch $\vec{x}(t)$ die Bewegung eines Massepunktes auf einer Bahnkurve Γ vorgegeben, so sind gemäß Abschnitt 7.3 die folgenden physikalischen Größen bestimmt:

- der **Geschwindigkeitsvektor**

$$\boxed{\vec{v}(t) := \dot{\vec{x}}(t) = v(t)\vec{t},}$$

  worin $v(t) := \|\vec{v}(t)\|$ die **Absolutgeschwindigkeit** der Bahnbewegung ist

- der **Beschleunigungsvektor**

$$\boxed{\vec{b}(t) := \ddot{\vec{x}}(t) = \Big(\frac{d}{dt}\,v(t)\Big)\vec{t} + v(t)\frac{d}{dt}\,\vec{t} = \dot{v}(t)\vec{t} + v(t)\vec{t}\,'\cdot\frac{ds}{dt} = \dot{v}(t)\vec{t} + v^2(t)\kappa\vec{n}.}$$

  Hierbei heißen die Komponenten

  - $\dot{v}(t)\vec{t}$: die **Tangentialkomponente** der Beschleunigung

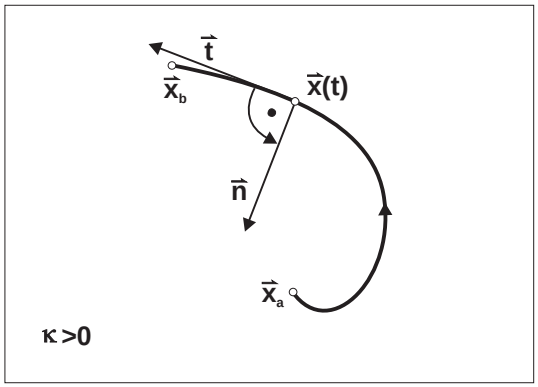
- $v^2(t)\kappa\vec{n}$: die **Normalkomponente** der Beschleunigung

Im Falle einer gleichförmigen Bewegung $\dot{v} = 0$ hat man also $\|\vec{b}(t)\| = v^2(t)|\kappa|$.

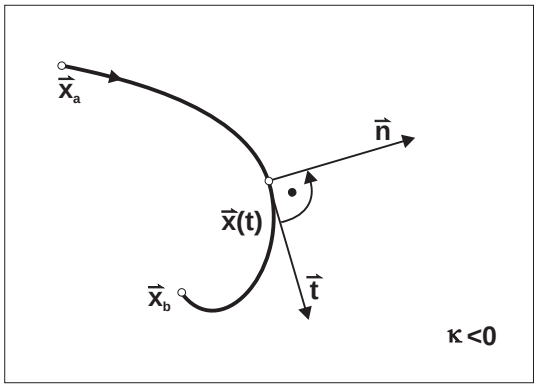
In der Definition von Krümmung und Normale sind die Größen κ und $\vec{n}$ nur bis auf das Vorzeichen eindeutig festgelegt. Um zu einer eindeutigen Definition zu gelangen, hat man die Wahl, entweder das Voreichen von κ zu fixieren, oder eine Orientierung der Normalen $\vec{n}$ vorzugeben. Es hat sich in der Literatur eingebürgert, bei ebenen Kurven und bei räumlichen Kurven jeweils auf verschiedene Weise vorzugehen, wobei wohl historische Gründe ausschlaggebend sind. Wir betrachten deshalb zunächst

(I) Kurven in $\mathbf{R}^2$ (ebene Kurven)

Die ebene Kurve $\Gamma \subset \mathbf{R}^2$ habe eine reguläre Parameterdarstellung $\vec{x}(t)$, wobei für das Folgende hinreichende Differenzierbarkeit von $\vec{x}$ vorausgesetzt sei. Eine **Orientierung** der Normalen $\vec{n}$ sei so festgelegt, dass $\vec{n}$ stets durch eine Drehung um $+\frac{\pi}{2}$ aus der Tangente $\vec{t}$ hervorgehe:



Normale $\vec{n}$ und Krümmung κ sind orientierungsabhängig



Normale $\vec{n}$ und Krümmung κ sind orientierungsabhängig

Definition 12.7 *Für ebene reguläre Parameterkurven $\vec{x} = \vec{x}(t)$ mit Tangente $\vec{t} = \vec{x}\,'(s) = \dot{\vec{x}}(t)/\|\dot{\vec{x}}(t)\|$ sind die* **Normale** *$\vec{n}$ und die* **Krümmung** *κ der Kurve wie folgt definiert:*

$$\boxed{\vec{n} := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \vec{t}, \quad \kappa := \langle \vec{t}\,', \vec{n} \rangle.}$$

Das heißt, $\vec{n}$ geht aus $\vec{t}$ durch eine Drehung um $+\frac{\pi}{2}$ hervor.

Bemerkung 12.4 (a) In dieser Definition sind $\vec{n}$ und κ orientierungsabhängig: ändert sich die Orientierung der Kurve Γ, so ändern sich auch die Orientierung von $\vec{n}$ und das Vorzeichen von κ.

(b) Die Normale $\vec{n}$ ist auch dann wohldefiniert, wenn $\vec{t}\,' = \vec{0}$ gilt.

(c) Es gilt natürlich wieder $\vec{t}\,' = \kappa\vec{n}$. Die Vektoren $\vec{t}$ und $\vec{n}$ bilden eine ON–Basis des $\mathbf{R}^2$, so dass jeder Vektor $\vec{y} \in \mathbf{R}^2$ die Darstellung

$$\vec{y} = \langle \vec{y}, \vec{t} \rangle\, \vec{t} + \langle \vec{y}, \vec{n} \rangle\, \vec{n}$$

gestattet. Speziell für $\vec{y} := \vec{t}\,'$ gilt ja $\langle \vec{t}\,', \vec{t} \rangle = 0$ und somit $\vec{t}\,' = \langle \vec{t}\,', \vec{n} \rangle\, \vec{n}$. □

Um zu expliziten Formeln für die Krümmung κ zu gelangen, sei $\vec{x}(t) = (x(t), y(t))^T$ die reguläre Parameterdarstellung der ebenen Kurve Γ. Nachfolgend unterdrücken wir jeweils das Argument t, und ein $''\cdot''$ bezeichne stets die Ableitung nach t. Für $\vec{x} \in C^2(I; \mathbf{R}^2)$ existieren die folgenden Ausdrücke:

$$\vec{t} = \frac{1}{\sqrt{\dot{x}^2 + \dot{y}^2}} \begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix}, \quad \vec{n} = \frac{1}{\sqrt{\dot{x}^2 + \dot{y}^2}} \begin{bmatrix} -\dot{y} \\ \dot{x} \end{bmatrix}, \quad \vec{t}\,' = \frac{1}{\sqrt{\dot{x}^2 + \dot{y}^2}} \frac{d}{dt} \vec{t} = \frac{\dot{x}\ddot{y} - \dot{y}\ddot{x}}{(\dot{x}^2 + \dot{y}^2)^2} \begin{bmatrix} -\dot{y} \\ \dot{x} \end{bmatrix}.$$

Hieraus resultiert

$$\kappa = \langle \vec{t}\,', \vec{n} \rangle = \frac{\dot{x}\ddot{y} - \dot{y}\ddot{x}}{(\dot{x}^2 + \dot{y}^2)^{3/2}}.$$

Analoge Formeln erhält man für Kurven Γ in expliziter Darstellung oder in Polarkoordinaten.

Folgerung 12.2 *Die Krümmung κ einer regulären ebenen Kurve $\Gamma \subset \mathbf{R}^2$ ist wie folgt gegeben:*

(a) *bei* **Parameterdarstellung** $\vec{x}(t) = (x(t), y(t))^T$ *durch*

$$\boxed{\kappa = \frac{\dot{x}\ddot{y} - \dot{y}\ddot{x}}{(\dot{x}^2 + \dot{y}^2)^{3/2}}}$$

(b) *bei* **expliziter Darstellung** $y = f(x)$ *durch*

$$\boxed{\kappa = \frac{y''}{(1 + y'^2)^{3/2}}}$$

(c) *bei* **Polardarstellung** $r = r(\varphi)$ *durch*

$$\boxed{\kappa = \frac{r^2 + 2\dot{r}^2 - r\ddot{r}}{(r^2 + \dot{r}^2)^{3/2}}}$$

BSP. (12.2.4) Wir betrachten eine **Ellipse** mit Parameterdarstellung $\vec{x}(t) = (a\cos t, b\sin t)^T$. Es gelten hier

$$\dot{x} = -a\sin t, \quad \ddot{x} = -a\cos t, \quad \dot{y} = b\cos t, \quad \ddot{y} = -b\sin t.$$

Somit folgen $\dot{x}\ddot{y} - \dot{y}\ddot{x} = ab$, $\dot{x}^2 + \dot{y}^2 = a^2\sin^2 t + b^2\cos^2 t =: \Delta(t)$, und daraus resultieren:

$$\boxed{\vec{t} = \frac{1}{\sqrt{\Delta(t)}} \begin{bmatrix} -a\sin t \\ b\cos t \end{bmatrix}, \quad \vec{n} = \frac{1}{\sqrt{\Delta(t)}} \begin{bmatrix} -b\cos t \\ -a\sin t \end{bmatrix}, \quad \kappa = \frac{ab}{(\Delta(t))^{3/2}}.}$$

BSP. (12.2.5) Wir betrachten eine **Kettenlinie** mit expliziter Darstellung $y = f(x) = \cosh x$, die auch als spezielle Parameterdarstellung $\vec{x}(x) = (x, y(x))^T$ gedeutet werden kann. Es gelten hier

$$y'(x) = \sinh x, \quad y''(x) = \cosh x, \quad 1 + y'^2(x) = \cosh^2 x := \Delta(x),$$

und daraus resultieren

$$\boxed{\vec{t} = \frac{1}{\sqrt{\Delta(x)}} \begin{bmatrix} 1 \\ y' \end{bmatrix} = \frac{1}{\cosh x} \begin{bmatrix} 1 \\ \sinh x \end{bmatrix}, \quad \vec{n} = \frac{1}{\sqrt{\Delta(x)}} \begin{bmatrix} -y' \\ 1 \end{bmatrix} = \frac{1}{\cosh x} \begin{bmatrix} -\sinh x \\ 1 \end{bmatrix}, \quad \kappa = \frac{1}{\cosh^2 x}.}$$

BSP. (12.2.6) Wir betrachten eine **Kardioide** mit Polardarstellung $r = r(\varphi) = a(1+\cos\varphi),\ 0 \le \varphi \le 2\pi$. Es gelten hier

$$\dot r = -a\sin\varphi, \quad \ddot r = -a\cos\varphi, \quad \dot x = \dot r\cos\varphi - r\sin\varphi, \quad \dot y = \dot r\sin\varphi + r\cos\varphi.$$

Somit folgen $r^2 + 2\dot r^2 - r\ddot r = 3a^2(1+\cos\varphi) = 3ar,\ \dot x^2 + \dot y^2 = r^2 + \dot r^2 = 2a^2(1+\cos\varphi) = 2ar =: \Delta(\varphi)$, und daraus resultieren:

$$\boxed{\vec t = \frac{1}{\sqrt{\Delta(\varphi)}}\begin{bmatrix}\dot x\\ \dot y\end{bmatrix} = \frac{1}{2|\cos\frac{\varphi}{2}|}\begin{bmatrix}-\sin\varphi - \sin 2\varphi\\ \cos\varphi + \cos 2\varphi\end{bmatrix},\ \ \vec n = -\frac{1}{2|\cos\frac{\varphi}{2}|}\begin{bmatrix}\cos\varphi + \cos 2\varphi\\ \sin\varphi + \sin 2\varphi\end{bmatrix}}$$

sowie

$$\boxed{\kappa = \frac{3}{2}\frac{1}{\sqrt{2ar}}.}$$

BSP. (12.2.7) Wir betrachten einen **Kreis** vom Radius ρ mit Polardarstellung $r = r(\varphi) = \rho,\ 0 \le \varphi \le 2\pi$. Es gelten hier

$$\dot{\vec x}(\varphi) = (-\rho\sin\varphi, \rho\cos\varphi)^T, \quad \|\dot{\vec x}(\varphi)\| = \rho,$$

und daraus resultieren

$$\boxed{\vec t = \begin{bmatrix}-\sin\varphi\\ \cos\varphi\end{bmatrix}, \quad \vec t\,' = \frac{1}{\rho}\frac{d}{d\varphi}\vec t = \frac{1}{\rho}\begin{bmatrix}-\cos\varphi\\ -\sin\varphi\end{bmatrix} = \kappa\vec n, \quad \vec n = \begin{bmatrix}-\cos\varphi\\ -\sin\varphi\end{bmatrix}, \quad \kappa = \frac{1}{\rho}.}$$

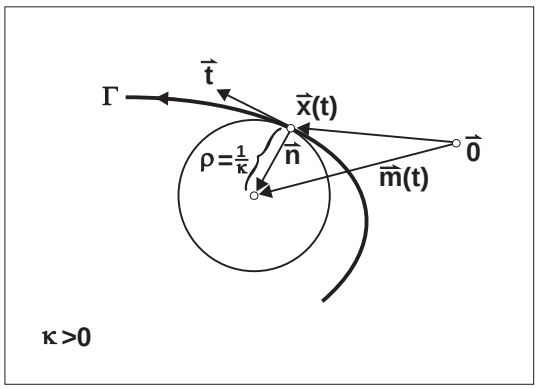
Die **Krümmung des Kreises** ist der reziproke Radius. Dieser Zusammenhang wird in der folgenden Definition verwendet.

Definition 12.8 *Für eine ebene reguläre Kurve $\Gamma \subset \mathbf{R}^2$ mit Parameterdarstellung $\vec x = \vec x(t)$ heiße die Zahl*

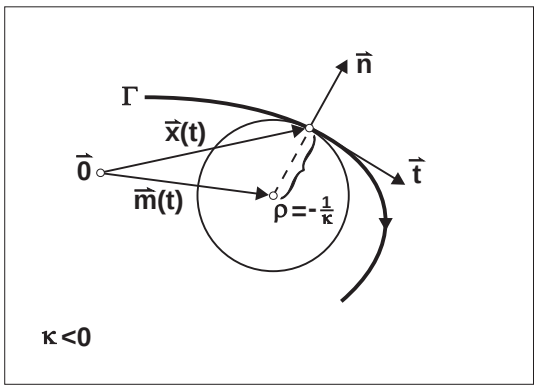
$$\boxed{\rho := \frac{1}{|\kappa|}}$$

der **Krümmungsradius** *von Γ im Punkte $\vec x(t)$, sofern $\kappa = \kappa(t) \ne 0$ gilt. Die Zahl ρ ist der Radius desjenigen Kreises, der die Kurve Γ im Punkte $\vec x(t)$ berührt und der dieselbe Krümmung κ besitzt wie die Kurve. Dieser Kreis heißt der* **Krümmungskreis** *oder* **Schmiegkreis** *von Γ im Punkte $\vec x(t)$. Er liegt auf der konkaven Seite von Γ, und sein Mittelpunkt ist gegeben durch*

$$\boxed{\vec m(t) = \vec x(t) + \frac{1}{\kappa(t)}\,\vec n(t).} \tag{2.4}$$



$\vec m(t) = \vec x(t) + \frac{1}{\kappa(t)}\,\vec n(t)$



$\vec m(t) = \vec x(t) - \left(-\frac{1}{\kappa(t)}\right)\vec n(t)$

Definition 12.9 *Diejenige ebene Kurve*

$$\boxed{\Gamma_1 := \{\vec{m}(t) \ : \ \vec{m}(t) = \vec{x}(t) + \frac{1}{\kappa(t)}\,\vec{n}(t), \ \ t \in I\},}$$

die von den Krümmungsmittelpunkten einer gegebenen ebenen Kurve Γ mit Parameterdarstellung $\vec{x} = \vec{x}(t)$, $t \in I$, beschrieben wird, heiße die **Evolute** *von Γ. Die Ausgangskurve Γ selbst nennt man auch die* **Evolvente** *von Γ_1.*

Folgerung 12.3 *Die Evolute einer regulären ebenen Kurve $\Gamma \subset \mathbf{R}^2$ mit nichtverschwindender Krümmung $\kappa \neq 0$ ist wie folgt gegeben:*

(a) *bei* **Parameterdarstellung** $\vec{x}(t) = (x(t), y(t))^T$ *durch*

$$\boxed{X(t) = x(t) - \dot{y}(t)\,\frac{\dot{x}^2 + \dot{y}^2}{\dot{x}\ddot{y} - \dot{y}\ddot{x}}, \quad Y(t) = y(t) + \dot{x}(t)\,\frac{\dot{x}^2 + \dot{y}^2}{\dot{x}\ddot{y} - \dot{y}\ddot{x}}}$$

(b) *bei* **expliziter Darstellung** $y = f(x)$ *durch*

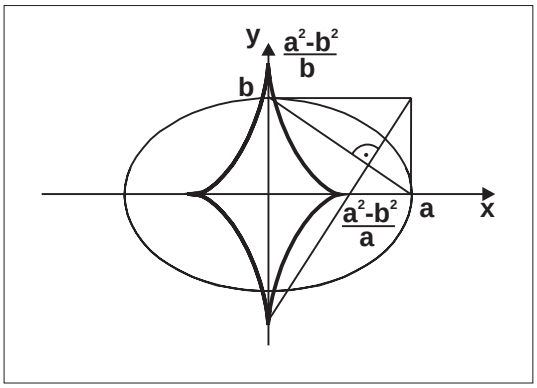
$$\boxed{X(x) = x - y'(x)\,\frac{1 + y'^2}{y''}, \quad Y(x) = y(x) + \frac{1 + y'^2}{y''}}$$

(c) *bei* **Polardarstellung** $r = r(\varphi)$ *durch*

$$\boxed{X(\varphi) = x - \frac{(r^2 + \dot{r}^2)(\dot{r}\sin\varphi + r\cos\varphi)}{r^2 + 2\dot{r}^2 - r\ddot{r}}, \quad Y(\varphi) = y + \frac{(r^2 + \dot{r}^2)(\dot{r}\cos\varphi - r\sin\varphi)}{r^2 + 2\dot{r}^2 - r\ddot{r}}}$$

BSP. (12.2.8) Wir bestimmen die Evolute der **Ellipse** aus BSP. (12.2.4). Wir setzen wie dort wieder $\dot{x}^2(t) + \dot{y}^2(t) = a^2\sin^2 t + b^2\cos^2 t =: \Delta(t)$. Gemäß Definition 12.9 erhalten wir die Evolute

$$\boxed{\vec{m}(t) = \vec{x}(t) + \frac{1}{\kappa(t)}\,\vec{n}(t) = \begin{bmatrix} a\cos t \\ b\sin t \end{bmatrix} + \Delta(t)\begin{bmatrix} -\frac{1}{a}\cos t \\ -\frac{1}{b}\sin t \end{bmatrix} = \begin{bmatrix} \frac{1}{a}\,(a^2 - b^2)\cos^3 t \\ \frac{1}{b}\,(b^2 - a^2)\sin^3 t \end{bmatrix} =: \begin{bmatrix} X(t) \\ Y(t) \end{bmatrix}.}$$



Die Evolute einer Ellipse

Das heißt, die Evolute einer Ellipse ist eine Astroide.

BSP. (12.2.9) Wir bestimmen die Evolute der **Traktrix**

$$\vec{x}(t) = \Big(t - \tanh t, \frac{1}{\cosh t}\Big)^T, \ t \geq 0,$$

aus BSP. (12.1.2). Wir haben hier $\dot{x}(t) = 1 - \cosh^{-2} t = \tanh^2 t$, $\ddot{x}(t) = 2\sinh t/\cosh^3 t$, $\dot{y}(t) = -\sinh t/\cosh^2 t$, $\ddot{y}(t) = (\sinh^2 t - 1)/\cosh^3 t$, und somit

$$\dot{x}^2 + \dot{y}^2 = \tanh^2 t, \quad \dot{x}\ddot{y} - \dot{y}\ddot{x} = \frac{\sinh^2 t}{\cosh^3 t}, \quad \frac{\dot{x}^2 + \dot{y}^2}{\dot{x}\ddot{y} - \dot{y}\ddot{x}} = \cosh t.$$

Wir erhalten aus Folgerung 12.3(a) die Evolute

$$\boxed{\vec{m}(t) = \begin{bmatrix} t - \tanh t \\ 1/\cosh t \end{bmatrix} + \cosh t \begin{bmatrix} \sinh t/\cosh^2 t \\ \tanh^2 t \end{bmatrix} = \begin{bmatrix} t \\ \cosh t \end{bmatrix} =: \begin{bmatrix} X(t) \\ Y(t) \end{bmatrix}, \quad t \geq 0.}$$

Das heißt, die Evolute der Traktrix ist die Kettenlinie.

Bemerkung 12.5 (a) Die beiden Vektoren $\vec{t}$ und $\vec{n}$ bilden in jedem Punkt $\vec{x}(t)$ einer regulären ebenen Kurve Γ eine ON–Basis des $\mathbf{R}^2$. Man nennt die Einheitsvektoren $\vec{t}$ und $\vec{n}$ auch das **begleitende Zweibein** von Γ.

(b) Hat die reguläre ebene Kurve Γ eine Parameterdarstellung $\vec{x} \in C^2(I; \mathbf{R}^2)$, so ist die Ableitung $\vec{n}\,'$ wohldefiniert. Mit $Q := \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ folgern wir wegen $\vec{n} = Q\vec{t}$ und wegen $Q^2 = -Id$:

$$\vec{n}\,' = Q\vec{t}\,' = Q\kappa\vec{n} = \kappa Q^2\vec{t} = -\kappa\vec{t}.$$

Insgesamt liegt das folgende Paar von Ableitungsgleichungen vor:

$$\boxed{\begin{aligned} \vec{t}\,' &= \kappa\vec{n}, \\ \vec{n}\,' &= -\kappa\vec{t}. \end{aligned}} \tag{2.5}$$

Die Gleichungen (2.5) heißen die FRENETschen Formeln der ebenen Kurventheorie. Sie bilden die Grundlage der gesamten Theorie ebener Kurven. □

(II) Kurven in $\mathbf{R}^3$ (Raumkurven)

Sei $\Gamma \subset \mathbf{R}^3$ eine *räumliche* Kurve mit einer regulären Parameterdarstellung $\vec{x} = \vec{x}(t)$. Dann ist in jedem Kurvenpunkt der Tangenteneinheitsvektor $\vec{t} = \dot{\vec{x}}(t)/\|\dot{\vec{x}}(t)\|$ eindeutig erklärt. Da eine Drehung von $\vec{t}$ um $+\frac{\pi}{2}$ im Vektorraum $\mathbf{R}^3$ unendlich vieldeutig ist, muss die Normale $\vec{n}$ hier anders als im ebenen Fall erklärt werden. Nachfolgend sei $\vec{x}(t)$ hinreichend oft differenzierbar vorausgesetzt. Anders als im $\mathbf{R}^2$ definieren wir nun die Krümmung einer räumlichen Kurve stets als nichtnegative Größe:

Definition 12.10 *Für reguläre räumliche Kurven Γ mit einer Parameterdarstellung $\vec{x} \in C^2(I; \mathbf{R}^3)$ heiße die Länge $\|\vec{t}\,'\| =: \kappa \geq 0$ des Krümmungsvektors die* **Krümmung** *der Kurve Γ. Gilt $\kappa > 0$, so heiße der Einheitsvektor in Richtung des Krümmungsvektors*

$$\boxed{\vec{n} := \frac{1}{\kappa}\,\vec{t}\,' = \frac{1}{\kappa}\,\vec{x}\,''(s)}$$

die **Hauptnormale** *von Γ. Der auf $\vec{t}$ und $\vec{n}$ senkrecht stehende Einheitsvektor*

$$\boxed{\vec{b} := \vec{t} \times \vec{n}}$$

heiße die **Binormale** *von Γ. Für $\kappa = 0$ sind $\vec{n}$ und $\vec{b}$ nicht erklärt. Das in jedem Kurvenpunkt $\vec{x}(t)$ mit $\kappa > 0$ existierende Vektortripel $\vec{t}, \vec{n}, \vec{b}$ heiße* **begleitendes Dreibein** *der Kurve Γ; es bildet eine ON–Basis des Vektorraums $\mathbf{R}^3$.*

Bemerkung 12.6 (a) Wir haben hier wie im ebenen Fall die Beziehung

$$\boxed{\vec{t}\,' = \kappa \vec{n}.} \tag{2.6}$$

(b) Die Einheitsvektoren $\vec{t}, \vec{n}, \vec{b}$ spannen paarweise je eine Ebene in $\mathbf{R}^3$ auf, und zwar

- $\vec{t}$ und $\vec{n}$ die **Schmiegebene** Σ mit der HNF $\langle \vec{y} - \vec{x}(t), \vec{b} \rangle = 0$
- $\vec{n}$ und $\vec{b}$ die **Normalebene** N mit der HNF $\langle \vec{y} - \vec{x}(t), \vec{t} \rangle = 0$
- $\vec{b}$ und $\vec{t}$ die **Streckebene** P mit der HNF $\langle \vec{y} - \vec{x}(t), \vec{n} \rangle = 0$

Ist $\vec{x}(t)$ die Bewegung eines Massepunktes auf der Bahn Γ und t der physikalische Zeitparameter, so hatten wir bereits in Abschnitt 7.3 festgestellt, dass der Beschleunigungsvektor $\ddot{\vec{x}}(t)$ in der Schmiegebene Σ liegt. □

BSP. (12.2.10) Wir betrachten die gemeine **Schraubenlinie** $\vec{x}(t) := (a\cos t, a\sin t, bt)^T,\ t \geq 0$, worin $a, b > 0$ fest gewählt sind. Es gelten hier

$$\dot{\vec{x}}(t) = (-a\sin t, a\cos t, b)^T, \quad \|\dot{\vec{x}}(t)\| = \sqrt{a^2 + b^2},$$

und somit

$$\vec{t} = \frac{1}{\sqrt{a^2+b^2}} \begin{bmatrix} -a\sin t \\ a\cos t \\ b \end{bmatrix}, \quad \vec{t}\,' = \frac{1}{\|\dot{\vec{x}}(t)\|} \frac{d}{dt}\vec{t} = \frac{1}{a^2+b^2} \begin{bmatrix} -a\cos t \\ -a\sin t \\ 0 \end{bmatrix}, \quad \kappa = \frac{a}{a^2+b^2} = \|\vec{t}\,'\| \neq 0.$$

Die Vektoren $\vec{n}$ und $\vec{b}$ sind also erklärt:

$$\boxed{\vec{n} = \begin{bmatrix} -\cos t \\ -\sin t \\ 0 \end{bmatrix}, \quad \vec{b} = \vec{t} \times \vec{n} = \frac{1}{\sqrt{a^2+b^2}} \begin{bmatrix} b\sin t \\ -b\cos t \\ a \end{bmatrix}.}$$

Für eine reguläre räumliche Kurve $\Gamma \subset \mathbf{R}^3$ mit der Parameterdarstellung $\vec{x} = \vec{x}(t) \in C^3(I; \mathbf{R}^3)$ können die Vektoren $\vec{t}\,', \vec{n}\,'$ und $\vec{b}\,'$ in jedem Punkt $\vec{x}(t)$ mit $\kappa > 0$ berechnet werden. Da in jedem solchen Punkt das begleitende Dreibein $\vec{t}, \vec{n}, \vec{b}$ eine ON–Basis des $\mathbf{R}^3$ bildet, müssen sich $\vec{t}\,', \vec{n}\,'$ und $\vec{b}\,'$ in dieser Basis aufspannen lassen. Gleichung (2.6) belegt diese Tatsache bereits für den Vektor $\vec{t}\,'$. Darüber hinaus gilt:

Satz 12.3 *Unter den oben genannten Voraussetzungen sei* $\tau := \langle \vec{n}\,', \vec{b} \rangle$ *gesetzt. Dann gelten die drei folgenden Ableitungsformel von* FRENET (1847)*:*

$$\boxed{\begin{array}{lcccl} \vec{t}\,' & = & & \kappa\vec{n}, & \\ \vec{n}\,' & = & -\kappa\vec{t} & & +\tau\vec{b}, \\ \vec{b}\,' & = & & -\tau\vec{n}. & \end{array}} \tag{2.7}$$

Die Zahl $\tau = \langle \vec{n}\,', \vec{b} \rangle$*, die in jedem Punkt* $\vec{x}(t)$ *mit* $\kappa > 0$ *definiert ist, heiße die* **Windung** *oder* **Torsion** *der Kurve* Γ*.*

Begründung: (i) Es gilt die FOURIER–Entwicklung

$$\vec{b}\,' = \langle \vec{b}\,', \vec{t}\rangle \vec{t} + \langle \vec{b}\,', \vec{n}\rangle \vec{n} + \langle \vec{b}\,', \vec{b}\rangle \vec{b}.$$

Aus $1 = \|\vec{b}\|^2$ ergibt sich durch Differentiation $0 = \frac{d}{ds}\|\vec{b}\|^2 = 2\langle \vec{b}\,', \vec{b}\rangle$. Ferner folgt aus $\vec{t} \perp \vec{b} \perp \vec{n}$ und aus (2.6):

$$0 = \langle \vec{b}, \vec{t}\,\rangle \;\Rightarrow\; 0 = \langle \vec{b}\,', \vec{t}\,\rangle + \langle \vec{b}, \vec{t}\,'\rangle = \langle \vec{b}\,', \vec{t}\,\rangle + \kappa \langle \vec{b}, \vec{n}\rangle = \langle \vec{b}\,', \vec{t}\,\rangle.$$

In gleicher Weise erhalten wir aus $\vec{b} \perp \vec{n}$:

$$0 = \langle \vec{b}, \vec{n}\rangle \;\Rightarrow\; 0 = \langle \vec{b}\,', \vec{n}\rangle + \langle \vec{b}, \vec{n}\,'\rangle = \langle \vec{b}\,', \vec{n}\rangle + \tau.$$

Hieraus folgt die behauptete Relation $\vec{b}\,' = -\tau \vec{n}$.

(ii) Die zweite FRENETsche Formel leitet man unter Berücksichtigung von $0 = \langle \vec{n}\,', \vec{n}\rangle$ und der folgenden Relation her:

$$0 = \langle \vec{n}, \vec{t}\,\rangle \;\Rightarrow\; 0 = \langle \vec{n}\,', \vec{t}\,\rangle + \langle \vec{n}, \vec{t}\,'\rangle = \langle \vec{n}\,', \vec{t}\,\rangle + \kappa.$$

BSP. (12.2.11) Im Falle der gemeinen **Schraubenlinie** aus BSP. (12.2.10) erhalten wir

$$\boxed{\vec{n}\,' = \frac{1}{\|\dot{\vec{x}}(t)\|}\frac{d}{dt}\vec{n} = \frac{1}{\sqrt{a^2+b^2}}\begin{bmatrix} \sin t \\ -\cos t \\ 0 \end{bmatrix}, \quad \tau = \langle \vec{n}\,', \vec{b}\rangle = \frac{b}{a^2+b^2}.}$$

Wir verifizieren mit diesen Größen unter Berücksichtigung der Ergebnisse aus BSP. (12.2.10) die FRENETschen Formeln (2.7):

$$\begin{aligned} \vec{b}\,' &= \frac{1}{\|\dot{\vec{x}}(t)\|}\frac{d}{dt}\vec{b} = \frac{b}{a^2+b^2}\begin{bmatrix} \cos t \\ \sin t \\ 0 \end{bmatrix} = -\tau\vec{n}, \\ -\kappa\vec{t} + \tau\vec{b} &= \frac{1}{(a^2+b^2)^{3/2}}\begin{bmatrix} +a^2\sin t + b^2 \sin t \\ -a^2\cos t - b^2\cos t \\ -ab + ab \end{bmatrix} = \frac{1}{\sqrt{a^2+b^2}}\begin{bmatrix} \sin t \\ -\cos t \\ 0 \end{bmatrix} = \vec{n}\,'. \end{aligned}$$

BSP. (12.2.12) Aus der dritten FRENETschen Formel folgt, dass $\tau = 0$ genau für $\vec{b}\,' = \vec{0}$ gilt, also für $\vec{b} = \vec{b}_0 = const.$ Wegen $0 = \langle \vec{b}, \vec{t}\,\rangle = \langle \vec{b}_0, \vec{x}\,'(s)\rangle$ tritt dieser Fall genau dann ein, wenn $\langle \vec{b}_0, \vec{x}(t)\rangle = d = const$ gilt, wenn also die räumliche Kurve $\vec{x}(t)$ ganz in einer Ebene $E \perp \vec{b}_0$ liegt.

Wir erschließen hieraus:

Folgerung 12.4 (a) *Die Torsion $\tau = \langle \vec{n}\,', \vec{b}\rangle$ ist ein Maß für die Abweichung einer räumlichen Kurve $\vec{x} = \vec{x}(t)$ vom ebenen Verlauf. Eine glatte räumliche Kurve ist genau dann eben, wenn ihre Torsion überall verschwindet.*

(b) *Die Krümmung $\kappa = \|\vec{t}\,'\|$ ist ein Maß für die Abweichung einer räumlichen Kurve $\vec{x} = \vec{x}(t)$ von der geraden Richtung der Tangente $\vec{t}$. Eine glatte räumliche Kurve ist genau dann eine Gerade, wenn ihre Krümmung überall verschwindet.*

Bemerkung 12.7 (a) Die wiederholt verwendete Ableitungsregel $\frac{d}{dt}\|\vec{x}(t)\|^2 = 2\langle \dot{\vec{x}}(t), \vec{x}(t)\rangle = 2\|\vec{x}(t)\|\frac{d}{dt}\|\vec{x}(t)\|$ führt auf

$$\frac{d}{dt}\|\vec{x}(t)\| = \frac{1}{\|\vec{x}(t)\|}\langle \dot{\vec{x}}(t), \vec{x}(t)\rangle. \tag{2.8}$$

(b) Es sei $\vec{x} = \vec{x}(t)$ die Parameterdarstellung einer regulären Raumkurve mit $\vec{x} \in C^3(I; \mathbf{R}^3)$. Dann gilt zunächst $\vec{x}\,'(s) = \dot{\vec{x}}(t)/\|\dot{\vec{x}}(t)\|$. Differenziert man mit Hilfe der Regel (1.7) nochmals nach s, so erhält man mittels (2.8) unter Fortlassung der Argumente:

$$\vec{x}\,'' = \frac{1}{\|\dot{\vec{x}}\|}\,\frac{d}{dt}\left(\frac{\dot{\vec{x}}}{\|\dot{\vec{x}}\|}\right) \overset{(2.8)}{=} \frac{1}{\|\dot{\vec{x}}\|^3}\left(\|\dot{\vec{x}}\| \cdot \ddot{\vec{x}} - \frac{\dot{\vec{x}}\,\langle \dot{\vec{x}}, \ddot{\vec{x}}\rangle}{\|\dot{\vec{x}}\|}\right) = \frac{\ddot{\vec{x}}}{\|\dot{\vec{x}}\|^2} - \frac{\langle \dot{\vec{x}}, \ddot{\vec{x}}\rangle}{\|\dot{\vec{x}}\|^3}\,\vec{t}\,. \tag{2.9}$$

(c) Nun gilt ja $\vec{x}\,''(s) = \kappa\vec{n}$. Unter Verwendung von (2.9) ergibt sich daraus:

$$\kappa\,(\vec{n} \times \vec{t}\,) = \vec{x}\,'' \times \vec{t}\, = \frac{1}{\|\dot{\vec{x}}\|^3}\,(\ddot{\vec{x}} \times \dot{\vec{x}}) = -\kappa\vec{b}.$$

Werden in der letzten Gleichung die Normen gebildet, so resultiert die folgende Berechnungsformel für die Krümmung κ:

$$\boxed{\kappa = \frac{\|\dot{\vec{x}} \times \ddot{\vec{x}}\|}{\|\dot{\vec{x}}\|^3}.} \tag{2.10}$$

Mit ähnlicher – aber geringfügig aufwendiger – Rechnung findet man auch die folgende Berechnungsvorschrift für die Torsion τ: □

$$\boxed{\tau = \frac{\det(\dot{\vec{x}}, \ddot{\vec{x}}, \dddot{\vec{x}})}{\|\dot{\vec{x}} \times \ddot{\vec{x}}\|^2} = \frac{\det(\dot{\vec{x}}, \ddot{\vec{x}}, \dddot{\vec{x}})}{\kappa^2\|\dot{\vec{x}}\|^6}.} \tag{2.11}$$

BSP. (12.2.13) Man zeige, dass die räumliche Kurve Γ mit der Parameterdarstellung

$$\vec{x}(t) := \begin{bmatrix} 2t^2 + t + 1 \\ t^2 - 2t \\ 2t^2 - 3 \end{bmatrix}, \quad t \geq 0,$$

eben ist, und man gebe die Gleichung derjenigen Ebene $E \subset \mathbf{R}^3$ an, in der Γ verläuft.
Zur *Lösung* dieser Aufgabe berechnen wir zunächst die Torsion τ von Γ gemäß der Vorschrift (2.11):

$$\dot{\vec{x}}(t) = \begin{bmatrix} 4t + 1 \\ 2t - 2 \\ 4t \end{bmatrix}, \quad \ddot{\vec{x}}(t) = \begin{bmatrix} 4 \\ 2 \\ 4 \end{bmatrix}, \quad \dddot{\vec{x}}(t) = \vec{0}, \quad \det(\dot{\vec{x}}, \ddot{\vec{x}}, \dddot{\vec{x}}) = \det(\dot{\vec{x}}, \ddot{\vec{x}}, \vec{0}) = 0.$$

Es folgt $\tau = 0$ für alle $t \geq 0$. Also ist die Kurve Γ eben, und sie muss in der Ebene $E \perp \vec{b}$ liegen, wie in BSP. (12.2.12) gezeigt wurde. Wir bestimmen drei Punkte auf Γ durch Wahl von $t = 0, 1, 2$:

$$\vec{p} := \vec{x}(0) = \begin{bmatrix} 1 \\ 0 \\ -3 \end{bmatrix}, \quad \vec{x}_1 := \vec{x}(1) = \begin{bmatrix} 4 \\ -1 \\ -1 \end{bmatrix}, \quad \vec{x}_2 := \vec{x}(2) = \begin{bmatrix} 11 \\ 0 \\ 5 \end{bmatrix}.$$

Nun gelten $\vec{p}$, $\vec{x}_1$, $\vec{x}_2 \in E$. Setzen wir also

$$\vec{u} := \vec{x}_1 - \vec{p} = \begin{bmatrix} 3 \\ -1 \\ 2 \end{bmatrix}, \quad \vec{v} := \vec{x}_2 - \vec{p} = \begin{bmatrix} 10 \\ 0 \\ 8 \end{bmatrix},$$

so erhalten wir, da die Vektoren $\vec{u}$ und $\vec{v}$ offenbar linear unabhängig sind, die folgende Parameterdarstellung der gesuchten Ebene E:

$$E = \{\vec{y} \in \mathbf{R}^3 \;:\; \vec{y} = \vec{p} + \lambda\vec{u} + \mu\vec{v},\;\; \lambda, \mu \in \mathbf{R}\}.$$

Da der Vektor $\vec{v} \times \vec{u} = (8, 4, -10)^T$ in Richtung der Normalen von E weist, erhalten wir gemäß BSP. (12.2.12) durch Normierung die Binormale $\vec{b} = \frac{1}{3\sqrt{5}}(4, 2, -5)^T$. Es resultiert die HESSEsche Normalform von E,

$$E = \Big\{ \vec{y} \in \mathbf{R}^3 \ : \ \langle \vec{y}, \vec{b} \rangle = \langle \vec{p}, \vec{b} \rangle = \frac{19}{3\sqrt{5}} \Big\}.$$

Das heißt, E hat den Abstand $d(\vec{0}, E) = \frac{19}{3\sqrt{5}}$ vom Ursprung, und in der allgemeinen Form gestattet E die Darstellung

$$E: \ 4x + 2y - 5z = 19.$$

12.3 Ergänzungen

Ist $\vec{x} = \vec{x}(s)$ die *natürliche* Parametrisierung einer regulären ebenen Kurve $\Gamma \subset \mathbf{R}^2$, und gilt $\vec{x} \in C^2(I; \mathbf{R}^2)$, so resultiert aus den Beziehungen

$$\vec{t} = \vec{x}\,'(s) = \begin{bmatrix} x'(s) \\ y'(s) \end{bmatrix}, \quad \vec{t}\,' = \vec{x}\,''(s) = \begin{bmatrix} x''(s) \\ y''(s) \end{bmatrix}, \quad \vec{n} = \begin{bmatrix} -y'(s) \\ x'(s) \end{bmatrix}$$

für die Krümmung $\kappa = \kappa(s)$ die Darstellung

$$\boxed{\kappa(s) = \langle \vec{t}\,', \vec{n} \rangle = x'(s)y''(s) - y'(s)x''(s).} \tag{3.1}$$

Der Tangenteneinheitsvektor $\vec{t} = \vec{t}(s)$ kann offenbar in der Form

$$\vec{t}(s) = \begin{bmatrix} \cos\alpha(s) \\ \sin\alpha(s) \end{bmatrix} = \begin{bmatrix} x'(s) \\ y'(s) \end{bmatrix}, \quad \alpha(s) := \operatorname{arc\,tan}_H \frac{y'(s)}{x'(s)}, \tag{3.2}$$

geschrieben werden. Wir erhalten durch Differentiation

$$\frac{d\alpha}{ds} = \frac{1}{1 + (y'/x')^2} \, \frac{x'y'' - y'x''}{x'^2} = x'y'' - y'x''.$$

Gemäß (3.1) gilt also die Beziehung

$$\boxed{\frac{d\alpha}{ds} = \kappa(s).} \tag{3.3}$$

Man leitet aus dieser Gleichung ab, dass eine reguläre ebene Kurve bereits durch die Vorgabe ihrer Krümmung als Funktion des Bogenlängen–Parameters s bis auf ihre Lage in der Ebene eindeutig festgelegt ist:

Satz 12.4 *Die Krümmung $\kappa = \kappa(s)$ sei als stetige Funktion des Bogenlängen–Parameters s vorgegeben. Dann kann die zugeordnete ebene Kurve $\vec{x} = \vec{x}(s) \in C^2(I; \mathbf{R}^2)$ aus κ und den Anfangsvorgaben $\vec{x}_0 := \vec{x}(0)$, $\vec{t}_0 := \vec{x}\,'(0)$ eindeutig rekonstruiert werden.*

Begründung: Es seien

$$\vec{t}_0 = \begin{bmatrix} x_0' \\ y_0' \end{bmatrix} := \begin{bmatrix} x'(0) \\ y'(0) \end{bmatrix}, \quad \alpha_0 := \operatorname{arc\,tan}_H \frac{y_0'}{x_0'}$$

gesetzt. Durch Integration von Gleichung (3.3) resultiert

$$\alpha(s) = \alpha_0 + \int_0^s \kappa(t)\, dt.$$

Wird dies in (3.2) eingesetzt, so führt abermalige Integration mit $\vec{x}_0 = (x_0, y_0)^T$ auf die Relationen

$$x(s) = x_0 + \int_0^s \cos\alpha(t)\, dt, \quad y(s) = y_0 + \int_0^s \sin\alpha(t)\, dt,$$

die bereits eine natürliche Parameterdarstellung der gesuchten Kurve liefern. □

BSP. (12.3.1) Es ist diejenige ebene Kurve Γ zu bestimmen, die zu gegebener Krümmung $\kappa = \kappa(s) := \frac{a}{a^2+s^2}$, $a \neq 0$, die folgenden Anfangsbedingungen erfüllt:

$$\vec{x}(0) = \vec{x}_0 := (0, a)^T, \quad \vec{x}\,'(0) = \vec{t}_0 = (1, 0)^T.$$

Lösung: Wegen $\alpha_0 = 0$ erhält man durch eine erste Integration

$$\alpha(s) = \int_0^s \frac{a}{a^2 + t^2}\, dt = \operatorname{arc\,tan}_H \frac{s}{a}, \quad s = a \tan\alpha.$$

Somit gelten

$$x'(s) = \cos\alpha = \cos\left(\operatorname{arc\,tan}_H \frac{s}{a}\right) = \frac{a}{\sqrt{a^2+s^2}}, \quad y'(s) = \sin\alpha = \frac{s}{\sqrt{a^2+s^2}}.$$

Integriert man diese beiden Gleichungen unter Berücksichtigung der Anfangsbedingung, so erhält man

$$x(s) = \int_0^s \frac{a\, dt}{\sqrt{a^2+t^2}} = a \operatorname{Ar\,sinh} \frac{s}{a}, \quad y(s) = a + \int_0^s \frac{t\, dt}{\sqrt{a^2+t^2}} = \sqrt{a^2+s^2}.$$

Die erste der zwei Gleichungen führt auf $s = a \sinh\frac{x}{a}$, und aus der zweiten Gleichung folgt

$$y(x) = a\sqrt{1 + \sinh^2\frac{x}{a}} = a\, \cosh\frac{x}{a}.$$

Das heißt, die gesuchte ebene Kurve Γ ist eine **Kettenlinie**.

Ein Analogon zu Satz 12.4 gilt auch für räumliche Kurven $\Gamma \subset \mathbf{R}^3$. Dort reichen die Kenntnis von Krümmung $\kappa(s)$ und Torsion $\tau(s)$ aus, um eine Raumkurve bis auf Bewegungen eindeutig festzulegen. Wir zitieren hier:

Satz 12.5 *Die Krümmung $\kappa = \kappa(s)$ und die Torsion $\tau = \tau(s)$ seien als stetige Funktionen des Bogenlängen–Parameters s vorgegeben. Dann kann die zugeordnete räumliche Kurve $\vec{x} = \vec{x}(s) \in C^3(I; \mathbf{R}^3)$ aus κ, τ und den Anfangsvorgaben $\vec{x}_0 := \vec{x}(0)$, $\vec{t}_0 := \vec{x}\,'(0)$ und $\vec{n}_0 \perp \vec{t}_0$ eindeutig rekonstruiert werden.*

Der Beweis ist nicht elementar. Er verwendet Sätze aus der Theorie nichtlinearer Differentialgleichungen. Wir verweisen auf die Literatur am Ende dieses Abschnitts.

BSP. (12.3.2) Wir wollen diskutieren, welche geometrische Figur $\Gamma \subset \mathbf{R}^3$ durch die Parameterkurve

$$\vec{x}(t) := 2\left(1 - t^2, t\sqrt{1-t^2}, \frac{1}{4}\operatorname{arc\,sin}_H t\right)^T, \quad -1 < t < 1,$$

dargestellt wird. Dazu berechnen wir ihre Krümmung und ihre Torsion. Aus den Relationen

$$\dot{\vec{x}}(t) = 2\begin{bmatrix} -2t \\ (1-2t^2)/\sqrt{1-t^2} \\ 1/4\sqrt{1-t^2} \end{bmatrix}, \quad \|\dot{\vec{x}}(t)\| = \frac{\sqrt{17}}{2\sqrt{1-t^2}}, \quad \vec{t} = \frac{4}{\sqrt{17}}\begin{bmatrix} -2t\sqrt{1-t^2} \\ 1-2t^2 \\ 1/4 \end{bmatrix},$$

resultieren der Krümmungsvektor und die Krümmung gemäß

$$\boxed{\vec{t}\,' = \frac{8}{17}\begin{bmatrix} -2(1-2t^2) \\ -4t\sqrt{1-t^2} \\ 0 \end{bmatrix}, \quad \kappa = \|\vec{t}\,'\| = \frac{16}{17} = const.}$$

Wir bestimmen die Normale und die Binormale gemäß

$$\vec{n} = \frac{1}{\kappa}\vec{t}\,' = \begin{bmatrix} -(1-2t^2) \\ -2t\sqrt{1-t^2} \\ 0 \end{bmatrix}, \quad \vec{b} = \vec{t}\times\vec{n} = \frac{1}{\sqrt{17}}\begin{bmatrix} 2t\sqrt{1-t^2} \\ -(1-2t^2) \\ 4 \end{bmatrix}, \quad \vec{b}\,' = \frac{2}{17}\begin{bmatrix} 2(1-2t^2) \\ 4t\sqrt{1-t^2} \\ 0 \end{bmatrix}.$$

Wir erhalten die Torsion

$$\boxed{\tau = -\langle \vec{b}\,', \vec{n}\rangle = \frac{4}{17} = const.}$$

Wir wir bereits in den Beispielen (12.2.10) und (12.2.11) gesehen haben, hat die gemeine **Schraubenlinie** ebenfalls eine konstante Krümmung und eine konstante Torsion. Wir können deshalb aus Satz 12.5 folgern, dass Γ Teil einer Schraubenlinie sein muss. In der Tat, durch die Parametertransformation $t = \sin\xi,\ -\frac{\pi}{2} < \xi < \frac{\pi}{2}$, gewinnen wir die Darstellung

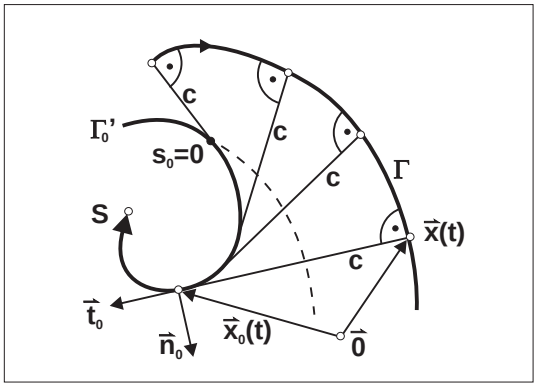
$$\vec{x}(t(\xi)) = \begin{bmatrix} 2\cos^2\xi \\ 2\sin\xi\cos\xi \\ \frac{1}{2}\xi \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \cos 2\xi \\ \sin 2\xi \\ \frac{1}{4}2\xi \end{bmatrix}, \quad -\frac{\pi}{2} < \xi < \frac{\pi}{2}.$$

Dies ist ein Gang der Schraubenlinie um die Achse $(1,0,z)^T$.

Wir diskutieren abschließend ein weiteres Umkehrproblem der Differentialgeometrie. In Abschnitt 12.2 wurde gezeigt, dass der geometrische Ort der Krümmungsmittelpunkte

$$\boxed{\vec{m}(t) = \vec{x}(t) + \frac{1}{\kappa(t)}\vec{n}(t)} \tag{3.4}$$

einer regulären ebenen Kurve $\Gamma \subset \mathbf{R}^2$ mit der Parameterdarstellung $\vec{x}(t)$ und mit nichtverschwindender Krümmung $\kappa(t) \neq 0$ die Evolute von Γ beschreibt. Wir fragen uns, ob aus der Vorgabe der Evolute $\vec{m}(t)$ die ebene Kurve Γ selbst rekonstruiert werden kann. Dazu stellen wir eine Vorüberlegung an.



Die Fadenkonstruktion der Evolvente

Um eine gegebene ebene Kurve Γ_0 werde ein im Punkt $S \in \Gamma_0$ befestigter Faden gelegt, der stets in Richtung der Tangente $\vec{t}_0$ gespannt sei. Das Fadenende beschreibt dann, wenn der Faden auf der Kurve Γ_0 abgewickelt wird, eine ebene Kurve Γ.

Definition 12.11 *Die so konstruierte Kurve* Γ *heiße die* **Filarevolvente** *der gegebenen Kurve* Γ_0 (*lat.* filum = *Faden*).

Nach obiger Skizze gilt für den Ortsvektor $\vec{x}(t)$, der die Kurve Γ beschreibt, die Relation

$$\boxed{\vec{x}(t) = \vec{x}_0(t) - (c + s_0(t))\,\vec{t}_0.} \tag{3.5}$$

Dabei liegt der Ortsvektor $\vec{x}_0(t)$ auf der vorgegebenen Kurve Γ. Nach dieser Vorbetrachtung können wir das oben angesprochene Umkehrproblem in der folgenden Weise lösen:

Satz 12.6 *Die Evolute der Filarevolvente* Γ *ist die ebene Kurve* Γ_0 *selbst. Das heißt, die Kurve* Γ *kann aus ihrer Evolute durch die Vorschrift* (3.5) *rekonstruiert werden, wenn ein einziger Punkt von* Γ *bekannt ist. Mit diesem Punkt wird der Parameter* c *in* (3.5) *festgelegt.*

Begründung: Es darf ohne Beschränkung der Allgemeinheit angenommen werden, dass t ein natürlicher Parameter der Ausgangskurve Γ_0 ist. Sei also $\vec{x}_0(t)$ deren natürliche Parameterdarstellung. Anstelle von (3.5) gilt nun

$$\vec{x}(t) = \vec{x}_0(t) - (c + t)\,\vec{t}_0(t), \tag{3.6}$$

und hieraus gewinnt man durch Differentiation

$$\dot{\vec{x}}(t) = \vec{x}\,'_0(t) - (c+t)\vec{t}\,'_0 - \vec{t}_0 = -(c+t)\kappa_0\vec{n}_0.$$

Es folgen $\|\dot{\vec{x}}(t)\| = |(c+t)\kappa_0|$ sowie $\vec{t} = \dot{\vec{x}}(t)/\|\dot{\vec{x}}(t)\| = -\operatorname{sign}\left((c+t)\kappa_0\right)\vec{n}_0$. Unter Verwendung der FRENETschen Formeln (2.5) ergibt sich nun

$$\frac{d}{dt}\,\vec{t} = -\operatorname{sign}\left((c+t)\kappa_0\right)\vec{n}\,'_0 \overset{(2.5)}{=} \operatorname{sign}(c+t)|\kappa_0|\vec{t}_0\,, \quad \vec{t} = \frac{\frac{d}{dt}\,\vec{t}}{\|\dot{\vec{x}}(t)\|} = \frac{1}{c+t}\,\vec{t}_0 \overset{!}{=} \kappa\vec{n}.$$

Also gilt

$$|\kappa|^2 = \kappa^2 = \frac{1}{(c+t)^2}.$$

Gemäß (3.4) hat die Filarevolvente Γ die Evolute

$$\vec{m}(t) = \vec{x}(t) + \frac{1}{\kappa}\,\vec{n} = \vec{x}(t) + \frac{1}{\kappa^2}\,\kappa\vec{n} = \vec{x}(t) + (c+t)^2\frac{1}{c+t}\,\vec{t}_0 = \vec{x}(t) + (c+t)\,\vec{t}_0 = \vec{x}_0(t).$$

Das heißt, die Evolute von Γ ist, wie behauptet, die Kurve Γ_0. □

BSP. (12.3.3) Zu bestimmen sind die Evolventen der **Zykloide**

$$\vec{x}_0(t) = a(t - \sin t, 1 - \cos t)^T, \quad 0 < t < 2\pi,$$

gemäß der Formel (3.5). Es gelten hier

$$\dot{\vec{x}}_0(t) = a\begin{bmatrix} 1-\cos t \\ \sin t \end{bmatrix}, \quad \|\dot{\vec{x}}_0(t)\| = 2a\sin\frac{t}{2}, \quad \vec{t}_0 = \frac{\dot{\vec{x}}_0(t)}{\|\dot{\vec{x}}_0(t)\|} = \frac{1}{2\sin\frac{t}{2}}\begin{bmatrix} 2\sin^2\frac{t}{2} \\ 2\sin\frac{t}{2}\cos\frac{t}{2} \end{bmatrix} = \begin{bmatrix} \sin\frac{t}{2} \\ \cos\frac{t}{2} \end{bmatrix}.$$

Die Bogenlänge auf Γ_0 ergibt sich nun zu

$$ds_0 = \|\dot{\vec{x}}_0(t)\|\,dt = 2a\sin\frac{t}{2}\,dt, \quad s_0(t) = 4a\Big(1 - \cos\frac{t}{2}\Big).$$

Werden diese Größen in (3.5) eingesetzt, so resultieren die Evolventen der Zykloide mit den Parameterdarstellungen

$$\boxed{\vec{x}(t) = a \begin{bmatrix} t - \sin t \\ 1 - \cos t \end{bmatrix} - 4a\Big(c - \cos\frac{t}{2}\Big) \begin{bmatrix} \sin\frac{t}{2} \\ \cos\frac{t}{2} \end{bmatrix} = a \begin{bmatrix} t + \sin t \\ 3 + \cos t \end{bmatrix} - 4ac \begin{bmatrix} \sin\frac{t}{2} \\ \cos\frac{t}{2} \end{bmatrix}.}$$

Für $c = 0$ erhält man wiederum eine Zykloide mit der Parameterdarstellung

$$\vec{x}(t) = a \begin{bmatrix} t + \sin t \\ 3 + \cos t \end{bmatrix}.$$

Für diese berechnen wir zur Probe die Evolute $\vec{m}(t) = \vec{x}(t) + \frac{1}{\kappa(t)}\,\vec{n}(t)$. Es gelten nun

$$\dot{\vec{x}}(t) = a \begin{bmatrix} 1 + \cos t \\ -\sin t \end{bmatrix}, \quad \|\dot{\vec{x}}(t)\| = 2a\Big|\cos\frac{t}{2}\Big|, \quad \vec{t} = \frac{1}{2|\cos\frac{t}{2}|} \begin{bmatrix} 2\cos^2\frac{t}{2} \\ -2\sin\frac{t}{2}\cos\frac{t}{2} \end{bmatrix} = \operatorname{sign}\big[\cos\frac{t}{2}\big] \begin{bmatrix} \cos\frac{t}{2} \\ -\sin\frac{t}{2} \end{bmatrix}.$$

Des weiteren haben wir

$$\vec{n} = \operatorname{sign}\big[\cos\frac{t}{2}\big] \begin{bmatrix} \sin\frac{t}{2} \\ \cos\frac{t}{2} \end{bmatrix}, \quad \vec{t}\,' = \frac{\operatorname{sign}(\cos\frac{t}{2})}{a|\cos\frac{t}{2}|} \begin{bmatrix} -\sin\frac{t}{2} \\ -\cos\frac{t}{2} \end{bmatrix} = -\frac{1}{4a} \begin{bmatrix} \tan\frac{t}{2} \\ 1 \end{bmatrix}.$$

Somit resultieren

$$\kappa = \langle \vec{t}\,', \vec{n} \rangle = -\frac{1}{4a|\cos\frac{t}{2}|},$$

und schließlich

$$\vec{m}(t) = a \begin{bmatrix} t + \sin t \\ 3 + \cos t \end{bmatrix} - 4a\cos\frac{t}{2} \begin{bmatrix} \sin\frac{t}{2} \\ \cos\frac{t}{2} \end{bmatrix} = a \begin{bmatrix} t - \sin t \\ 1 - \cos t \end{bmatrix},$$

wie vorgegeben.

Literatur zur Differentialgeometrie:

- K. Strubecker, Differentialgeometrie I. Sammlung Göschen, Band 1113/1113a.
- K. Habetha, Höhere Mathematik für Ingenieure und Physiker, Band 3. Klett Studienbücher.
- D. Laugwitz, Differentialgeometrie. Teubner Verlag, Stuttgart.

Kapitel 13

Funktionen von mehreren reellen Veränderlichen

13.1 Vorbetrachtungen

Der allgemeine Funktionsbegriff wurde bereits in Abschnitt 6.1 eingeführt, und zwar als Abbildung f von einer Menge X in eine Menge Y mit der Symbolik $f : X \to Y$. Dabei heißen die Mengen

- $X \equiv D(f)$ der **Definitionsbereich** von f
- $f(X) := \{y \in Y \, : \, y = f(x), \, x \in X\}$ der **Bildbereich** oder **Wertebereich** von f.

Nachfolgend untersuchen wir Funktionen, bei denen X und Y Teilmengen euklidischer Vektorräume sind. Hier werden wir uns insbesondere mit dem Fall $X \subset \mathbf{R}^n$, $n > 1$, auseinandersetzen, denn die Theorie der *Funktionen einer reellen Veränderlichen* $X \subset \mathbf{R}$ wurde ja bereits in den vorangegangenen Kapiteln erschöpfend behandelt. Wir beschränken uns zunächst auf die Untersuchung *skalarwertiger Funktionen*, also auf den Fall $Y \subset \mathbf{R}$.

Definition 13.1 *Funktionen $u = f(\vec{x})$ mit $\vec{x} \in D(f) \subset \mathbf{R}^n$ und $u \in \mathbf{R}$ heißen* **reelle Funktionen von** n *(unabhängigen)* **Veränderlichen**. *Dabei bezeichne $\vec{x}$ das geordnete n–Tupel $\vec{x} = (x_1, x_2, \ldots, x_n)$. Wir schreiben auch äquivalent*

$$u = f(x_1, x_2, \ldots, x_n), \quad (x_1, x_2, \ldots, x_n) \in D(f) \subset \mathbf{R}^n,$$

oder $u = f(x, y)$ bzw. $u = f(x, y, z)$, wenn nur wenige unabhängige Veränderliche auftreten.

BSP. (13.1.1) Physikalische Gesetze, bei denen mehrere physikalische Größen miteinander verknüpft werden, sind typische Vertreter von Funktionen mehrerer Veränderlicher, *zum Beispiel* die **Zustandsgleichung** für ideale Gase (das Boyle–Mariotte–Gesetz). Hier werden der *Druck p* eines idealen Gases, die *Stoffmenge n*, die *Temperatur T* und das *Volumen V* gemäß der Relation

$$p = \frac{nRT}{V}, \quad R\text{: ideale Gaskonstante},$$

miteinander verknüpft. Da sich die Stoffmenge n in der Regel nicht ändert, resultiert eine Funktion $p = p(V, T)$, bei der aus physikalischen Gründen $T > 0$ und $V > 0$ gelten muss.

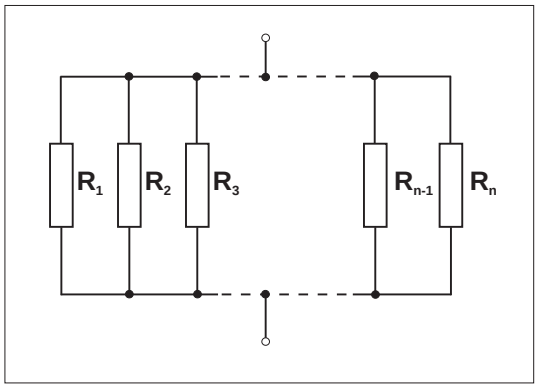
Mit dem Ohmschen Gesetz liegt ein weiteres physikalisches *Beispiel* vor. In einem elektrischen Leiter sind *Stromstärke I*, *Widerstand R* und *Spannung U* durch die folgende Relation verknüpft:

$$U = U(R, I) := R \cdot I.$$

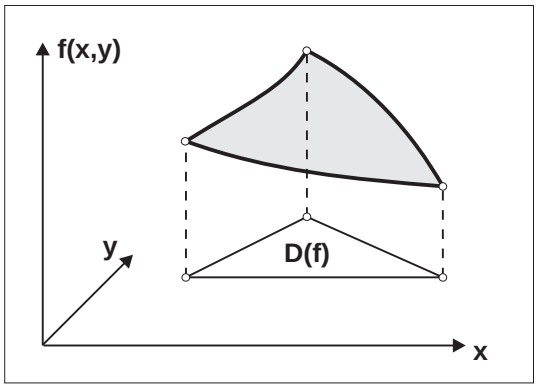
Wir führen als letztes *Beispiel* die **Parallelschaltung** von n Widerständen $R_1, R_2, \ldots, R_n$ an, deren Gasamtwiderstand R sich nach den KIRCHHOFFschen Gesetzen wie folgt ergibt:

$$R = R(R_1, R_2, \ldots, R_n) := \Big(\sum_{j=1}^{n} \frac{1}{R_j} \Big)^{-1}.$$

Auch hier gilt aus physikalischen Gründen stets $R_j > 0$.



Parallelschaltung von n Widerständen



Der Graph der Funktion $u = f(x,y)$ ist eine Fläche in $\mathbf{R}^3$

BSP. (13.1.2) Die Beschreibung von orts– und zeitabhängigen physikalischen Größen führt auf Funktionen mehrerer Veränderlicher. *Zum Beispiel* ist die **Temperatur** T eines wärmeleitenden Mediums bei Abkühlungs– bzw. Aufheizungsprozessen eine Funktion von *Ort* (x, y, z) und *Zeit* t: $T = T(x, y, z, t)$.

Ein weiteres *Beispiel* ist die transversale **Auslenkung** u einer an beiden Enden eingespannten Saite. Die Auslenkung ist eine Funktion des (eindimensionalen) *Ortes* x und der *Zeit* t: $u = u(x, t)$.

Gilt $D(f) \subset \mathbf{R}^2$, so kann die funktionale Beziehung

$$u = f(x, y), \quad (x, y) \in D(f), \tag{1.1}$$

auch *geometrisch* gedeutet werden: Der Graph von f ist eine Teilmenge des dreidimensionalen Anschauungsraumes $\mathbf{R}^3$ und somit unserer Anschauung zugänglich.

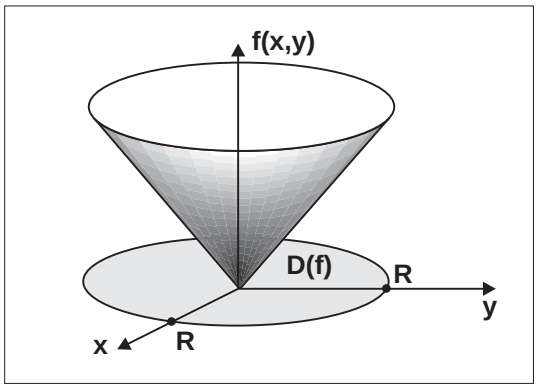
Definition 13.2 *Der Graph* $G(f) := \{(x, y, u) \in \mathbf{R}^3 \,:\, u = f(x, y), \quad (x, y) \in D(f)\}$ *einer Funktion f von zwei unabhängigen Veränderlichen heiße eine* **Fläche** *in* $\mathbf{R}^3$.

BSP. (13.1.3) Die **Kegelfläche** in $\mathbf{R}^3$ mit der Gleichung

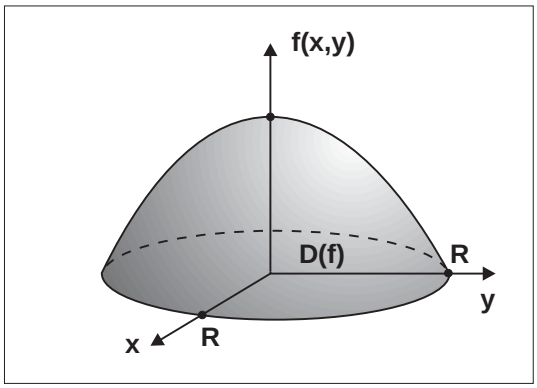
$$\boxed{u = f(x, y) := \sqrt{x^2 + y^2}, \quad (x, y) \in D(f) := \{(x, y) \in \mathbf{R}^2 \,:\, 0 \le x^2 + y^2 \le R^2\}.}$$

Die **Halbsphäre** in $\mathbf{R}^3$ mit der Gleichung

$$\boxed{u = f(x, y) := \sqrt{R^2 - (x^2 + y^2)}, \quad (x, y) \in D(f) := \{(x, y) \in \mathbf{R}^2 \,:\, 0 \le x^2 + y^2 \le R^2\}.}$$



Kegelfläche



Halbsphäre

Eine weitere Darstellungsmöglichkeit der funktionalen Beziehung (1.1) eröffnet sich für $D(f) \subset \mathbf{R}^2$, wenn der Graph $G(f)$ mit Ebenen $u = const$ geschnitten wird. Die entstehenden Schnittlinien werden orthogonal auf die (x, y)–Ebene projiziert: Man erhält ein Höhenlinien–Portrait der Funktion f.

Definition 13.3 *Für $u = f(x, y)$ mit $(x, y) \in D(f) \subset \mathbf{R}^2$ heißen die implizit definierten Kurven*

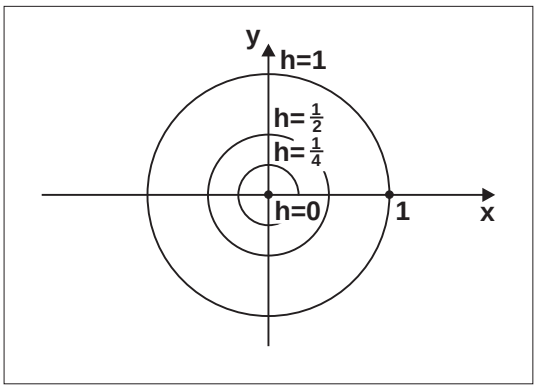
$$\Gamma_h := \{(x, y) \in D(f) \,:\, f(x, y) = h\}, \quad h \in \mathbf{R},$$

die **Höhenlinien** *oder* **Niveaulinien** *oder* **Äquipotentiallinien** *von f. Das graphische Gebilde, das durch Darstellung einer Funktion mittels ihrer Höhenlinien entsteht, heiße* **Karte** *von f.*

BSP. (13.1.4) Die Höhenlinien der **Kegelfläche** $u = f(x, y) := \sqrt{x^2 + y^2}$ sind die Linien

$$\sqrt{x^2 + y^2} = h = const, \quad h \geq 0.$$

Diese bilden eine Schar konzentrischer Kreise vom Radius h um den Mittelpunkt O.



Höhenlinien der Kegelfläche

Funktionen $f : \mathbf{R}^n \to \mathbf{R}$ mit $n > 2$ sind einer Veranschaulichung nicht mehr unmittelbar zugänglich; ihr Graph ist eine Teilmenge des $\mathbf{R}^{n+1}$. Im Fall $n = 3$ definiert man das folgende Analogon zu den Höhenlinien:

Definition 13.4 *Für $u = f(x, y, z)$ mit $(x, y, z) \in D(f) \subset \mathbf{R}^3$ heißen die implizit definierten Flächen*

$$F_h := \{(x, y, z) \in D(f) \,:\, f(x, y, z) = h\}, \quad h \in \mathbf{R},$$

die **Niveauflächen** *oder* **Äquipotentialflächen** *von f.*

Niveauflächen sind nun wieder einer graphischen Darstellung in $\mathbf{R}^3$ zugänglich.

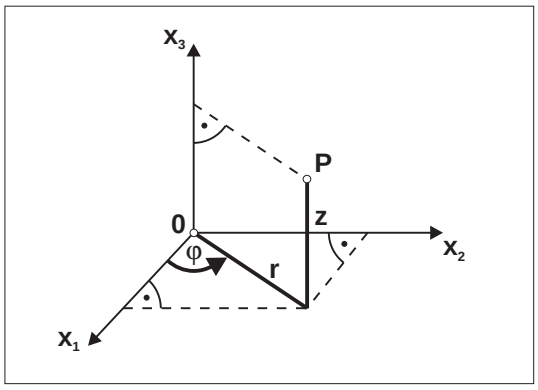
BSP. (13.1.5) Die Niveauflächen der Funktion $f(x,y,z) := x^2 + y^2 - 2z$ sind die **Rotationsparaboloide** $2z + h = x^2 + y^2$, $h = const$. Deren Rotationsachse ist die z–Achse.

In den obigen Definitionen der Niveaulinien bzw. –flächen haben wir bereits apostrophiert, dass diese Gebilde in **impliziter Darstellung** durch Gleichungen vom Typ $f(\vec{x}) = h$ vorliegen. Solche Gleichungen sind im allgemeinen nicht mehr eindeutig **explizit** auflösbar. Die damit verbundene Problematik werden wir in einem der folgenden Abschnitte erörtern, und zwar mit dem Satz über implizite Funktionen.

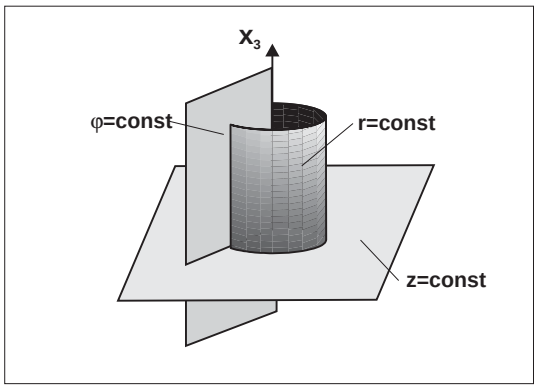
Räumliche Koordinatensysteme

Das kartesische Koordinatensystem in $\mathbf{R}^3$ haben wir bereits zur geometrischen Veranschaulichung von Funktionszusammenhängen benutzt. Für spezielle Probleme ist es oft vorteilhaft, andere Koordinatensysteme zu verwenden. In der Ebene $\mathbf{R}^2$ konnten wir zum Beispiel *Polarkoordinaten* sinnvoll einsetzen. *Räumliche Polarkoordinaten* sind in zwei Varianten bekannt:

(I) Zylinderkoordinaten (r, φ, z): Es sei $(O; x_1, x_2, x_3)$ ein kartesisches Koordinatensystem in $\mathbf{R}^3$. Dann kann die Lage eines Punktes $O \neq P \in \mathbf{R}^3$ auch durch die drei Größen r, φ, z eindeutig beschrieben werden, vgl. folgende Skizze.



Die Zylinderkoordinaten in $\mathbf{R}^3$



Koordinatenflächen $r = const$, $\varphi = const$, $z = const$

Beide Koordinatensysteme stehen in folgender Relation zueinander:

$$\boxed{x_1 = r\cos\varphi, \quad x_2 = r\sin\varphi, \quad x_3 = z, \quad 0 < r,\ 0 \le \varphi < 2\pi}$$

sowie

$$\boxed{r = \sqrt{x_1^2 + x_2^2}, \quad \tan\varphi = \frac{x_2}{x_1}, \quad z = x_3, \quad 0 < r,\ 0 \le \varphi < 2\pi.}$$

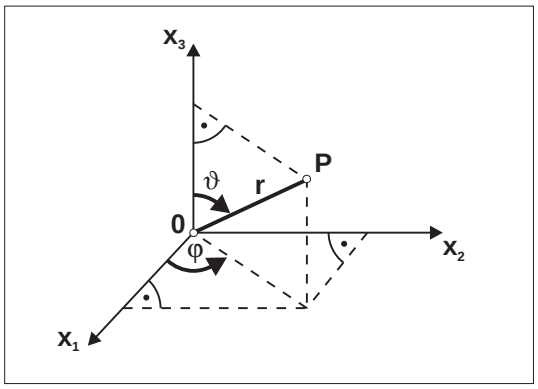
Für $r = 0$ ist φ nicht erklärt; alle Tripel $(0, \varphi, z)$ identifiziert man mit $(0, 0, z)$. Die Koordinatenflächen $r = const,\ \varphi = const,\ z = const$ sind paarweise orthogonal.

BSP. (13.1.6) Durch die Gleichung $f(x_1, x_2, x_3) := x_1^2 + x_2^2 - 2x_3 = 0,\ x_3 \ge 0$, wird ein nach oben geöffnetes **Rotationsparaboloid** beschrieben. In Zylinderkoordinaten gelangen wir zur expliziten Darstellung

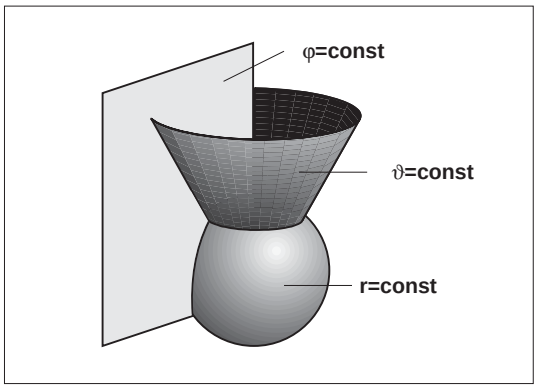
$$\boxed{z(r) = \frac{r^2}{2}, \quad r \ge 0.}$$

Das Fehlen der Veränderlichen φ bedeutet *Rotationssymmetrie*.

(II) Kugelkoordinaten (r, ϑ, φ): Es sei $(O; x_1, x_2, x_3)$ wieder ein kartesisches Koordinatensystem in $\mathbf{R}^3$. Die Lage eines Punktes $O \neq P \in \mathbf{R}^3$ läßt sich nun auch durch die drei Größen r, ϑ, φ eindeutig beschreiben, vgl. folgende Skizze.



Die Kugelkoordinaten in $\mathbf{R}^3$



Koordinatenflächen $r = const$, $\vartheta = const$, $\varphi = const$

Beide Koordinatensysteme stehen in folgender Relation zueinander:

$$x_1 = r\cos\varphi\sin\vartheta, \quad x_2 = r\sin\varphi\sin\vartheta, \quad x_3 = r\cos\vartheta, \quad 0 < r,\ 0 < \vartheta < \pi,\ 0 \leq \varphi < 2\pi$$

sowie

$$r = \sqrt{x_1^2 + x_2^2 + x_3^2}, \quad \cos\vartheta = \frac{x_3}{\sqrt{x_1^2 + x_2^2 + x_3^2}}, \quad \tan\varphi = \frac{x_2}{x_1}.$$

Für $r = 0$ sind ϑ und φ nicht erklärt; alle Tripel $(0, \vartheta, \varphi)$ identifiziert man mit $(0, 0, 0)$. Anderen Punkten der x_3–Achse wird ein Winkel $\vartheta = 0$ oder $\vartheta = \pi$ zugeordnet, während φ unbestimmt bleibt. Die Koordinatenflächen $r = const,\ \vartheta = const,\ \varphi = const$ sind paarweise orthogonal.

BSP. (13.1.7) Das durch die Gleichung $f(x_1, x_2, x_3) := x_1^2 + x_2^2 - 2x_3 = 0,\ x_3 \geq 0$, beschriebene **Rotationsparaboloid** gestattet in Kugelkoordinaten die explizite Darstellung

$$r(\vartheta) = \frac{2\cos\vartheta}{\sin^2\vartheta}, \quad 0 < \vartheta \leq \frac{\pi}{2}.$$

Das Fehlen der Veränderlichen φ bedeutet auch hier *Rotationssymmetrie*.

13.2 Metrische Räume, Stetigkeit

Viele Begriffe der Analysis wie Stetigkeit, Differenzierbarkeit etc. sind untrennbar verbunden mit dem Begriff der Konvergenz von Folgen. Dies hatten wir für Funktionen einer reellen Veränderlichen in den Abschnitten 6.3 – 6.5 und 7.1 begründet. Darüber hinaus hatten wir in Abschnitt 3.1 die Feststellung getroffen, dass ein Konvergenzbegriff auf einer nichtleeren Menge M immer dann erklärt werden kann, wenn M eine **Metrik** trägt, das heißt, wenn eine Abbildung $d(\cdot,\cdot) : M \times M \to \mathbf{R}$ existiert mit den folgenden Eigenschaften

(M1)	$d(x,y) > 0 \quad \Leftrightarrow \quad x \neq y,$	(Definitheit)
(M2)	$d(x,y) = d(y,x)\ \forall\ x, y \in M,$	(Symmetrie)
(M3)	$d(x,y) \leq d(x,z) + d(z,y)\ \forall\ x, y, z \in M,$	(Dreiecksungleichung)

Wir nennen eine nichtleere Menge M mit einer Metrik $d(\cdot,\cdot)$ einen **metrischen Raum**.

BSP. (13.2.1) Beispiele metrischer Räume sind

- der Körper $\mathbf{R}$ der reellen Zahlen mit der folgenden Metrik (vgl. Satz 1.15)

(D1) $$\boxed{d(x,y) := |x-y| \;\; \forall\, x,y \in \mathbf{R}}$$

- der Körper $\mathbf{C}$ der komplexen Zahlen mit der folgenden Metrik (vgl. Bemerkung 2.2)

(D2) $$\boxed{d(z_1,z_2) := |z_1-z_2| = \sqrt{(x_1-x_2)^2+(y_1-y_2)^2} \;\; \forall\, z_j := x_j + iy_j \in \mathbf{C}}$$

- der Vektorraum $\mathbf{R}^n$ mit einer der drei folgenden Metriken (vgl. Bemerkung 4.11)

(D3) $$\boxed{d(\vec{x},\vec{y}) := \|\vec{x}-\vec{y}\| = \Big(\sum_{k=1}^{n} |x_k-y_k|^2\Big)^{1/2} \;\; \forall\, \vec{x},\vec{y} \in \mathbf{R}^n,}$$

(D4) $$\boxed{d^*(\vec{x},\vec{y}) := \max_{1\le k\le n} |x_k-y_k| \;\; \forall\, \vec{x},\vec{y} \in \mathbf{R}^n,}$$

(D5) $$\boxed{d^{**}(\vec{x},\vec{y}) := \sum_{k=1}^{n} |x_k-y_k| \;\; \forall\, \vec{x},\vec{y} \in \mathbf{R}^n.}$$

Man überzeugt sich durch Nachrechnen sehr einfach von der Gültigkeit der Relationen

$$d^*(\vec{x},\vec{y}) \;\le\; d^{**}(\vec{x},\vec{y}) \;\le\; \sqrt{n}\, d(\vec{x},\vec{y}) \;\; \forall\, \vec{x},\vec{y} \in \mathbf{R}^n, \tag{2.1}$$

$$d(\vec{x},\vec{y}) \;\le\; d^{**}(\vec{x},\vec{y}) \;\le\; n\, d^*(\vec{x},\vec{y}) \;\; \forall\, \vec{x},\vec{y} \in \mathbf{R}^n. \tag{2.2}$$

Das Beispiel belegt, dass eine Menge M auch mehrere verschiedene Metriken tragen kann. Man definiert jedoch:

Definition 13.5 *Zwei Metriken $d_1(\cdot,\cdot)$ und $d_2(\cdot,\cdot)$ auf derselben Menge M heißen* **äquivalent**, *wenn es Zahlen $\alpha,\beta > 0$ gibt mit*

$$\boxed{\alpha\, d_1(x,y) \le d_2(x,y) \le \beta\, d_1(x,y) \;\; \forall\, x,y \in M.} \tag{2.3}$$

In diesem Sinne sind die obigen Metriken (D3), (D4) und (D5) auf dem Vektorraum $M := \mathbf{R}^n$ paarweise äquivalent, wie man aus den Ungleichungen (2.1) und (2.2) ersieht.

BSP. (13.2.2) Auf dem Funktionenraum $M := C([a,b];\mathbf{R}^n)$ der über dem abgeschlossenen Intervall $[a,b] \subset \mathbf{R}$ stetigen vektorwertigen Funktionen $\vec{x}(t), \vec{y}(t), \ldots$ sind in der folgenden Weise Metriken erklärt:

(D6) $$\boxed{d(\vec{x},\vec{y}) := \max_{t\in[a,b]} \|\vec{x}(t)-\vec{y}(t)\| \;\; \forall\, \vec{x},\vec{y} \in M,}$$

(D7) $$\boxed{d^*(\vec{x},\vec{y}) := \int_a^b \|\vec{x}(t)-\vec{y}(t)\|\, dt \;\; \forall\, \vec{x},\vec{y} \in M.}$$

Zwar gilt stets $d^*(\vec{x},\vec{y}) \le (b-a) d(\vec{x},\vec{y})$, aber eine Ungleichung in der Form $d(\vec{x},\vec{y}) \le c\, d^*(\vec{x},\vec{y})$ mit einer von $\vec{x},\vec{y} \in M$ unabhängigen Konstanten $c > 0$ gilt im allgemeinen nicht. Deshalb sind die beiden Metriken $d(\cdot,\cdot)$ und $d^*(\cdot,\cdot)$ **nicht** äquivalent.

Ist M ein metrischer Raum, so heißt eine Folge $(a_n)_{n\in\mathbf{N}} \subset M$ gemäß Definition 3.3 **konvergent** zum Grenzwert $a \in M$: $\lim\limits_{n\to\infty} a_n = a$, wenn gilt:

$$\boxed{\forall\, k \in \mathbf{N}\ \exists\, N \in \mathbf{R}\ :\ d(a_n, a) < 10^{-k}\ \forall\, n > N.} \tag{2.4}$$

Bemerkung 13.1 (a) Sind auf der Menge M äquivalente Metriken vorgelegt, so hängt der Konvergenzbegriff **nicht** von der speziellen Wahl der Metrik ab. Dies folgt aus (2.3) und (2.4).

(b) Die Aussage (a) ist im allgemeinen falsch, wenn Metriken **nicht** äquivalent sind. Wir belegen diese Aussage durch das folgende □

BSP. (13.2.3) Der Funktionenraum $M := C([0,1])$ der über $[0,1]$ stetigen Funktionen sei mit den zwei Metriken (D6) und (D7) versehen, wobei wir dort $n = 1$ zu setzen haben. Wir betrachten in M die Funktion $x(t) := 0$ sowie die Folge

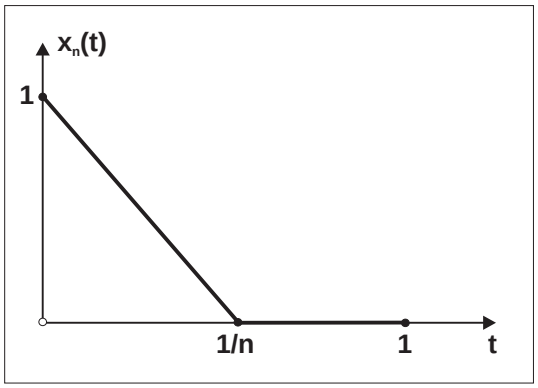
$$x_n(t) := \begin{cases} 1 - nt & :\ 0 \le t \le \frac{1}{n}, \\ 0 & :\ t > \frac{1}{n}, \end{cases} \qquad n \in \mathbf{N}.$$

Man verifiziert sofort die Konvergenzeigenschaft

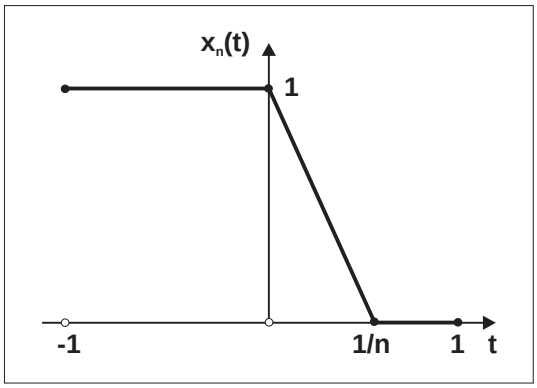
$$d^*(x_n, x) = \int_0^{1/n} (1 - nt)\, dt = \frac{1}{2n} \to 0\ (n \to \infty).$$

Das heißt, es gilt $\lim\limits_{n\to\infty} x_n = x$ in der Metrik $d^*(\cdot,\cdot)$. Hingegen gilt $\lim\limits_{n\to\infty} x_n \neq x$ in der Metrik $d(\cdot,\cdot)$:

$$d(x_n, x) = \max_{t\in[0,1]} |x_n(t) - 0| = 1 \to 1\ (n \to \infty).$$



Die Funktion $x_n(t)$ aus BSP. (13.2.3)



Die Funktion $x_n(t)$ aus BSP. (13.2.4)

Die Konvergenzbedingung (2.4) setzt die Kenntnis des Grenzwertes $a \in M$ voraus. Dieses Faktum ist immer dann von Nachteil, wenn man darauf angewiesen ist, Konvergenz ohne explizite Kenntnis des Grenzwertes nachprüfen zu müssen. Einen Ausweg aus diesem Dilemma haben wir für Zahlenfolgen in dem Konvergenzkriterium von CAUCHY (Satz 3.6) aufgezeigt und dabei den Begriff der CAUCHY**–Folge** geprägt. Gemäß Definition 3.9 heißt eine Folge $(a_n)_{n\in\mathbf{N}} \subset M$ eine CAUCHY–Folge, wenn gilt:

$$\boxed{\forall\, k \in \mathbf{N}\ \exists\, N \in \mathbf{R}\ :\ d(a_n, a_m) < 10^{-k}\ \forall\, n, m > N.} \tag{2.5}$$

Jede *konvergente* Folge $(a_n)_{n\in\mathbf{N}} \subset M$ ist eine CAUCHY–Folge, denn es gilt unter Verwendung der Dreiecksungleichung

$$d(a_n, a_m) \overset{(M3)}{\le} d(a_n, a) + d(a, a_m) \overset{(2.4)}{<} 2 \cdot 10^{-k}\ \forall\, n, m > N.$$

In Satz 3.6 hatten wir gezeigt, dass für Zahlenfolgen auch die Umkehrung dieser Aussage gilt: Jede CAUCHY–Folge in $\mathbf{K} := \mathbf{R}$ oder $\mathbf{K} := \mathbf{C}$ konvergiert gegen einen Grenzwert aus $\mathbf{K}$. In allgemeinen metrischen Räumen ist diese Aussage falsch, wie das folgende Beispiel belegt.

BSP. (13.2.4) Der Funktionenraum $M := C([-1,1])$ der über $[-1,1]$ stetigen Funktionen sei mit der Metrik (D7) versehen (man setze dort $n = 1$). Wir betrachten in M die Funktionenfolge

$$x_n(t) := \begin{cases} 1 & : \ -1 \le t \le 0, \\ 1 - nt & : \ 0 < t \le \frac{1}{n}, \\ 0 & : \ t > \frac{1}{n}, \end{cases} \qquad n \in \mathbf{N}.$$

Dann folgt für $k \in \mathbf{N}$ und $n > m \ge 10^k$

$$d^*(x_n, x_m) = \int_{-1}^{1} |x_n(t) - x_m(t)|\, dt = \int_0^{1/m} (1 - mt)\, dt - \int_0^{1/n} (1 - nt)\, dt = \frac{1}{2}\left(\frac{1}{m} - \frac{1}{n}\right) < 10^{-k}.$$

Also ist $(x_n)_{n\in\mathbf{N}} \subset M$ eine CAUCHY–Folge. Setzen wir

$$x(t) := \begin{cases} 1 \ : \ -1 \le t < 0, \\ 0 \ : \ 0 \le t \le 1, \end{cases}$$

so kann wie in BSP. (13.2.3) die Konvergenz $\lim\limits_{n\to\infty} x_n = x$ in der Metrik $d^*(\cdot,\cdot)$ gezeigt werden. Da $x(t)$ aber unstetig ist, gilt $x \notin M$. Die CAUCHY–Folge $(x_n)_{n\in\mathbf{N}}$ konvergiert **nicht** in M.

Metrische Räume M, in denen jede CAUCHY–Folge ihren Grenzwert in M annimmt, wurden durch die Definition 3.10 ausgezeichnet: Es sind die **vollständigen** metrischen Räume.

Definition 13.6 *Ein metrischer Raum M heiße* **vollständig** *genau dann, wenn* **jede** CAUCHY*-Folge $(a_n)_{n\in\mathbf{N}} \subset M$ ihren Grenzwert a in M annimmt.*

Die Vollständigkeit von M bleibt bei Übergang zu äquivalenten Metriken stets erhalten.

BSP. (13.2.5) Beispiele vollständiger metrischer Räume sind

- der Körper $\mathbf{R}$ der reellen Zahlen mit der Metrik (D1); dies folgt aus Satz 3.6
- der Körper $\mathbf{C}$ der komplexen Zahlen mit der Metrik (D2); dies folgt ebenfalls aus Satz 3.6
- der Vektorraum $\mathbf{R}^n$ mit jeder der Metriken (D3), (D4) oder (D5)
- der Funktionenraum $C([a,b];\mathbf{R}^n)$ mit der Metrik (D6), **nicht aber** mit der Metrik (D7). Die Konvergenz in der Metrik (D6) ist im Falle $n = 1$ genau die gleichmäßige Konvergenz von Funktionenfolgen auf dem Intervall $[a,b]$, man vgl. Definition 9.4. Die Vollständigkeit erhält man nun aus Satz 9.4
- der Funktionenraum $L(a,b)$ der über dem (nicht notwendig beschränkten) Intervall (a,b) LEBESGUE–integrierbaren Funktionen mit der Metrik (D7) ($n = 1$). Das dortige Integral ist als LEBESGUE–Integral zu interpretieren. Eine Begründung für die Vollständigkeit können wir hier nicht geben.

Stetigkeit in metrischen Räumen

Es seien zwei metrische Räume (M_1, d_1) und (M_2, d_2) gegeben sowie eine Abbildung $T \in \mathrm{Abb}\,(M_1, M_2)$. Insbesondere können $M_1 = M_2 =: M$ und $d_1 = d_2 := d$ gelten. Ist $T : D(T) \to \mathbf{K}$ mit $D(T) \subset \mathbf{R}$ eine skalarwertige Funktion einer reellen Veränderlichen, so sind sicher $M_1 := D(T)$ und $M_2 := \mathbf{K}$ zwei solche metrische Räume. Die in Abschnitt 6.5 getroffene Stetigkeitsdefinition 6.17 übertragen wir nun auf den hier vorliegenden allgemeinen Fall.

Definition 13.7 *Eine Abbildung* $T \in \mathrm{Abb}\,(M_1, M_2)$ *heiße* **stetig im Punkte** $x_0 \in M_1$*, wenn für jede Folge* $(x_n)_{n\in\mathbf{N}} \subset M_1$ *mit* $\lim\limits_{n\to\infty} x_n = x_0$ *gilt:*

$$\boxed{\lim_{n\to\infty} Tx_n = Tx_0.} \tag{2.6}$$

Ist T *in* **jedem Punkte** $x_0 \in M_1$ *stetig, so heiße* T **stetig** *(auf* M_1*).*

In Satz 6.8 wurde für $T \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K})$ die Äquivalenz der obigen Definition mit der $\epsilon - \delta$–Definition der Stetigkeit gezeigt. Diese gilt auch im allgemeinen Fall uneingeschränkt:

Satz 13.1 ($\epsilon - \delta$–Definition der Stetigkeit)
Eine Abbildung $T \in \mathrm{Abb}\,(M_1, M_2)$ *ist genau dann im Punkt* $x_0 \in M_1$ *stetig, wenn gilt:*

$$\boxed{\forall\,\epsilon > 0\ \exists\,\delta = \delta(\epsilon, x_0) > 0\ :\ d_2(Tx, Tx_0) < \epsilon\ \forall\, x \in M_1\ \ \textit{mit}\ \ 0 < d_1(x, x_0) < \delta.} \tag{2.7}$$

Begründung: (a) Gelte zunächst die Relation (2.7). Man wähle zu $\epsilon > 0$ ein $\delta = \delta(\epsilon, x_0)$ gemäß der Vorschrift (2.7), wobei ohne Einschränkung $\delta < 1$ angenommen werden darf. Zu jeder Folge $(x_n)_{n\in\mathbf{N}} \subset M_1 \setminus \{x_0\}$ mit $\lim\limits_{n\to\infty} x_n = x_0$ existieren nun Zahlen $k, N \in \mathbf{N}$ mit

$$0 < d_1(x_n, x_0) < \delta \le 10^{-k}\ \ \forall\ n > N.$$

Gemäß (2.7) muss dann $d_2(Tx_n, Tx_0) < \epsilon\ \forall\ n > N$ gelten und somit auch $\lim\limits_{n\to\infty} Tx_n = Tx_0$.

(b) Gelte nun die Relation (2.6). Wäre (2.7) nicht erfüllt, so hätten wir im Gegenteil

$$\exists\,\epsilon_0 > 0\ \forall\, k \in \mathbf{N}\ :\ d_2(Tx_k, Tx_0) \ge \epsilon_0\ \ \text{für ein}\ x_k \in M_1\ \text{mit}\ \ 0 < d_1(x_k, x_0) < \frac{1}{k}.$$

Es wäre somit $(x_k)_{k\in\mathbf{N}} \subset M_1 \setminus \{x_0\}$ eine konvergente Folge mit Grenzwert x_0, aber mit $d_2(Tx_k, Tx_0) \ge \epsilon_0 > 0$, was im Widerspruch zur Konvergenzbedingung (2.6) steht. □

Der Bedingung (2.7) entnehmen wir, dass der allgemeinen Stetigkeitsdefinition wiederum das Prinzip der **Abstandsmessung** durch eine Metrik zugrunde liegt. Bei den Funktionen $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K})$ wurden die Metriken (D1) und (D2) verwendet. Es ist nun klar, dass diejenigen Begriffe, in denen es nur auf Abstandsmessungen ankommt, sofort von dem skalaren Fall $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K})$ auf den allgemeinen Fall $T \in \mathrm{Abb}\,(M_1, M_2)$ übertragen werden können. Dazu müssen nur die Metriken in $\mathbf{R}$ und $\mathbf{K}$ durch die entsprechenden Metriken in M_1 und M_2 ersetzt werden. Insbesondere übertragen sich somit Begriffe wie **gleichmäßige Stetigkeit**, (gleichmäßige) LIPSCHITZ**–Stetigkeit, Kontraktion**, usw. sowie alle daraus abgeleiteten Eigenschaften. Als eine der bedeutendsten Verallgemeinerungen in dieser Richtung gilt das folgende Pendant zum Fixpunktsatz 6.25.

Satz 13.2 (BANACHscher Fixpunktsatz)**
Sei M *ein* **vollständiger** *metrischer Raum mit einer Metrik* $d(\cdot,\cdot)$*. Ist die Abbildung* $T : M \to M$ **kontrahierend***:*

$$\boxed{\exists\, q \in [0, 1)\ :\ d(Tx, Ty) \le q\, d(x, y)\ \ \forall\ x, y \in M,} \tag{2.8}$$

so hat T *in* M *genau einen Fixpunkt* $x = Tx \in M$*. Dieser ist der Grenzwert der Folge der Iterationen*

$$\boxed{x_{n+1} := Tx_n, \quad n \in \mathbf{N}_0, \quad x_0 \in M\ \textit{beliebig},} \tag{2.9}$$

und es gelten für alle $n \in \mathbf{N}$ die Fehlerabschätzungen (FA):

$$\boxed{d(x, x_n) \leq \begin{cases} \dfrac{q^n}{1-q}\, d(x_1, x_0) & : \textbf{a–priori FA}, \\ \dfrac{q}{1-q}\, d(x_n, x_{n-1}) & : \textbf{a–posteriori FA}, \end{cases}} \tag{2.10}$$

mit der Kontraktionskonstanten q aus (2.8).

Eine *Begründung* erhält man durch wörtliche Übertragung des Beweises von Satz 6.25, wenn man dort das Intervall $[a, b]$ durch den metrischen Raum M und die Metrik (D1) durch $d(\cdot, \cdot)$ ersetzt.

Als **Anwendung** des BANACHschen Fixpunktsatzes betrachten wir für feste reelle Zahlen $a < b$ den *Streifen* $M_1 := [a, b] \times \mathbf{R}^n \subset \mathbf{R}^{n+1}$, den wir mit der Metrik (D3) versehen. Wir betrachten ferner $\vec{f} \in \mathrm{Abb}\,(M_1, \mathbf{R}^n)$, also eine Vektorfunktion von $n+1$ reellen Veränderlichen $\vec{f} = \vec{f}(t, \vec{x})$, $t \in [a, b]$, $\vec{x} \in \mathbf{R}^n$. Nach den obigen Ausführungen ist ein Stetigkeitsbegriff für $\vec{f}$ wohldefiniert. Weiterhin sind *Komposita* stetiger Funktionen wieder stetige Funktionen. Das heißt, ist $\vec{x} : [a, b] \to \mathbf{R}^n$ stetig, so folgt aus der Stetigkeit von $\vec{f}$ auch die Stetigkeit der Abbildung

$$\vec{f}(\cdot, \vec{x}(\cdot)) : [a, b] \to \mathbf{R}^n.$$

Somit wird der folgende Satz sinnvoll.

Satz 13.3 *Der Funktionenraum $M := C([a, b]; \mathbf{R}^n)$ sei mit der Metrik* (D6) *versehen. Gegeben seien ferner ein $\vec{u} \in M$ und eine stetige Vektorfunktion $\vec{f} : [a, b] \times \mathbf{R}^n \to \mathbf{R}^n$, die einer* LIPSCHITZ*–Bedingung*

$$\boxed{\exists\, L \geq 0 \;:\; \|\vec{f}(t, \vec{x}) - \vec{f}(t, \vec{y})\| \leq L\, \|\vec{x} - \vec{y}\| \;\; \forall\, t \in [a, b] \;\; \forall\, \vec{x}, \vec{y} \in \mathbf{R}^n} \tag{2.11}$$

genügt. Dann wird durch die Vorschrift

$$\boxed{(T\vec{x})(t) := \vec{u}(t) + \int_{t_0}^{t} \vec{f}(s, \vec{x}(s))\, ds, \quad \vec{x} \in M, \;\; t \in [a, b],} \tag{2.12}$$

eine Abbildung $T : M \to M$ erklärt, die für jedes Teilintervall $[t_0, t_1] \subseteq [a, b]$ mit der Längenbeschränkung $L(t_1 - t_0) < 1$ **kontrahierend** *ist.*

Begründung: Auf Grund der Vorbemerkung ist $T\vec{x}$ stetig. Das heißt, T bildet den metrischen Raum M in sich ab. Um die Kontraktionseigenschaft zu zeigen, seien $\vec{x}, \vec{y} \in M$ und $t \in [t_0, t_1]$ fixiert:

$$\|(T\vec{x})(t) - (T\vec{y})(t)\| \;=\; \Big\| \int_{t_0}^{t} \Big(\vec{f}(s, \vec{x}(s)) - \vec{f}(s, \vec{y}(s)) \Big)\, ds \Big\| \leq \int_{t_0}^{t} \Big\| \vec{f}(s, \vec{x}(s)) - \vec{f}(s, \vec{y}(s)) \Big\|\, ds$$

$$\overset{(2.11)}{\leq} L \int_{t_0}^{t} \|\vec{x}(s) - \vec{y}(s)\|\, ds \leq L\, d(\vec{x}, \vec{y}) \int_{t_0}^{t_1} ds = L(t_1 - t_0)\, d(\vec{x}, \vec{y}).$$

Demnach hat T auf dem Intervall $[t_0, t_1]$ die Kontraktionskonstante $q := L(t_1 - t_0) < 1$. □

Der BANACHsche Fixpunktsatz 13.2 trifft also auf die Abbildung $T : C([a, b]; \mathbf{R}^n) \to C([a, b]; \mathbf{R}^n)$ aus (2.12) zu, sofern die einschränkende Bedingung $L(b - a) < 1$ erfüllt ist. Wir befreien uns von dieser unerwünschten Längenbeschränkung des Intervalls $[a, b]$ durch einen einfachen Trick.

Trick mit der gewichteten Metrik

Anstelle der Metrik (D6) erklären wir auf dem Funktionenraum $M := C([a,b];\mathbf{R}^n)$ die **gewichtete Metrik**

(D8) $$\boxed{d_r(\vec{x},\vec{y}) := \max_{t\in[a,b]} e^{-r(t-a)}\|\vec{x}(t)-\vec{y}(t)\|, \quad \vec{x},\vec{y}\in M.}$$

Hierin ist $r>0$ eine noch zu wählende Konstante. Es gilt offenbar

$$d_r(\vec{x},\vec{y}) \le d(\vec{x},\vec{y}) \le e^{r(b-a)} d_r(\vec{x},\vec{y}), \quad \vec{x},\vec{y}\in M,$$

so dass die Metriken (D6) und (D8) äquivalent sind.

Satz 13.4 (Existenzsatz von PICARD)
Der Funktionenraum $M:=C([a,b];\mathbf{R}^n)$ sei mit der gewichteten Metrik (D8) *versehen. Gegeben seien ferner ein $\vec{u}\in M$ und eine stetige Funktion $\vec{f}:[a,b]\times\mathbf{R}^n\to\mathbf{R}^n$, die der* LIPSCHITZ*–Bedingung* (2.11) *genügt. Dann wird durch die Vorschrift*

$$\boxed{(T\vec{x})(t) := \vec{u}(t) + \int_a^t \vec{f}(s,\vec{x}(s))\,ds, \quad \vec{x}\in M, \;\; t\in[a,b],} \tag{2.13}$$

eine Abbildung $T: M\to M$ erklärt, die genau einen Fixpunkt $\vec{x}=T\vec{x}\in M$ besitzt.

Begründung: Da M ein vollständiger metrischer Raum ist, können wir den BANACHschen Fixpunktsatz anwenden, sofern wir die Kontraktionseigenschaft (2.8) für T gezeigt haben. Dazu verfügen wir in geeigneter Weise über den Parameter $r>0$. Zunächst gilt für alle $\vec{x},\vec{y}\in M$ und $t\in[a,b]$ die Abschätzung

$$e^{-r(t-a)}\|(T\vec{x})(t)-(T\vec{y})(t)\| = e^{-r(t-a)}\Big\|\int_a^t \Big(\vec{f}(s,\vec{x}(s))-\vec{f}(s,\vec{y}(s))\Big)\,ds\Big\|$$

$$\overset{(2.11)}{\le} L\int_a^t e^{-r(t-s)}e^{-r(s-a)}\|\vec{x}(s)-\vec{y}(s)\|\,ds$$

$$\le L\,d_r(\vec{x},\vec{y})\int_a^t e^{-r(t-s)}\,ds = \frac{L}{r}\,(1-e^{-r(t-a)})\,d_r(\vec{x},\vec{y}) \le \frac{L}{r}\,(1-e^{-r(b-a)})\,d_r(\vec{x},\vec{y}).$$

Wir können nun $r>0$ so groß wählen, dass die Bedingung

$$\boxed{\frac{L}{r}\,(1-e^{-r(b-a)}) := q < 1} \tag{2.14}$$

erfüllt ist. Dann ist T eine Kontraktion mit der Kontraktionskonstanten q. □

Bemerkung 13.2 Gemäß (2.9) kann der Fixpunkt $\vec{x}$ der Abbildung T durch die Iterationsvorschrift

$$\vec{x}_{n+1}(t) := (T\vec{x}_n)(t) = \vec{u}(t) + \int_a^t \vec{f}(s,\vec{x}_n(s))\,ds, \quad n=0,1,2,\ldots$$

(näherungsweise) berechnet werden. Man wählt dabei zweckmäßigerweise den Startwert $\vec{x}_0(t)$ $:= \vec{u}(t)$. □

BSP. (13.2.6) Wir betrachten den skalaren Fall $n=1$ und spezifizieren dazu in (2.13) die Größen gemäß

$$f(t,x) := x, \quad u(t) := 1, \quad a := 0,\; b>0.$$

Es ist offenbar $L = 1$. Der eindeutig bestimmte Fixpunkt $x \in C([a,b])$ der Gleichung

$$x(t) = 1 + \int_0^t x(s)\, ds$$

ist Lösung der Anfangswertaufgabe

$$\dot{x}(t) - x(t) = 0, \quad t > 0; \qquad x(0) = 1,$$

und man verifiziert mit Standardargumenten $x(t) = e^t$. Hingegen liefert die obige Iterationsvorschrift die folgenden Näherungen

$$\begin{aligned}
x_0(t) &= 1, \\
x_1(t) &= 1 + \int_0^t ds = 1 + t, \\
x_2(t) &= 1 + \int_0^t (1+s)\, ds = 1 + t + \tfrac{1}{2!}\, t^2, \\
x_3(t) &= 1 + \int_0^t (1 + s + \tfrac{1}{2}\, s^2)\, ds = 1 + t + \tfrac{1}{2!}\, t^2 + \tfrac{1}{3!}\, t^3, \\
&\vdots
\end{aligned}$$

Offenbar liegt der Folge $(x_n)_{n\geq 0}$ das Bildungsgesetz

$$x_n(t) = 1 + t + \frac{1}{2!}\, t^2 + \cdots + \frac{1}{n!}\, t^n = \sum_{k=0}^{n} \frac{t^k}{k!}, \quad n \geq 0,$$

zugrunde, und dieses liefert im Limes $n \to \infty$ ebenfalls die oben behauptete Lösung $x(t) = e^t$.

13.3 Eigenschaften stetiger Funktionen $f \in \mathrm{Abb}(\mathbf{R}^n, \mathbf{R})$

Der Vektorraum $\mathbf{R}^n$ mit der Metrik (D3) ist ein metrischer Raum, und dies trifft in gleicher Weise für jede nichtleere Teilmenge $M \subset \mathbf{R}^n$ zu. Ebenso ist der Körper $\mathbf{R}$ mit der Metrik (D1) ein metrischer Raum. Demzufolge ist ein Stetigkeitsbegriff für Funktionen $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ mit $D(f) =: M \subset \mathbf{R}^n$ wohldefiniert: f ist gemäß Definition 13.7 genau dann in $\vec{x}_0 \in D(f)$ stetig, wenn für jede Folge $(\vec{x}_n)_{n\in\mathbf{N}} \subset D(f)$ mit $\lim\limits_{n\to\infty} \vec{x}_n = \vec{x}_0$ gilt:

$$\boxed{\lim_{n\to\infty} |f(\vec{x}_n) - f(\vec{x}_0)| = 0.} \tag{3.1}$$

Gemäß Satz 13.1 gilt äquivalent

$$\boxed{\forall\, \epsilon > 0\ \exists\, \delta = \delta(\epsilon, \vec{x}_0)\ :\ |f(\vec{x}) - f(\vec{x}_0)| < \epsilon\ \forall\, \vec{x} \in D(f) \text{ mit } 0 < \|\vec{x} - \vec{x}_0\| < \delta.} \tag{3.2}$$

Hierin bezeichnet $\|\vec{x}\| := \Big(\sum\limits_{j=1}^{n} |x_j|^2 \Big)^{1/2}$ wieder die euklidische Norm des Vektors $\vec{x} \in \mathbf{R}^n$, durch die die Metrik (D3) nach der Vorschrift $d(\vec{x}, \vec{y}) = \|\vec{x} - \vec{y}\|$ induziert wird. Die Funktion f heißt stetig (schlechthin), wenn Stetigkeit in jedem Punkt $\vec{x}_0 \in D(f)$ vorliegt.

Für je zwei stetige Funktionen $f, g \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ sind wiederum die folgenden Funktionen auf $D(f) \cap D(g)$ stetig:

$$\lambda f + \mu g \;\; \forall\, \lambda, \mu \in \mathbf{R}, \quad f \cdot g, \quad \frac{f}{g} \;\text{ in allen Punkten } \vec{x} \text{ mit } g(\vec{x}) \neq 0.$$

In gleicher Weise sind für stetige Funktionen $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$, $g \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R})$ und $\vec{y} \in \mathrm{Abb}(\mathbf{R}, \mathbf{R}^n)$ die folgenden Komposita dort stetig, wo sie definiert sind:

$$(g \circ f)(\vec{x}) := g(f(\vec{x})), \qquad (f \circ \vec{y})(t) := f(\vec{y}(t)).$$

Funktionen $f(\vec{x})$ mit einer **Unbestimmtheitsstelle** $\vec{x}_0 \in D(f)$ von der Form

$$f(\vec{x}_0) = \frac{0}{0}$$

bedürfen einer gesonderten Stetigkeitsbetrachtung. Wir erörten in den folgenden Beispielen die auftretenden Probleme.

BSP. (13.3.1) Wir betrachten $f \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R})$ mit

$$f(x,y) := \begin{cases} \dfrac{xy^2}{x^2+y^2} & : \; (x,y) \neq (0,0), \\ 0 & : \; (x,y) = (0,0). \end{cases}$$

Nach dem oben Gesagten über die Stetigkeit von Komposita ist f sicher in allen Punkten $(x,y) \neq (0,0)$ stetig. Im Punkt $(x_0, y_0) := (0,0)$ erhalten wir

$$|f(x,y) - f(0,0)| = \frac{y^2}{x^2+y^2}\,|x| \leq 1 \cdot |x| \leq \sqrt{x^2+y^2} = \|\vec{x}\| < \epsilon \;\; \forall\, 0 < \|\vec{x}\| < \delta := \epsilon.$$

Somit gilt (3.2), das heißt, f ist auch in $(0,0)$ **stetig.**

BSP. (13.3.2) Wir betrachten $f \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R})$ mit

$$f(x,y) := \begin{cases} \dfrac{xy}{x^2+y^2} & : \; (x,y) \neq (0,0), \\ 0 & : \; (x,y) = (0,0). \end{cases}$$

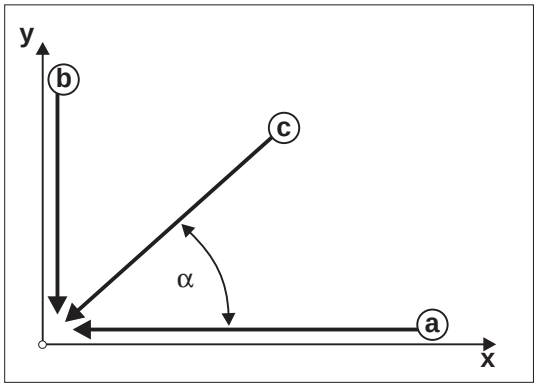
Es seien $\boxed{a}$, $\boxed{b}$, $\boxed{c}$ die unten skizzierten Wege, längs deren wir die Limites $(x,y) \to (0,0)$ vollziehen werden. Es gelten

$$\boxed{a} - \lim_{x \to 0} f(x,0) = 0 = \boxed{b} - \lim_{y \to 0} f(0,y),$$

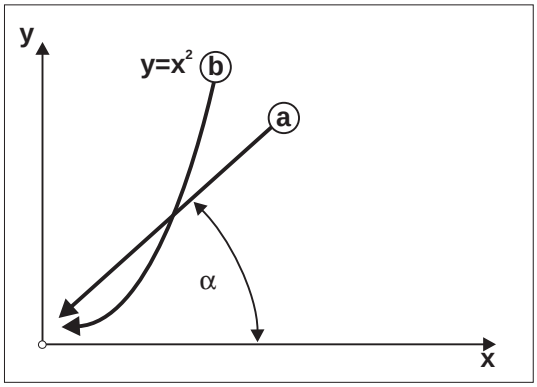
hingegen aber für $\alpha \neq j \cdot \frac{\pi}{2}$, $j = 0,1,2,3$:

$$\boxed{c} - \lim_{t \to 0} f(t\cos\alpha, t\sin\alpha) = \frac{1}{2}\,\sin 2\alpha \neq 0 = f(0,0).$$

Das heißt, f ist **unstetig** in $(0,0)$, denn die Relation (3.1) gilt nicht.



Zur Stetigkeit der Funktion f aus BSP. (13.3.2) in $(0,0)$



Zur Stetigkeit der Funktion f aus BSP. (13.3.3) in $(0,0)$

Durch dieses Beispiel wird das folgende **Unstetigkeitskriterium** motiviert: Ergeben sich für verschiedene Folgen $(\vec{x}_n)_{n\in\mathbf{N}} \subset D(f)$ mit $\lim\limits_{n\to\infty} \vec{x}_n = \vec{x}_0$ verschiedene Grenzwerte $\lim\limits_{n\to\infty} f(\vec{x}_n)$, so ist f sicher unstetig in $\vec{x}_0 \in D(f)$.

BSP. (13.3.3) Wir betrachten $f \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R})$ mit

$$f(x,y) := \begin{cases} \dfrac{1}{y} & : \; x = \sqrt{y}, \;\; y > 0, \\ 0 & : \; 0 \neq x \neq \sqrt{y}. \end{cases}$$

Es seien $\boxed{a}$, $\boxed{b}$ die oben skizzierten Wege, längs deren wir die Limites $(x,y) \to (0,0)$ vollziehen werden. Es gilt zunächst für jedes feste $\alpha \in [0, 2\pi]$

$$\boxed{a} - \lim_{t\to 0} f(t\cos\alpha, t\sin\alpha) = 0,$$

hingegen aber

$$\boxed{b} - \lim_{y\to 0+} f(\sqrt{y}, y) = \lim_{y\to 0+} \frac{1}{y} = +\infty.$$

Die Funktion f ist wiederum **unstetig** in $(0,0)$, jedoch genügt es **nicht**, Stetigkeit in $\vec{x}_0$ entlang strahlenförmiger Wege durch den Punkt $\vec{x}_0$ zu überprüfen. Insbesondere gilt folgende Feststellung:

> Stetigkeit in jeder Variablen einzeln impliziert keinesfalls bereits die Stetigkeit in allen Variablen zusammen.

Außer dem Grenzwert (3.1) können auch **iterierte Grenzwerte** gebildet werden.

BSP. (13.3.4) Wir betrachten $f \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R})$ mit

$$f(x,y) := \frac{x^2 - y^2 + x^3 + y^3}{x^2 + y^2}, \quad (x,y) \neq (0,0).$$

Offenbar berechnet man hier

$$\lim_{x\to 0} \Big(\lim_{y\to 0} f(x,y) \Big) = 1, \quad \text{aber} \quad \lim_{y\to 0} \Big(\lim_{x\to 0} f(x,y) \Big) = -1.$$

Diese Beobachtung führt auf ein weiteres **Unstetigkeitskriterium:**

> Hat $f(\vec{x})$ in einem Punkt $\vec{x}_0$ verschiedene iterierte Grenzwerte, so ist f in $\vec{x}_0$ stets unstetig. Ist jedoch f stetig in $\vec{x}_0$, so sind alle iterierten Grenzwerte gleich, sofern diese existieren.

BSP. (13.3.5) Weitere Beispiele stetiger Funktionen $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ sind:

- Die **konstante Funktion** $f(\vec{x}) := const, \;\; \vec{x} \in \mathbf{R}^n$.
- Die **affine Funktion** $f(\vec{x}) := \langle \vec{a}, \vec{x} \rangle + b, \quad \vec{x} \in \mathbf{R}^n$, mit festem Vektor $\vec{0} \neq \vec{a} \in \mathbf{R}^n$ und fester Zahl $b \in \mathbf{R}$. Die Äquipotentialflächen $f(\vec{x}) = h$ dieser Funktion sind offensichtlich die Hyperebenen $H : \; \langle \vec{a}, \vec{x} \rangle = \alpha := h - b$. Diese haben die Normalenrichtung $\vec{a}$ und den Abstand $d(H, \vec{0}) = |\alpha| / \|\vec{a}\|$ vom Ursprung.
- Die **quadratische Funktion** $f(\vec{x}) := \vec{x}^T A \vec{x} = \langle \vec{x}, A\vec{x} \rangle, \quad \vec{x} \in \mathbf{R}^n$, mit symmetrischer Matrix $A = A^T \in \mathbf{R}^{(n,n)}$. Zum Beispiel für $n = 2$

  $$A := \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}, \quad f(x,y) = a_{11}x^2 + 2a_{12}xy + a_{22}y^2, \quad (x,y) \in \mathbf{R}^2.$$

  Hier sind die Äquipotentialflächen offenbar *Kegelschnitte*. Im Fall $n = 3$ sind die Äquipotentialflächen die in Abschnitt 11.6 diskutierten *Quadriken*.

- Die **homogenen Polynome** vom Grade k. Man vergegenwärtige sich zunächst, dass ein n–dimensionaler **Multiindex** ρ ein geordnetes n–Tupel nichtnegativer ganzer Zahlen ist: $\rho := (\rho_1, \rho_2, \ldots, \rho_n) \in \mathbf{N}_0^n$. Wir werden im nächsten Abschnitt, Definition 13.12, detaillierter auf Multiindizes eingehen. Die **Ordnung** $|\rho|$ von $\rho = (\rho_1, \rho_2, \ldots, \rho_n)$ ist definiert durch

$$|\rho| := \sum_{j=1}^{n} \rho_j.$$

In diesem Sinne ist nun ein homogenes Polynom vom Grade k in den Variablen $\vec{x} = (x_1, x_2, \ldots, x_n)$ eine Funktion

$$f(\vec{x}) := \sum_{|\rho|=k} a_{\rho_1\rho_2\cdots\rho_n} x_1^{\rho_1} x_2^{\rho_2} \cdots x_n^{\rho_n} =: \sum_{|\rho|=k} a_\rho \vec{x}^\rho, \quad \vec{x} \in \mathbf{R}^n.$$

Zum Beispiel ist $f(x, y, z) := 4x^2z^3 + 3xy^2z^2 - 2y^4z$ ein homogenes Polynom vom Grade 5 in den Variablen (x, y, z). Offenbar gilt für homogene Polynome $f(\vec{x})$ vom Grade k die Relation

$$\boxed{f(\lambda\vec{x}) = \lambda^k f(\vec{x}) \quad \forall\ \lambda \in \mathbf{R}\ \ \forall\ \vec{x} \in \mathbf{R}^n.} \tag{3.3}$$

- Die **Polynome** vom Grade m:

$$f(\vec{x}) := \sum_{k=0}^{m} \sum_{|\rho|=k} a_{\rho_1\rho_2\cdots\rho_n} x_1^{\rho_1} x_2^{\rho_2} \cdots x_n^{\rho_n} =: \sum_{|\rho|\le m} a_\rho \vec{x}^\rho, \quad \vec{x} = (x_1, x_2, \ldots, x_n) \in \mathbf{R}^n.$$

Das sind gerade die Summen von homogenen Polynomen k–ten Grades mit $0 \le k \le m$. Zum Beispiel ist $f(x, y, z) := 4x^2z^3 + 3xy^2z^2 + 2x + y - z - 7$ ein Polynom vom Grade 5.

- Die **rationalen Funktionen**:

$$R(\vec{x}) := \frac{f(\vec{x})}{g(\vec{x})}, \quad \vec{x} \in \mathbf{R}^n \ \text{ und } \ g(\vec{x}) \neq 0,$$

worin f und g Polynome vom Grade m_f bzw. m_g sind.

Hängen Eigenschaften stetiger Funktionen $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R})$ in **einer** reellen Veränderlichen nur von einer **Abstandsmessung** ab, so können solche Eigenschaften unmittelbar auf den allgemeinen Fall einer stetigen Funktion $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ übertragen werden. Ein solches Beispiel ist der Satz 6.17 über die Erhaltung von Ungleichungen:

Satz 13.5 *Es sei $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ eine im Punkte $\vec{x}_0 \in D(f)$ stetige Funktion, und es gelte $f(\vec{x}_0) > g \in \mathbf{R}$. Dann folgt*

$$\boxed{\exists\ \delta > 0\ :\ f(\vec{x}) > g\ \forall\ \vec{x} \in D(f)\ \text{ mit }\ 0 < \|\vec{x} - \vec{x}_0\| < \delta.} \tag{3.4}$$

Eine analoge Aussage gilt natürlich auch in der Form $f(\vec{x}) < g$.

Wir sehen, dass auf der *Bildmenge* von $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ Ungleichungsrelationen uneingeschränkt verwendet werden dürfen, nicht aber auf dem Definitionsbereich $D(f) \subset \mathbf{R}^n$. Der Vektorraum $\mathbf{R}^n$, $n > 1$, ist ja kein Körper, und insbesondere nicht geordnet. Deshalb kann der Begriff der **monotonen Funktionen** aus Abschnitt 6.7 **nicht** auf den allgemeinen Fall $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ übertragen werden. Probleme mit dem Prinzip der Übertragung von Eigenschaften aus dem Eindimensionalen auf $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ müssen auch dort auftreten, wo der Auswahlsatz von Bolzano–Weierstrass verwendet wurde. Als Beispiel nennen wir den Satz 6.15 über die Beschränktheit einer stetigen Funktion. Dort hatten wir die Tatsache benutzt, dass jede (unendliche) beschränkte Folge mindestens einen Häufungspunkt besitzt – eine Aussage des Satzes von Bolzano–Weierstrass. Im Vektorraum $\mathbf{R}^n$ schaffen wir eine vergleichbare Situation durch Einführung eines neuen Begriffes:

Definition 13.8 *Eine Teilmenge $K \subset \mathbf{R}^n$ heiße* **kompakt**, *wenn jede unendliche Folge $(\vec{x}_n)_{n\in\mathbf{N}} \subset K$ mindestens einen Häufungspunkt $\vec{x}_0 \in K$ besitzt.*

Nun gilt die wörtliche Übertragung des Satzes 6.15.

Satz 13.6 *Es sei $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ eine auf der kompakten Teilmenge $K \subset \mathbf{R}^n$ stetige Funktion. Dann ist f* **beschränkt**:

$$\boxed{\exists\, M > 0 \;:\; |f(\vec{x})| \le M \;\; \forall\, \vec{x} \in K.} \tag{3.5}$$

Da also das Bild $f(K) \subset \mathbf{R}$ einer kompakten Teilmenge $K \subset \mathbf{R}^n$ unter einer stetigen Funktion f beschränkt ist, gilt das **Supremumsprinzip**: Es existieren $\sup f(K)$ und $\inf f(K)$. In Analogie zu Satz 6.16 erhalten wir deshalb:

Satz 13.7 (Extremalsatz)
Gegeben sei eine stetige Funktion $f : K \to \mathbf{R}$ über einer kompakten Teilmenge $K \subset \mathbf{R}^n$. Dann nimmt die Funktion f das Maximum und das Minimum ihrer Funktionswerte jeweils in einem Punkt der Menge K an:

$$\boxed{\exists\, \vec{x}_*, \vec{x}^* \in K \;:\; f(\vec{x}_*) = \min_{\vec{x}\in K} f(\vec{x}), \quad f(\vec{x}^*) = \max_{\vec{x}\in K} f(\vec{x}).} \tag{3.6}$$

Demgemäß gilt $f(\vec{x}_*) \le f(\vec{x}) \le f(\vec{x}^*) \;\forall\, \vec{x} \in K$.

BSP. (13.3.6) Wir betrachten bei festem $0 < R_1 < R_2$ die Menge $K := \{(x,y) \in \mathbf{R}^2 \;:\; R_1 \le \sqrt{x^2+y^2} \le R_2\}$, und darauf die Funktion

$$f(x,y) := \frac{x}{x^2+y^2}, \quad (x,y) \in K.$$

In ebenen *Polarkoordinaten* $x = r\cos\varphi$, $y = r\sin\varphi$ resultiert die Darstellung

$$\tilde{f}(r,\varphi) := f(r\cos\varphi, r\sin\varphi) = \frac{1}{r}\cos\varphi, \qquad r \in [R_1, R_2], \quad \varphi \in [0, 2\pi],$$

und somit

$$-\frac{1}{R_1} \le \tilde{f}(r,\varphi) \le \frac{1}{R_1},$$

wobei Minimum $-\frac{1}{R_1}$ und Maximum $\frac{1}{R_1}$ in den Punkten $(r_*, \varphi_*) := (R_1, \pi)$ bzw. $(r^*, \varphi^*) := (R_1, 0)$ angenommen werden. Beide Punkte liegen auf dem inneren Kreisrand $r = R_1$ des Kreisrings K. Entfernt man diesen Rand, das heißt, betrachtet man $\tilde{K} := \{(x,y) \in \mathbf{R}^2 \;:\; R_1 < \sqrt{x^2+y^2} \le R_2\}$, so nimmt f weder Maximum noch Minimum auf $\tilde{K}$ an: Satz 13.7 trifft nicht mehr zu. Da aber f auf $\tilde{K}$ stetig ist, darf die Menge $\tilde{K}$ nicht mehr kompakt sein!

Das obige Beispiel wirft die Frage auf, welches genau die kompakten Teilmengen von $\mathbf{R}^n$ sind. Zur Klärung dieser Frage führen wir Begriffe ein, die in allgemeinen metrischen Räumen M die Begriffe *offenes Intervall, abgeschlossenes Intervall, Randpunkt* ersetzen.

Definition 13.9 *Es sei M ein metrischer Raum mit einer Metrik $d(\cdot,\cdot)$, und es sei $x \in M$ fest.*

(a) *Die Menge*

$$\boxed{B_\epsilon(x) := \{y \in M \;:\; d(x,y) < \epsilon\}, \quad \epsilon > 0,}$$

heiße **offene ϵ–Kugel** *um den Punkt x, und entsprechend heiße*

$$\boxed{\overline{B_\epsilon}(x) := \{y \in M \,:\, d(x,y) \le \epsilon\}, \quad \epsilon \ge 0,}$$

abgeschlossene ϵ–Kugel *um den Punkt x.*

(b) *Es heiße $x \in M$*

- *ein* **innerer Punkt** *einer Teilmenge $\Omega \subset M$, wenn Ω eine offene ϵ–Kugel $B_\epsilon(x) \subset \Omega$ um den Punkt x enthält. In diesem Fall heiße Ω eine* **Umgebung** *von x,*
- *ein* **äußerer Punkt** *von $\Omega \subset M$, wenn x innerer Punkt der Komplementärmenge $M \setminus \Omega$ ist,*
- *ein* **Randpunkt** *von $\Omega \subset M$, wenn für jedes $\epsilon > 0$ sowohl $B_\epsilon(x) \not\subset \Omega$ als auch $B_\epsilon(x) \not\subset M \setminus \Omega$ gelten.*

Die Menge

$$\boxed{\Omega^\circ := \{x \in M \,:\, x \text{ ist innerer Punkt von } \Omega\}}$$

heiße **offener Kern** *oder* **Inneres** *von Ω; die Menge*

$$\boxed{\partial\Omega := \{x \in M \,:\, x \text{ ist Randpunkt von } \Omega\}}$$

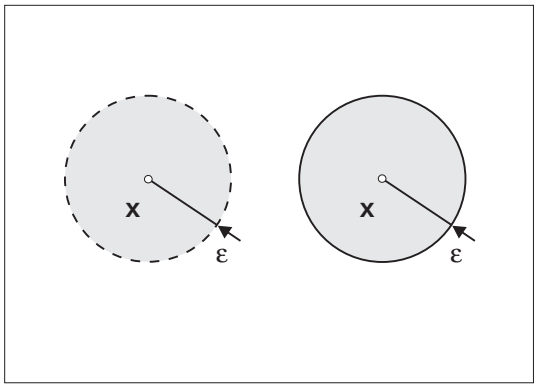
heiße der **Rand** *von Ω; schließlich heiße die Menge*

$$\boxed{\overline{\Omega} := \partial\Omega \cup \Omega}$$

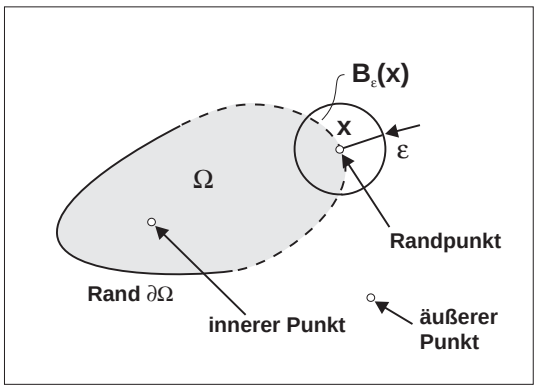
die **abgeschlossene Hülle** *von Ω.*

(c) *Eine Teilmenge $\Omega \subset M$ heiße* **offen**, *wenn $\Omega = \Omega^\circ$ gilt. Gilt $\Omega = \overline{\Omega}$, so heiße Ω* **abgeschlossen**.

(d) *Eine Teilmenge $\Omega \subset M$ heiße* **beschränkt**, *wenn es Elemente $R > 0$ und $x_0 \in M$ gibt mit $\Omega \subset B_R(x_0)$.*



Offene und abgeschlossene ϵ–Kugeln



Randpunkt, äußerer Punkt, innerer Punkt

Bemerkung 13.3 (a) Randpunkte einer Teilmenge $\Omega \subset M$ müssen nicht zu Ω gehören.

(b) Ist $M := \mathbf{R}^n$, so kann als Metrik die **euklidische Metrik** (D3) verwendet werden:

$$d(\vec{x}, \vec{y}) := \|\vec{x} - \vec{y}\| = \Big(\sum_{j=1}^{n} |x_j - y_j|^2 \Big)^{1/2}.$$

(c) Sämtliche Begriffe aus der Definition 13.9 sind invariant gegenüber Wechsel zu äquivalenten Metriken: Ist zum Beispiel eine Menge Ω offen in der einen Metrik, so ist sie offen in jeder dazu äquivalenten Metrik.

(d) In metrischen Räumen M wird der Begriff der **Kompaktheit** einer Teilmenge $K \subset M$ in gleicher Weise erklärt wie in $M := \mathbf{R}^n$: Genau dann ist K kompakt, wenn jede M–Folge $(x_n)_{n\in\mathbf{N}} \subset K$ mindestens einen Häufungspunkt $x_0 \in K$ hat. □

Es gilt jetzt:

Satz 13.8 *Es seien M ein metrischer Raum mit einer Metrik $d(\cdot,\cdot)$ und $K \subset M$ eine kompakte Teilmenge. Dann ist K abgeschlossen und beschränkt.*

Begründung: (a) Wäre K nicht abgeschlossen, so gäbe es einen Punkt $x_0 \in \partial K$ mit $x_0 \notin K$, und dazu eine Folge $(x_n)_{n\in\mathbf{N}} \subset K$ mit $\lim\limits_{n\to\infty} x_n = x_0$. Da x_0 einziger Häufungspunkt der Folge $(x_n)_{n\in\mathbf{N}}$ ist, ergäbe sich ein Widerspruch zur Kompaktheit von K.

(b) Wäre K unbeschränkt, so gäbe es in jeder Kugel $B_j(x_0)$, $j \in \mathbf{N}$, $x_0 \in K$ fest, einen Punkt $x_j \in K$ mit $j - 1 < d(x_j, x_0) < j$. Die Folge $(x_j)_{j\in\mathbf{N}}$ kann aber keine konvergente Teilfolge besitzen, also auch keinen Häufungspunkt $x \in K$. □

Liegt der **Sonderfall** des metrischen Raumes $M := \mathbf{R}^n$ vor, so gilt sogar die Umkehrung des obigen Satzes 13.8.

Satz 13.9 (von Heine–Borel)
Eine Teilmenge $K \subset \mathbf{R}^n$ ist genau dann kompakt, wenn K abgeschlossen und beschränkt ist.

Auf eine *Begründung* können wir hier nicht eingehen. Dieser Satz liefert die gewünschte Verallgemeinerung des Satzes von Bolzano–Weierstrass, und somit die gesuchte Charakterisierung der kompakten Teilmengen des $\mathbf{R}^n$.

BSP. (13.3.7) Die abgeschlossene ϵ–Kugel $\overline{B_\epsilon}(\vec{y})$ um einen festen Punkt $\vec{y} \in \mathbf{R}^n$ ist für jedes $\epsilon > 0$ kompakt. Bezeichnet $d(\vec{x}, \vec{y})$ wieder die euklidische Metrik (D3) in $\mathbf{R}^n$, so ist die Funktion $f(\vec{x}) := d(\vec{x}, \vec{x}_0)$, $\vec{x}_0 \in \mathbf{R}^n$ fest, auf ganz $\mathbf{R}^n$ stetig. Denn es gilt ja wegen der Dreiecksungleichung

$$|f(\vec{x}_1) - f(\vec{x}_2)| = |d(\vec{x}_1, \vec{x}_0) - d(\vec{x}_2, \vec{x}_0)| \le d(\vec{x}_1, \vec{x}_2) < \epsilon \ \ \forall\, \vec{x}_1, \vec{x}_2 \text{ mit } d(\vec{x}_1, \vec{x}_2) < \delta := \epsilon.$$

Folglich ist f auf der kompakten Menge $\overline{B_\epsilon}(\vec{y})$ beschränkt und nimmt dort Maximum und Minimum der Funktionswerte an.

BSP. (13.3.8) Der **offene positive Kegel** in $\mathbf{R}^n$ ist die Menge $\Omega := \{\vec{x} \in \mathbf{R}^n : x_j > 0, \quad j = 1, 2, \ldots, n\}$. Wir betrachten das homogene Polynom vom Grade $n - 1$

$$g(\vec{x}) := \sum_{j=1}^{n} x_1 x_2 \cdots x_{j-1} x_{j+1} \cdots x_n, \quad \vec{x} = (x_1, x_2, \ldots, x_n) \in \mathbf{R}^n,$$

worin vereinbarungsgemäß $x_0 := x_{n+1} := 1$ gelte. Es folgt nun $g(\vec{x}) > 0$ für alle $\vec{x} \in \Omega$, und da g stetig ist, muss auch $\frac{1}{g(\vec{x})}$ in jedem Punkt $\vec{x} \in \Omega$ stetig sein. Fixiert man ein solches $\vec{x}$, so gibt es

wegen der Offenheit von Ω eine abgeschlossene R–Kugel $\overline{B_R}(\vec{x}) \subset \Omega$. Diese Kugel ist kompakt. Wir betrachten auf $\overline{B_R}(\vec{x})$ die stetigen Funktionen

$$h(\vec{z}) := \Big(\sum_{j=1}^{n} z_1^2 z_2^2 \cdots z_{j-1}^2 z_{j+1}^2 \cdots z_n^2 \Big)^{1/2}, \quad z_0 := z_{n+1} := 1,$$

wobei wir speziell $\vec{z} := (x_1 y_1, x_2 y_2, \ldots, x_n y_n) \in \mathbf{R}^n$ bei festem $\vec{x} \in \mathbf{R}^n$ und beliebigem $\vec{y} = (y_1, y_2, \ldots, y_n) \in \mathbf{R}^n$ annehmen. Da die Funktion

$$R(\vec{y}) := \frac{h(\vec{z})}{|g(\vec{x})|\,|g(\vec{y})|}, \quad \vec{y} \in \overline{B_R}(\vec{x}),$$

stetig ist, folgt aus der Kompaktheit von $\overline{B_R}(\vec{x})$ die Existenz von

$$C := \max_{\vec{y} \in \overline{B_R}(\vec{x})} R(\vec{y}).$$

Nach diesen Vorbetrachtungen untersuchen wir die Stetigkeit der Funktion

$$f(\vec{x}) := \Big(\sum_{j=1}^{n} \frac{1}{x_j} \Big)^{-1}$$

auf der Menge Ω. (Man vergleiche dazu BSP. (13.1.1) über den Gesamtwiderstand einer Parallelschaltung von n Widerständen.) Dazu seien $\vec{x} \in \Omega$ und $\overline{B_R}(\vec{x})$ wie oben angegeben. Es folgt für $\vec{y} \in \overline{B_R}(\vec{x})$:

$$\begin{aligned}
|f(\vec{y}) - f(\vec{x})| &= \Big| \Big(\sum_{j=1}^{n} \frac{1}{y_j} \Big)^{-1} - \Big(\sum_{j=1}^{n} \frac{1}{x_j} \Big)^{-1} \Big| = \frac{\Big| \sum\limits_{j=1}^{n} \frac{y_j - x_j}{x_j y_j} \Big|}{\Big| \sum\limits_{j=1}^{n} \frac{1}{x_j} \Big| \Big| \sum\limits_{j=1}^{n} \frac{1}{y_j} \Big|} \\
&\leq \frac{\Big(\sum\limits_{j=1}^{n} \Big(\frac{1}{x_j y_j} \Big)^2 \Big)^{1/2}}{\Big| \sum\limits_{j=1}^{n} \frac{1}{x_j} \Big| \Big| \sum\limits_{j=1}^{n} \frac{1}{y_j} \Big|} \Big(\sum_{j=1}^{n} |y_j - x_j|^2 \Big)^{1/2} = R(\vec{y})\, d(\vec{y}, \vec{x}) \\
&\leq C\, d(\vec{y}, \vec{x}) < \epsilon \;\; \forall\, \vec{y} \in \overline{B_R}(\vec{x}) \text{ mit } d(\vec{y}, \vec{x}) < \delta := \frac{\epsilon}{C}.
\end{aligned}$$

Somit ist $f : \Omega \to \mathbf{R}$ stetig.

Stetige Funktionen $f : [a, b] \to \mathbf{R}$ sind sogar gleichmäßig stetig. Dies wurde in Satz 6.20 gezeigt. Der dort geführte Beweis bleibt für Funktionen $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ richtig, wenn $D(f) =: K$ eine kompakte Teilmenge ist.

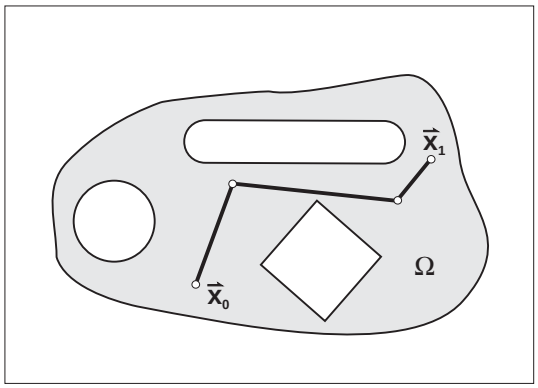
Satz 13.10 *Eine stetige Funktion $f : K \to \mathbf{R}$ ist auf der* **kompakten** *Teilmenge $K \subset \mathbf{R}^n$ sogar gleichmäßig stetig.*

Wir verallgemeinern schließlich den Zwischenwertsatz 6.19 von Bolzano. Die Aussage dieses Satzes lautet kurzgefasst, dass eine stetige Funktion $f : [a, b] \to \mathbf{R}$ das Intervall $D(f) := [a, b]$ wieder in ein Intervall abbildet. Ist $D(f)$ kein Intervall, so wird diese Aussage falsch.

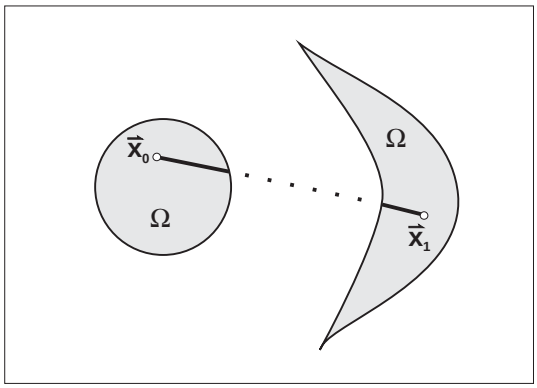
Wir betrachten *zum Beispiel* auf $D(f) := [-1, 1] \setminus \{0\}$ die stetige Funktion $f(x) := \frac{1}{x}$. Die Bildmenge zerfällt in zwei getrennte Intervalle

$$f(D(f)) = (-\infty, -1] \cup [1, +\infty).$$

Das Spezifische am Intervallbegriff ist die Tatsache, dass alle Punkte zwischen Intervallanfang und Intervallende zum Intervall gehören. Diese Eigenschaft können wir auch in $\mathbf{R}^n$ modellieren:



Zusammenhängende Teilmenge



Nicht zusammenhängende Teilmenge

Definition 13.10 *Eine Teilmenge $\Omega \subset \mathbf{R}^n$ heiße (weg–)***zusammenhängend***, wenn je zwei ihrer Punkte $\vec{x}_0, \vec{x}_1 \in \Omega$ durch einen stetigen Weg $\vec{x}(\cdot) : [0,1] \to \Omega$ verbunden werden können, der ganz in Ω verläuft: Es gelten $\vec{x}_0 = \vec{x}(0)$ und $\vec{x}_1 = \vec{x}(1)$ sowie $\vec{x}(t) \in \Omega$ für alle $t \in [0,1]$. Eine nichtleere, offene und zusammenhängende Menge $\Omega \subset \mathbf{R}^n$ heiße ein* **Gebiet***.*

Satz 13.11 (Zwischenwertsatz)
Sind $\vec{x}_0, \vec{x}_1 \in \Omega$ zwei Punkte einer **zusammenhängenden** *Teilmenge $\Omega \subset \mathbf{R}^n$, und ist eine stetige Funktion $f : \Omega \to \mathbf{R}$ gegeben, so nimmt f jeden Wert des Intervalls zwischen $f(\vec{x}_0)$ und $f(\vec{x}_1)$ mindestens einmal an.*

Begründung: Ist $\vec{x}(t)$ mit $\vec{x}(0) = \vec{x}_0$ und $\vec{x}(1) = \vec{x}_1$ sowie $\vec{x}(t) \in \Omega \ \forall\, t \in [0,1]$ der stetige Weg, der die Punkte $\vec{x}_0, \vec{x}_1 \in \Omega$ verbindet, so ist die Funktion

$$h(t) := f(\vec{x}(t)), \quad t \in [0,1],$$

eine stetige Funktion, auf die der Satz 6.19 zutrifft: h nimmt jeden Wert des Intervalls zwischen $h(0) = f(\vec{x}_0)$ und $h(1) = f(\vec{x}_1)$ mindestens einmal an. □

BSP. (13.3.9) Die Kugelfläche $S_1(\vec{0}) := \{\vec{x} \in \mathbf{R}^3 \,:\, \|\vec{x}\| = 1\} \subset \mathbf{R}^3$ ist kompakt und zusammenhängend. Hat man auf $S_1(\vec{0})$ eine stetige Temperaturverteilung $T(\vec{x})$ vorgegeben, so muss es gemäß Satz 13.7 Punkte $\vec{x}_*, \vec{x}^* \in S_1(\vec{0})$ mit minimaler bzw. maximaler Temperatur geben. Wegen Satz 13.11 wird auch jede Zwischentemperatur im Intervall zwischen $T(\vec{x}_*)$ und $T(\vec{x}^*)$ angenommen.

Aus den Sätzen 13.7 und 13.11 resultieren einige einfache Folgerungen.

Folgerung 13.1 *Es sei $f : \Omega \to \mathbf{R}$ auf der Teilmenge $\Omega \subset \mathbf{R}$ stetig.*

(a) *Ist Ω zusammenhängend, so ist $f(\Omega)$ ein Intervall (also selbst zusammenhängend).*

(b) *Ist Ω kompakt, so ist auch $f(\Omega)$ kompakt.*

(c) *Ist Ω kompakt und zusammenhängend, so ist $f(\Omega)$ ein abgeschlossenes Intervall:*

$$f(\Omega) = [f(\vec{x}_*), f(\vec{x}^*)], \quad f(\vec{x}_*) = \min_{\vec{x} \in \Omega} f(\vec{x}), \quad f(\vec{x}^*) = \max_{\vec{x} \in \Omega} f(\vec{x}).$$

13.4 Partielle Ableitungen

Da eine Division durch Vektoren im allgemeinen nicht erklärt ist, kann der Ableitungsbegriff für Funktionen $f \in \mathrm{Abb}(\mathbf{R}^n, \mathbf{R})$, $n > 1$, nicht wie im eindimensionalen Fall als Grenzwert der Folge der Differenzenquotienten definiert werden. Hingegen können $n-1$ der n vorhandenen Variablen $\vec{x} = (x_1, x_2, \ldots, x_n)$ festgehalten werden, *zum Beispiel* $\vec{x} = (a_1, \ldots, a_{j-1}, x_j, a_{j+1}, \ldots, a_n)$, und es kann $f(\vec{x}) =: g(x_j)$ als Funktion der eindimensionalen Veränderlichen x_j aufgefasst werden. Auf diesem Wege gelangt man zur folgenden

Definition 13.11 *Gegeben seien eine Funktion* $f \in \mathrm{Abb}(\mathbf{R}^n, \mathbf{R})$ *und ein* **innerer** *Punkt* $\vec{x} \in D(f)$. *Existiert der Grenzwert*

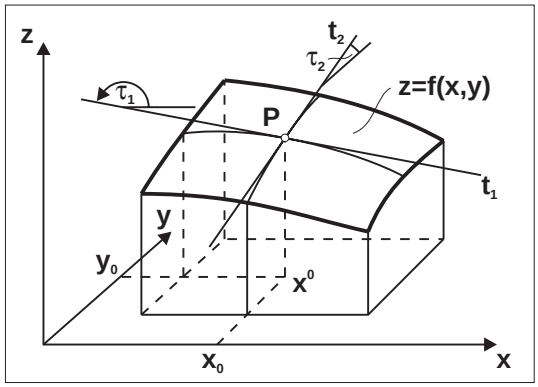
$$\boxed{\frac{\partial f}{\partial x_j}(\vec{x}) := \lim_{h \to 0} \frac{1}{h}\left(f(\vec{x} + h\vec{e_j}) - f(\vec{x})\right), \quad \vec{e_j} := (0, \ldots, 0, \underbrace{1}_{j-\text{te Stelle}}, 0, \ldots, 0),}$$

so heiße dieser **partielle Ableitung** *von* f *im Punkte* $\vec{x} \in D(f)$ *nach der* j*–ten Komponenten. Die Funktion* f *heiße in* $\vec{x} \in D(f)$ **partiell differenzierbar**, *wenn die partiellen Ableitungen* $\frac{\partial f}{\partial x_j}(\vec{x})$ *nach* **allen** *Komponenten* $j = 1, 2, \ldots, n$ *existieren. Die Funktion* f *heiße* **in** $D(f)$ **partiell differenzierbar**, *wenn* f *in allen inneren Punkten* $\vec{x} \in D(f)$ *partiell differenzierbar ist.*

Bemerkung 13.4 (a) Andere Bezeichnungen für partielle Ableitungen sind

$$f_{x_j} \equiv \frac{\partial}{\partial x_j} f \equiv \partial_j f \equiv D_j f \equiv f_{,j} \equiv \frac{\partial f}{\partial x_j}.$$

(b) Die partielle Ableitung $\frac{\partial f}{\partial x_j}$ der Funktion f ist genau die gewöhnliche Ableitung der Funktion $g(x_j) = f(\vec{x})$ bei $\vec{x} = (a_1, \ldots, a_{j-1}, x_j, a_{j+1}, \ldots, a_n)$.



Zur geometrischen Bedeutung der partiellen Ableitungen

(c) Im Falle $n = 2$ gilt die folgende geometrische Deutung der partiellen Ableitungen: Die Schnittkurve der Bildfläche von $z = f(x, y)$ mit der Ebene $y = y_0$ ist der Graph der Funktion $f_1(x) := f(x, y_0)$. Nun ist $f_1'(x_0) := f_x(x_0, y_0)$ die Steigung dieser Kurve im Punkte (x_0, y_0), also $\tan \tau_1 = f_1'(x_0) = f_x(x_0, y_0)$. Ganz entsprechend gilt $\tan \tau_2 = f_y(x_0, y_0)$, siehe obige Skizze. □

BSP. (13.4.1) Wir betrachten die Funktion

$$\boxed{f(x, y) := \operatorname{arc\,tan}_H \frac{y}{x}, \qquad D(f) := \mathbf{R}^2 \setminus \{(0, y)\}.}$$

Da $D(f) \subset \mathbf{R}^2$ eine offene Teilmenge ist, ist jeder Punkt $(x,y) \in D(f)$ ein innerer Punkt, und es existieren die partiellen Ableitungen

$$f_x(x,y) = \frac{1}{1+(y/x)^2}\left(-\frac{y}{x^2}\right) = -\frac{y}{x^2+y^2}, \quad f_y(x,y) = \frac{1}{1+(y/x)^2}\,\frac{1}{x} = \frac{x}{x^2+y^2}, \quad x \neq 0.$$

Die bekannten **Rechenregeln** für Ableitungen von Funktionen $f \in \mathrm{Abb}\,(\mathbf{R},\mathbf{R})$ übertragen sich sinngemäß auf Funktionen $f \in \mathrm{Abb}\,(\mathbf{R}^n,\mathbf{R})$, *zum Beispiel*

- die **Produktregel**

$$\boxed{\frac{\partial}{\partial x_j}(f\cdot g) = g\cdot\frac{\partial f}{\partial x_j} + f\cdot\frac{\partial g}{\partial x_j},}$$

- die **Kettenregel:** Für eine differenzierbare Funktion $h \in \mathrm{Abb}\,(\mathbf{R},\mathbf{R})$ gilt

$$\boxed{\frac{\partial}{\partial x_j}\,h(f(\vec{x})) = h'(f(\vec{x}))\cdot\frac{\partial f}{\partial x_j}\,(\vec{x}).}$$

  Diese Regel wurde in BSP. (13.4.1) mit $h(t) := \operatorname{arc\,tan}_H t$ und $f(x,y) := \frac{y}{x}$ benutzt.

BSP. (13.4.2) Zustandsgleichung für ein **ideales Gas**:

$$\left.\begin{array}{llll} p = n\frac{RT}{V} & \Rightarrow & \frac{\partial p}{\partial V} = -n\frac{RT}{V^2}, & \frac{\partial p}{\partial T} = n\frac{R}{V}, \\ V = n\frac{RT}{p} & \Rightarrow & \frac{\partial V}{\partial T} = n\frac{R}{p}, & \frac{\partial V}{\partial p} = -n\frac{RT}{p^2}, \\ T = \frac{pV}{nR} & \Rightarrow & \frac{\partial T}{\partial p} = \frac{V}{nR}, & \frac{\partial T}{\partial V} = \frac{p}{nR} \end{array}\right\} \quad \Rightarrow \quad \boxed{\frac{\partial p}{\partial V}\cdot\frac{\partial V}{\partial T}\cdot\frac{\partial T}{\partial p} = -n\,\frac{RT}{pV} = -1.}$$

Beachte: Die formale Behandlung der partiellen Ableitungen wie gewöhnliche Quotienten ist **unzulässig**! Sonst ergäbe das obige Produkt den Wert 1.

BSP. (13.4.3) Wir betrachten die Funktion $f(x,y)$ aus BSP. (13.3.2), nämlich

$$\boxed{f(x,y) := \begin{cases} \dfrac{xy}{x^2+y^2} & : \ (x,y) \neq (0,0), \\ 0 & : \ (x,y) = (0,0). \end{cases}}$$

Wie wir bereits gezeigt haben, ist f im Punkt $(0,0)$ **unstetig**. Hingegen existieren partielle Ableitungen in diesem Punkt:

$$\begin{aligned} f_x(0,0) &= \lim_{h\to 0}\frac{1}{h}\,(f(h,0)-f(0,0)) = \lim_{h\to 0}\frac{1}{h}\cdot 0 = 0, \\ f_y(0,0) &= \lim_{h\to 0}\frac{1}{h}\,(f(0,h)-f(0,0)) = \lim_{h\to 0}\frac{1}{h}\cdot 0 = 0. \end{aligned}$$

Beachte: Anders als bei der gewöhnlichen Ableitung folgt aus der partiellen Differenzierbarkeit einer Funktionen $f \in \mathrm{Abb}\,(\mathbf{R}^n,\mathbf{R})$ in einem Punkt $\vec{x}_0$ **nicht** schon die Stetigkeit von f in diesem Punkt:

partielle Differenzierbarkeit impliziert keinesfalls Stetigkeit!

Insofern ist der Begriff der partiellen Differenzierbarkeit unbefriedigend, und wir werden deshalb im nächsten Abschnitt 13.5 einen besseren Begriff einführen. In diesem Zusammenhang werden wir auch den folgenden Satz beweisen:

Satz 13.12 *Gegeben seien $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ und ein innerer Punkt $\vec{x}_0 \in D(f)$. Ist die Funktion f in einer Umgebung von $\vec{x}_0$ partiell differenzierbar und sind alle partiellen Ableitungen in $\vec{x}_0$ stetig, so ist auch f in $\vec{x}_0$ stetig.*

BSP. (13.4.4) Es sei $f(x,y)$ die Funktion aus dem vorangegangenen BSP. (13.4.3). Dann gilt für $(x,y) \neq (0,0)$:

$$f_x(x,y) = \frac{y(y^2-x^2)}{(x^2+y^2)^2}, \qquad f_y(x,y) = \frac{x(x^2-y^2)}{(x^2+y^2)^2}.$$

Wird nun für $|\alpha| \neq 1$ die Gerade $y = \alpha x$ betrachtet, so ergeben sich die folgenden Limites längs dieser Geraden:

$$\lim_{x\to 0} f_x(x,\alpha x) = \lim_{x\to 0} \frac{\alpha(\alpha^2-1)}{x(1+\alpha^2)^2} = \pm\infty, \quad \lim_{x\to 0} f_y(x,\alpha x) = \lim_{x\to 0} \frac{1-\alpha^2}{x(1+\alpha^2)^2} = \pm\infty,$$

während $f_x(0,0) = 0 = f_y(0,0)$ gelten. Die partiellen Ableitungen in $(0,0)$ sind unstetig; somit kann auch keine Stetigkeit von f im Punkte $(0,0)$ erwartet werden.

Partielle Ableitungen höherer Ordnung

Die partiellen Ableitungen $\frac{\partial f}{\partial x_j}$ einer Funktion $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ sind im allgemeinen wieder Funktionen der Variablen $\vec{x} = (x_1, x_2, \ldots, x_n)$, und sie können somit ebenfalls partielle Ableitungen nach den Veränderlichen x_k besitzen. Man wird zwangsläufig zum Begriff der **zweiten** (und höheren) partiellen Ableitung einer Funktion f geführt, *zum Beispiel*

$$\begin{aligned}
\frac{\partial^2 f}{\partial x_j \partial x_k}(\vec{x}_0) &:= \frac{\partial}{\partial x_j}\Big(\frac{\partial f}{\partial x_k}(\vec{x})\Big)\Big|_{\vec{x}=\vec{x}_0} := \frac{\partial}{\partial x_j} f_{x_k}(\vec{x}_0) := f_{x_k x_j}(\vec{x}_0), \quad j,k = 1,2,\ldots,n,\\
\frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}(\vec{x}_0) &:= f_{x_k x_j x_i}(\vec{x}_0), \quad \text{usw.}
\end{aligned}$$

BSP. (13.4.5) Wir hatten in BSP. (13.4.1) bereits die partiellen Ableitungen $f_x(x,y) = -\frac{y}{x^2+y^2}$ und $f_x(x,y) = \frac{x}{x^2+y^2}$ der Funktion $f(x,y) := \operatorname{arc\,tan}_H \frac{y}{x}$, $x \neq 0$, betrachtet. Es gilt des weiteren:

$$f_{xx}(x,y) = \frac{2xy}{(x^2+y^2)^2}, \quad f_{yy}(x,y) = \frac{-2xy}{(x^2+y^2)^2}, \quad f_{xy}(x,y) = \frac{y^2-x^2}{(x^2+y^2)^2} = f_{yx}(x,y).$$

Die Gleichheit der gemischten partiellen Ableitungen $f_{xy} = f_{yx}$ ist keinesfalls zufällig. Es gilt nämlich allgemein:

Satz 13.13 (von H.A. Schwarz**)**
Gegeben seien eine Funktion $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ und ein innerer Punkt $\vec{x}_0 \in D(f)$, in welchem f stetig partiell differenzierbar sei. Existieren in $\vec{x}_0$ die zweiten partiellen Ableitungen $\frac{\partial^2 f}{\partial x_j \partial x_k}$ sowie $\frac{\partial^2 f}{\partial x_k \partial x_j}$, und sind diese stetig in $\vec{x}_0$, so gilt stets

$$\boxed{\frac{\partial^2 f}{\partial x_j \partial x_k}(\vec{x}_0) = \frac{\partial^2 f}{\partial x_k \partial x_j}(\vec{x}_0).} \tag{4.1}$$

Begründung: Da hier nur die beiden Variablen x_j, x_k auftreten, genügt es, sich auf den Sonderfall einer Funktion $f(x,y)$ von zwei Veränderlichen zu beschränken. Es sei also $\vec{x}_0 = (x,y) \in D(f)$ ein innerer Punkt, und es sei $\epsilon > 0$ so bestimmt, dass die ϵ–Kugel $B_\epsilon(\vec{x}_0) \subset D(f)$ erfüllt. Für $(h,k) \neq (0,0)$ mit $\vec{x}_0 + (h,k) \in B_\epsilon(\vec{x}_0)$ setzen wir

$$u(y) := f(x+h,y) - f(x,y)$$

und bestimmen nach dem Mittelwertsatz Zahlen ϑ_1, $\vartheta_2 \in (0,1)$ mit

$$\begin{aligned} u(y+k) - u(y) &= ku_y(y+\vartheta_1 k) = k(f_y(x+h, y+\vartheta_1 k) - f_y(x, y+\vartheta_1 k)) \\ &= kh \cdot f_{yx}(x+\vartheta_2 h, y+\vartheta_1 k). \end{aligned} \tag{4.2}$$

Setzen wir hingegen

$$v(x) := f(x,y+k) - f(x,y),$$

so können wir in gleicher Weise Zahlen ϑ_3, $\vartheta_4 \in (0,1)$ bestimmen mit

$$\begin{aligned} v(x+h) - v(x) &= hv_x(x+\vartheta_3 h) = h(f_x(x+\vartheta_3 h, y+k) - f_x(x+\vartheta_3 h, y)) \\ &= kh \cdot f_{xy}(x+\vartheta_3 h, y+\vartheta_4 k). \end{aligned} \tag{4.3}$$

Nun ist

$$u(y+k) - u(y) = f(x+h,y+k) - f(x,y+k) - f(x+h,y) + f(x,y) = v(x+h) - v(x),$$

so dass aus (4.2) und (4.3) resultieren:

$$f_{yx}(x+\vartheta_2 h, y+\vartheta_1 k) = f_{xy}(x+\vartheta_3 h, y+\vartheta_4 k).$$

Wegen der vorausgesetzten Stetigkeit erhält man im Limes $h,k \to 0$ die behauptete Gleichheit. □

BSP. (13.4.6) Wir betrachten hier die Funktion

$$\boxed{f(x,y) := \begin{cases} \dfrac{x^2 y}{x^2+y^2} & : \ (x,y) \neq (0,0), \\ 0 & : \ (x,y) = (0,0). \end{cases}}$$

Wie in BSP. (13.4.3) zeigt man $f_x(0,0) = 0 = f_y(0,0)$. Außerhalb des Punktes $(0,0)$ gelten:

$$f_x(x,y) = \frac{2xy^3}{(x^2+y^2)^2}, \qquad f_y(x,y) = \frac{x^2(x^2-y^2)}{(x^2+y^2)^2}.$$

Hieraus erschließen wir

$$f_{xy}(0,0) = \lim_{h\to 0} \frac{1}{h}\left(f_x(0,h) - f_x(0,0)\right) = \lim_{h\to 0} \frac{1}{h} \cdot 0 = 0,$$

aber

$$f_{yx}(0,0) = \lim_{h\to 0} \frac{1}{h}\left(f_y(h,0) - f_y(0,0)\right) = \lim_{h\to 0} \frac{1}{h} \cdot 1 = \pm\infty.$$

Das heißt, die Ableitung $f_{yx}(0,0)$ existiert überhaupt nicht. Die SCHWARZsche Vertauschungsregel (4.1) gilt hier nicht, da die Voraussetzungen des Satzes 13.13 nicht erfüllt sind.

BSP. (13.4.7) Für einen festen Vektor $\vec{0} \neq \vec{a} \in \mathbf{R}^n$ betrachten wir unter Verwendung des Standardskalarproduktes $\langle \vec{a}, \vec{x} \rangle$ die Funktion

$$\boxed{f(\vec{x}) := \sin\langle \vec{a}, \vec{x} \rangle, \quad \vec{x} \in \mathbf{R}^n, \quad \vec{a} = (a_1, a_2, \ldots, a_n).}$$

Wir bilden die partiellen Ableitungen

$$\begin{aligned} f_{x_i}(\vec{x}) &= a_i \cos\langle \vec{a}, \vec{x}\rangle, \\ f_{x_i x_j}(\vec{x}) &= -a_i a_j \sin\langle \vec{a}, \vec{x}\rangle = f_{x_j x_i}(\vec{x}), \\ f_{x_i x_j x_k}(\vec{x}) &= -a_i a_j a_k \cos\langle \vec{a}, \vec{x}\rangle = f_{x_j x_i x_k}(\vec{x}) = f_{x_k x_j x_i}(\vec{x}) = f_{x_k x_i x_j}(\vec{x}) = f_{x_j x_k x_i}(\vec{x}), \\ &\vdots \end{aligned}$$

Das heißt, die SCHWARZsche Vertauschungsregel gilt unter entsprechenden Stetigkeitsvoraussetzungen auch für höhere Ableitungen.

Multiindizes

Multiindizes waren uns bereits in BSP. (13.3.5) begegnet. Wir formalisieren hier ihre Definition und erweitern ihren Anwendungsbereich. Denn wie das obige BSP. (13.4.7) zeigt, kann das Hinschreiben der höheren Ableitungen einer Funktion von der Schreibtechnik her sehr aufwendig sein. Die Verwendung von Multiindizes führt zu erheblichen Vereinfachungen.

Definition 13.12 (a) *Ein geordnetes n–Tupel nichtnegativer ganzer Zahlen*

$$\boxed{\rho = (\rho_1, \rho_2, \ldots, \rho_n) \in \mathbf{N}_0^n}$$

heiße n–dimensionaler **Multiindex**. *Die Zahlen $\rho_j \in \mathbf{N}_0$ heißen die Komponenten von ρ.*

(b) *Die Zahl*

$$\boxed{|\rho| := \sum_{j=1}^{n} \rho_j}$$

heiße die **Ordnung** *oder* **Länge** *des Multiindex $\rho = (\rho_1, \rho_2, \ldots, \rho_n)$.*

(c) *Ist ρ ein Multiindex und a eine Zahl (ein Koeffizient oder eine Funktion), so heiße a_ρ eine* **mehrfach indizierte Größe**. *Ist $\{a_\rho : |\rho| = k\}$ eine Familie mehrfach indizierter Größen, deren Multiindizes ρ alle dieselbe Länge k haben, so bedeute*

$$\boxed{\sum_{|\rho|=k} a_\rho := \text{ Summe aller } a_\rho \text{ mit Multiindex der Länge } k,}$$

zum Beispiel für $n = 3$:

$$\sum_{|\rho|=2} a_\rho = a_{200} + a_{020} + a_{002} + a_{110} + a_{101} + a_{011}.$$

Analog setzt man für festes $m \in \mathbf{N}_0$:

$$\boxed{\sum_{|\rho|\le m} a_\rho := \sum_{k=0}^{m} \sum_{|\rho|=k} a_\rho.}$$

(d) *Ist $\rho = (\rho_1, \rho_2, \ldots, \rho_n)$ ein n–dimensionaler Multiindex, so bezeichne $\vec{x}^\rho$ für $\vec{x} = (x_1, x_2, \ldots, x_n) \in \mathbf{R}^n$ das Polynom vom Grade $|\rho|$*

$$\boxed{\vec{x}^\rho := x_1^{\rho_1} \cdot x_2^{\rho_2} \cdots x_n^{\rho_n};}$$

zum Beispiel gilt in $\mathbf{R}^3$ *für* $\rho := (0,3,2)$ *und* $\vec{x} = (x,y,z)$*:* $\vec{x}^\rho = x^0 \cdot y^3 \cdot z^2 = y^3 z^2$.

(e) *In* $\mathbf{R}^n$ *bezeichne* $D_j := \frac{\partial}{\partial x_j}$, $j = 1,2,\ldots,n$, *den partiellen Differentialoperator nach der* j*–ten Variablen, zum Beispiel*

$$D_3 f(\vec{x}) := \frac{\partial f}{\partial x_3}(\vec{x}), \quad D_j f(\vec{x}) := \frac{\partial f}{\partial x_j}(\vec{x}), \quad D_j D_k f(\vec{x}) := \frac{\partial^2 f}{\partial x_j \partial x_k}(\vec{x}).$$

Ist $\rho = (\rho_1, \rho_2, \ldots, \rho_n)$ *ein* n*–dimensionaler Multiindex, so bezeichne* D^ρ *den* **partiellen Differentialoperator** *der Ordnung* $|\rho|$*:*

$$\boxed{D^\rho := D_1^{\rho_1} D_2^{\rho_2} \cdots D_n^{\rho_n} := \frac{\partial^{|\rho|}}{\partial x_1^{\rho_1} \partial x_2^{\rho_2} \cdots \partial x_n^{\rho_n}}.}$$

Zum Beispiel gilt in $\mathbf{R}^3$ *mit* $\rho := (0,3,2)$*:*

$$D^\rho f(x,y,z) = \frac{\partial^5 f(x,y,z)}{\partial y^3 \partial z^2} = f_{zzyyy}(x,y,z).$$

Partielle Ableitungen einer Funktion $f \in \mathrm{Abb}(\mathbf{R}^n, \mathbf{R})$ haben wir nur in inneren Punkten von $D(f)$ erklärt. Deshalb wird in der folgenden Definition eine **offene** Teilmenge $\Omega \subset \mathbf{R}^n$ bevorzugt, die nur aus inneren Punkten besteht.

Definition 13.13 *Es seien* $\Omega \subset \mathbf{R}^n$ *eine offene Teilmenge und* $m \in \mathbf{N}_0$ *fest. Dann setzt man*

$$\boxed{\begin{aligned} C^m(\Omega) &:= \{f : \Omega \to \mathbf{R} \,:\, D^\rho f : \Omega \to \mathbf{R} \text{ stetig für alle Multiindizes } \rho,\ |\rho| \le m\}, \\ C^\infty(\Omega) &:= \bigcap_{m \in \mathbf{N}_0} C^m(\Omega). \end{aligned}}$$

Insbesondere schreibt man $C(\Omega)$ *anstelle von* $C^0(\Omega)$*.*

Bemerkung 13.5 (a) Der Funktionenraum $C^m(\Omega)$ ist ein **Vektorraum** über dem Körper $\mathbf{R}$ unter der üblichen punktweisen Addition und der λ–Multiplikation.

(b) Für Funktionen $f \in C^m(\Omega)$ gilt die Schwarzsche Vertauschungsregel für alle partiellen Ableitungen $D^\rho f$ der Ordnung $|\rho| \le m$: Es kommt somit auf die Reihenfolge $D_1^{\rho_1} D_2^{\rho_2} \cdots D_n^{\rho_n}$ **nicht** an: □

$$D_1^{\rho_1} \cdots D_j^{\rho_j} \cdots D_k^{\rho_k} \cdots D_n^{\rho_n} f(\vec{x}) = D_1^{\rho_1} \cdots D_k^{\rho_k} \cdots D_j^{\rho_j} \cdots D_n^{\rho_n} f(\vec{x}), \quad |\rho| \le m.$$

Da eine Funktion $f \in C^m(\Omega)$ in beliebiger Reihenfolge mehrfach partiell abgeleitet werden kann, ist die "Multiplikation" $D^\rho D^\sigma$ kommutativ:

$$D^\rho D^\sigma f(\vec{x}) = D^\sigma D^\rho f(\vec{x}) \ \ \forall\, f \in C^m(\Omega), \quad |\rho| + |\sigma| \le m.$$

Allgemeiner kann man **lineare Differentialoperatoren der Ordnung** m betrachten; das sind Ausdrücke in der Form

$$\boxed{P_m(D) := \sum_{|\rho| \le m} a_\rho(\vec{x}) D^\rho, \qquad P_m(D) f(\vec{x}) = \sum_{|\rho| \le m} a_\rho(\vec{x}) (D^\rho f)(\vec{x}) \ \ \forall\, f \in C^m(\Omega).} \tag{4.4}$$

Einen Sonderfall bilden lineare Differentialoperatoren mit **konstanten** Koeffizienten $a_\rho(\vec{x}) = a_\rho = const.$ Für sie gelten ähnliche Rechenregeln wie für Polynome, *zum Beispiel:*

- **Quadratformeln**:

  $P_2(D) := D_1^2 - 2D_1D_2 + D_2^2 = (D_1 - D_2)^2$. Angewendet auf $f(x,y) := x^3y^4$ ergibt $P_2(D)f(x,y) = 6xy^4 - 24x^2y^3 + 12x^3y^2$,

- **binomische Formeln**:

$$(a_1D_1 + a_2D_2)^m = \sum_{k=0}^{m} \binom{m}{k} a_1^k a_2^{m-k} D_1^k D_2^{m-k},$$

- LAPLACE–Operator:

$$\boxed{\Delta := \sum_{j=1}^{n} D_j^2, \qquad \Delta f(\vec{x}) = \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_j^2}(\vec{x}) \ \ \forall\, f \in C^2(\Omega).}$$

Neben **skalaren** Differentialoperatoren (4.4) können formal auch **vektorielle** Differentialoperatoren betrachtet werden, *zum Beispiel*

$$\vec{\nabla} := (D_1, D_2, \ldots, D_n)^T, \qquad \vec{\nabla} f(\vec{x}) = (f_{x_1}(\vec{x}), f_{x_2}(\vec{x}), \ldots, f_{x_n}(\vec{x}))^T \ \ \forall\, f \in C^1(\Omega).$$

Definition 13.14 *Gegeben seien eine Funktion $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ und ein innerer Punkt $\vec{x}_0 \in D(f)$, in welchem f partiell differenzierbar ist. Dann heiße der Vektor*

$$\boxed{\mathrm{grad}\, f(\vec{x}_0) := (D_1 f(\vec{x}_0), D_2 f(\vec{x}_0), \ldots, D_n f(\vec{x}_0))^T \in \mathbf{R}^n} \tag{4.5}$$

der **Gradient** *von f im Punkte $\vec{x}_0$. Es gilt $\vec{\nabla} f(\vec{x}_0) = \mathrm{grad}\, f(\vec{x}_0)$, und der formale Differentialoperator*

$$\boxed{\vec{\nabla} := (D_1, D_2, \ldots, D_n)^T}$$

heiße der **Nabla–Operator**.

BSP. (13.4.8) Der Gradient von f ist für folgende Funktionen zu berechnen

$$f_1(x,y,z) := z \cdot \mathrm{arc\,tan}_H \frac{y}{x}, \ \ x \neq 0, \qquad f_2(\vec{x}) := \cos\langle \vec{a}, \vec{x}\rangle, \ \ \vec{a} = (a_1, a_2, \ldots, a_n) \text{ fest.}$$

Lösung: Man erhält mit einfacher Rechnung

$$\mathrm{grad}\, f_1(x,y,z) = \begin{bmatrix} -yz/(x^2+y^2) \\ xz/(x^2+y^2) \\ \mathrm{arc\,tan}_H \frac{y}{x} \end{bmatrix}, \qquad \mathrm{grad}\, f_2(\vec{x}) = (-\sin\langle \vec{a}, \vec{x}\rangle) \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = -\vec{a}\sin\langle \vec{a}, \vec{x}\rangle.$$

Man rechnet mit dem Nabla–Operator $\vec{\nabla}$ **formal** wie mit einem Vektor; es gelten *zum Beispiel*

$$\boxed{\begin{aligned} \langle \vec{a}, \vec{\nabla}\rangle f &= \sum_{j=1}^{n} a_j D_j f \quad \text{für jedes feste } \vec{a} \in \mathbf{R}^n, \\ \langle \vec{\nabla}, \vec{\nabla}\rangle f &= \sum_{j=1}^{n} D_j^2 f = \Delta f, \quad \text{also} \quad \langle \vec{\nabla}, \vec{\nabla}\rangle = \Delta. \end{aligned}}$$

Im Vektorraum $\mathbf{R}^3$ gilt darüber hinaus für jedes feste $\vec{a} \in \mathbf{R}^3$:

$$\boxed{(\vec{a} \times \vec{\nabla}) f = \begin{vmatrix} \vec{e}_1 & a_1 & D_1 \\ \vec{e}_2 & a_2 & D_2 \\ \vec{e}_3 & a_3 & D_3 \end{vmatrix} f = \begin{bmatrix} a_2 f_z - a_3 f_y \\ a_3 f_x - a_1 f_z \\ a_1 f_y - a_2 f_x \end{bmatrix}.}$$

Bemerkung 13.6 Die Abhängigkeit der obigen Definition 13.14 des Gradienten $\operatorname{grad} f(\vec{x})$ von der Willkür der Koordinatenwahl ist *untypisch* für Vektoren in $\mathbf{R}^n$. In der formalen Definition (4.5) erhält man bei Basis– und Koordinatenwechsel andere partielle Ableitungen. Es erhebt sich die Frage, ob dadurch auch der Vektor $\operatorname{grad} f(\vec{x})$ verändert wird. Eine Antwort werden wir in Abschnitt 13.5 geben. □

Man befreit sich durch die folgende Definition von der Willkür, partielle Ableitungen nur in Richtung der Standardbasisvektoren zu betrachten:

Definition 13.15 *Gegeben seien eine Funktion* $f \in \operatorname{Abb}(\mathbf{R}^n, \mathbf{R})$ *und ein innerer Punkt* $\vec{x}_0 \in D(f)$. *Ferner sei* $\vec{h} \in \mathbf{R}^n$, $\|\vec{h}\| = 1$, *ein* **Einheitsvektor**. *Existiert der Grenzwert*

$$\boxed{\frac{\partial f}{\partial \vec{h}}(\vec{x}_0) := \lim_{t \to 0} \frac{1}{t} \left(f(\vec{x}_0 + t\vec{h}) - f(\vec{x}_0) \right) = \frac{d}{dt} f(\vec{x}_0 + t\vec{h})|_{t=0},}$$

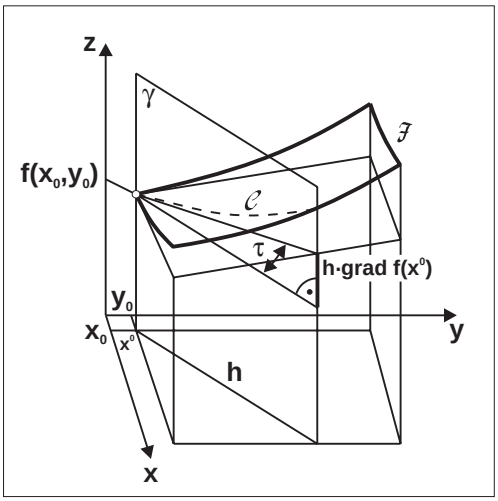
so heiße $\frac{\partial f}{\partial \vec{h}}(\vec{x}_0)$ *die* **Richtungsableitung** *von* f *im Punkte* $\vec{x}_0$ **in Richtung** $\vec{h}$.

Bemerkung 13.7 (a) Wir werden im nächsten Abschnitt, Satz 13.14, mit Hilfe des Gradienten eine einfachere Darstellung der Richtungsableitung beweisen, nämlich

$$\boxed{\frac{\partial f}{\partial \vec{h}}(\vec{x}_0) = \langle \operatorname{grad} f(\vec{x}_0), \vec{h} \rangle.} \tag{4.6}$$

(b) Im Vektorraum $\mathbf{R}^2$ gestattet die Richtungsableitung $\frac{\partial f}{\partial \vec{h}}(\vec{x}_0)$ die folgende geometrische Deutung. Wird die Fläche $\mathcal{F}$ der Funktion $z = f(x, y)$ mit der Parallelebene γ zur z–Achse durch den Punkt (x_0, y_0), die den Vektor $\vec{h}$ enthält, zum Schnitt gebracht, so erhält man eine Schnittkurve $\mathcal{C}$. Die Richtungsableitung $\frac{\partial f}{\partial \vec{h}}(\vec{x}_0)$ ist dann gerade die Steigung dieser Schnittkurve im Punkt (x_0, y_0): □

$$\tan \tau = \frac{\partial f}{\partial \vec{h}}(\vec{x}_0).$$



Zur geometrischen Deutung der Richtungsableitung

BSP. (13.4.9) Ist $\vec{h} := \vec{e}_j$ der j–te Einheitsvektor der Standardbasis, so folgt offenbar

$$\boxed{\frac{\partial f}{\partial \vec{e}_j}(\vec{x}_0) = \frac{\partial f}{\partial x_j}(\vec{x}_0), \quad j = 1, 2, \ldots, n.}$$

BSP. (13.4.10) Es seien in $\mathbf{R}^2$ die Funktion $f(x,y) := x^2 + x\cos y$ sowie der Einheitsvektor $\vec{h} := (h_1, h_2)^T$, $\sqrt{h_1^2 + h_2^2} = 1$, vorgelegt. Wir berechnen die Richtungsableitung von f in Richtung $\vec{h}$:

$$\begin{aligned}
\frac{\partial f}{\partial \vec{h}}(x,y) &= \frac{d}{dt}((x+th_1)^2 + (x+th_1)\cos(y+th_2))|_{t=0} \\
&= (2(x+th_1)h_1 + h_1\cos(y+th_2) - (x+th_1)h_2\sin(y+th_2))|_{t=0} \\
&= 2xh_1 + h_1\cos y - xh_2\sin y.
\end{aligned}$$

Andererseits gelangen wir mit

$$\operatorname{grad} f(x,y) = (2x + \cos y, -x\sin y)^T$$

und mit der Relation (4.6) wesentlich einfacher zum gleichen Resultat:

$$\frac{\partial f}{\partial \vec{h}}(x,y) = \langle \operatorname{grad} f(x,y), \vec{h}\rangle = (2x + \cos y)h_1 - xh_2\sin y.$$

13.5 Differenzierbarkeit, Ableitungen

Die bisherigen Betrachtungen über die partiellen Ableitungen einer Funktion $f \in \operatorname{Abb}(\mathbf{R}^n, \mathbf{R})$ in einem inneren Punkt $\vec{x}_0 \in D(f)$ sind mit dem Begriff der **Richtungsableitung** in $\vec{x}_0$ in Richtung $\vec{h}$ erschöpft. Da in jedem Punkt $\vec{x}_0 \in \mathbf{R}^n$ ∞–viele Richtungen vorgegeben werden können, ist die Situation völlig unbefriedigend, zumal aus der Richtungsdifferenzierbarkeit nicht einmal die Stetigkeit der Funktion f im Punkte $\vec{x}_0$ folgt. Wir suchen einen Ableitungsbegriff, der **alle** Richtungsableitungen enthält, und der wie im Falle einer skalaren Funktion $f \in \operatorname{Abb}(\mathbf{R}, \mathbf{R})$ die **Stetigkeit** von f in $\vec{x}_0$ impliziert. Es sei daran erinnert, dass im skalaren Fall $f \in \operatorname{Abb}(\mathbf{R}, \mathbf{R})$ die **Tangente**

$$\boxed{T(x) := f(x_0) + f'(x_0)\cdot(x - x_0)} \tag{5.1}$$

die einzige Lösung des Leibnizschen **Tangentenproblems** ist: Gesucht ist diejenige affine Funktion $T(x) := ax + b$, die den Graphen $G(f)$ in einem Punkte $(x_0, f(x_0))$, $x_0 \in D(f)$, *möglichst gut approximiert*:

$$f(x) = T(x) + R(x; x_0) \quad \text{mit} \quad \frac{R(x;x_0)}{x - x_0} \to 0 \text{ für } 0 < |x - x_0| \to 0, \tag{5.2}$$

man vergleiche Abschnitt 7.1. Unter Verwendung des Landau–Symbols $\mathcal{O}$ (klein oh) kann (5.2) auch in der folgenden Form geschrieben werden:

$$\boxed{f(x) = T(x) + \mathcal{O}(|x - x_0|) \quad \text{für} \quad x \to x_0.} \tag{5.3}$$

Auf dem Vektorraum $\mathbf{R}^n$ haben alle affinen Funktionen $T(\vec{x})$ durch einen Punkt $(\vec{x}_0, f(\vec{x}_0)) \in D(f) \times \mathbf{R}$ die Form

$$T(\vec{x}) = f(\vec{x}_0) + \langle \vec{a}, \vec{x} - \vec{x}_0 \rangle, \quad \vec{x} \in \mathbf{R}^n, \quad \vec{a} \in \mathbf{R}^n \text{ fest.} \tag{5.4}$$

Der Sonderfall $n = 1$ führt mit $a := f'(x_0)$ wieder auf (5.1). Durch diesen Zusammenhang wird es nahegelegt, den in (5.3) geprägten Ableitungsbegriff des Sonderfalls skalarer Funktionen auf Funktionen $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ zu übertragen:

Definition 13.16 *Gegeben seien eine Funktion $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ und ein innerer Punkt $\vec{x}_0 \in D(f)$. Genau dann heiße f in $\vec{x}_0$* **differenzierbar**, *wenn es einen Vektor $\vec{a} \in \mathbf{R}^n$ gibt mit*

$$\boxed{f(\vec{x}) = f(\vec{x}_0) + \langle \vec{a}, \vec{x} - \vec{x}_0 \rangle + \mathcal{O}(\|\vec{x} - \vec{x}_0\|) \quad \text{für} \quad \vec{x} \to \vec{x}_0.} \tag{5.5}$$

Hierbei bezeichnet $\|\vec{x} - \vec{x}_0\| = d(\vec{x}, \vec{x}_0)$ die euklidische Metrik (D3).

Da $\vec{x}_0 \in D(f)$ ein innerer Punkt ist, gibt es eine ϵ–Kugel $B_\epsilon(\vec{x}_0)$, die ganz in $D(f)$ liegt. Die Relation (5.5) macht also für alle $\vec{x} \in B_\epsilon(\vec{x}_0)$ einen Sinn.

BSP. (13.5.1) Für einen festen Vektor $\vec{0} \neq \vec{b} \in \mathbf{R}^n$ betrachten wir $f(\vec{x}) := \sin\langle \vec{b}, \vec{x} \rangle,\ \vec{x} \in \mathbf{R}^n$. Wir zeigen die Differenzierbarkeit der Funktion f in jedem Punkt $\vec{x}_0 \in \mathbf{R}^n$. Dazu sei $\vec{x} := \vec{x}_0 + \vec{h},\ \vec{h} \in \mathbf{R}^n$, gesetzt. Es gilt nun unter Verwendung des Additionstheorems der Sinus–Funktion:

$$f(\vec{x}) = f(\vec{x}_0 + \vec{h}) = \sin\langle \vec{b}, \vec{x}_0 \rangle \cdot \cos\langle \vec{b}, \vec{h} \rangle + \cos\langle \vec{b}, \vec{x}_0 \rangle \cdot \sin\langle \vec{b}, \vec{h} \rangle. \tag{5.6}$$

Beachten wir $|\langle \vec{b}, \vec{h} \rangle| \leq \|\vec{b}\|\,\|\vec{h}\| = \mathcal{O}(1)$ für $\vec{h} \to \vec{0}$ sowie $|\sin \cdots| \leq 1, |\cos \cdots| \leq 1$, so folgt aus den asymptotischen Entwicklungen

$$\sin \epsilon = \epsilon + \mathcal{O}(\epsilon), \qquad \cos \epsilon = 1 + \mathcal{O}(\epsilon) \quad \text{für} \quad \epsilon \to 0$$

zusammen mit (5.6) die Beziehung

$$f(\vec{x}) = \sin\langle \vec{b}, \vec{x}_0 \rangle + \langle \vec{b} \cdot \cos\langle \vec{b}, \vec{x}_0 \rangle, \vec{h} \rangle + \mathcal{O}(\|\vec{h}\|) \quad \text{für} \quad \vec{h} \to \vec{0}.$$

Das heißt, die Funktion f ist in $\vec{x}_0$ differenzierbar, und es gilt $\vec{a} = \vec{b} \cdot \cos\langle \vec{b}, \vec{x}_0 \rangle$.

In BSP. (13.4.7) hatten wir bereits die Richtungsableitungen der Funktion f in Richtung der Standardbasisvektoren des Vektorraums $\mathbf{R}^n$ berechnet: $f_{x_i}(\vec{x}) = b_i \cos\langle \vec{b}, \vec{x} \rangle$. Hieraus ergibt sich

$$\boxed{\mathrm{grad}\, f(\vec{x}_0) = \vec{b} \cdot \cos\langle \vec{b}, \vec{x}_0 \rangle = \vec{a},}$$

und dieser Zusammenhang ist nicht zufällig, wie wir gleich in Satz 13.14 zeigen werden. Er motiviert auch die folgende Definition.

Definition 13.17 *Der in der Beziehung* (5.5) *auftretende Vektor $\vec{a}$ heiße der* **Gradient** *der Funktion f im Punkte $\vec{x}_0 \in D(f)$ oder die* **Ableitung** *von f in $\vec{x}_0$. Wir verwenden dafür die bereits eingeführte Bezeichnung*

$$\boxed{\vec{a} = \mathrm{grad}\, f(\vec{x}_0) = \vec{\nabla} f(\vec{x}_0).}$$

Der Vektor $\mathrm{grad}\, f(\vec{x}_0) \in \mathbf{R}^n$ ist offenbar vollständig bestimmt, wenn seine Komponenten

$$\mathrm{grad}_j f(\vec{x}_0) = \langle \mathrm{grad}\, f(\vec{x}_0), \vec{e}_j \rangle, \quad j = 1, 2, \ldots, n,$$

in der Standardbasis $\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n$ bekannt sind. Eine Berechnungsvorschrift für diese Komponenten geben wir im folgenden Satz an.

Satz 13.14 *Gegeben seien eine Funktion* $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ *und ein innerer Punkt* $\vec{x}_0 \in D(f)$.

(a) *Ist* f *in* $\vec{x}_0$ *differenzierbar, so existieren die* **Richtungsableitungen** *von* f *im Punkte* $\vec{x}_0$ *in* **jeder** *Richtung* $\vec{h} \in \mathbf{R}^n$, *und es gilt*

$$\boxed{\frac{\partial f}{\partial \vec{h}}\,(\vec{x}_0) = \langle \mathrm{grad}\, f(\vec{x}_0), \vec{h}\rangle.} \tag{5.7}$$

Speziell für $\vec{h} := \vec{e}_j$ *existieren also die partiellen Ableitungen* $\frac{\partial f}{\partial x_j}\,(\vec{x}_0)$, *und es gilt*

$$\boxed{\frac{\partial f}{\partial x_j}\,(\vec{x}_0) = \langle \mathrm{grad}\, f(\vec{x}_0), \vec{e}_j\rangle = \mathrm{grad}_j f(\vec{x}_0), \quad 1 \le j \le n.} \tag{5.8}$$

Das heißt, die Ableitung $\mathrm{grad}\, f(\vec{x}_0)$ *hat in der Standardbasis des Vektorraumes* $\mathbf{R}^n$ *die Darstellung*

$$\boxed{\mathrm{grad}\, f(\vec{x}_0) = (D_1 f(\vec{x}_0), D_2 f(\vec{x}_0), \ldots, D_n f(\vec{x}_0))^T.} \tag{5.9}$$

(b) *Ist die Funktion* f *im Punkte* $\vec{x}_0$ *differenzierbar, so ist sie dort auch* **stetig**.

(c) *Besitzt* f *im Punkte* $\vec{x}_0$ **stetige** *partielle Ableitungen* $\frac{\partial f}{\partial x_j}\,(\vec{x}_0)$, $j = 1, 2, \ldots, n$, *so ist* f *in* $\vec{x}_0$ *auch differenzierbar.*

Begründungen: (a) Es sei $\vec{h} \in \mathbf{R}^n$, $\|\vec{h}\| = 1$, ein Einheitsvektor. Wir wählen $\epsilon > 0$ so, dass die ϵ–Kugel $B_\epsilon(\vec{x}_0)$ ganz in $D(f)$ liegt. Für $0 < |t| < \epsilon$ setzen wir $\vec{x} := \vec{x}_0 + t\vec{h}$ in (5.5) ein:

$$f(\vec{x}_0 + t\vec{h}) - f(\vec{x}_0) = t\,\langle \mathrm{grad}\, f(\vec{x}_0), \vec{h}\rangle + \mathcal{O}(|t|) \quad \text{für} \quad t \to 0.$$

Daraus folgt schon (5.7), und der Rest ist ohnehin klar.

(b) Wegen (5.5) haben wir

$$|f(\vec{x}) - f(\vec{x}_0)| \le \|\mathrm{grad}\, f(\vec{x}_0)\|\,\|\vec{x} - \vec{x}_0\| + \mathcal{O}(\|\vec{x} - \vec{x}_0\|) \to 0 \quad \text{für} \quad \vec{x} \to \vec{x}_0,$$

und somit die Stetigkeit von f in $\vec{x}_0$.

(c) Setzen wir $\vec{x} := \vec{x}_0 + \vec{h}$, so können wir den Mittelwertsatz der Differentialrechnung (in einer Veränderlichen) jeweils *komponentenweise* anwenden: Es existiert eine Zahl $\vartheta \in (0, 1)$ mit

$$\begin{aligned} &f(x_{01} + h_1, x_{02} + h_2, \ldots, x_{0n} + h_n) - f(x_{01}, x_{02} + h_2, \ldots, x_{0n} + h_n) \\ &\qquad = h_1 \frac{\partial f}{\partial x_1}\,(x_{01} + \vartheta h_1, x_{02} + h_2, \ldots, x_{0n} + h_n). \end{aligned}$$

Da f_{x_1} nach Voraussetzung in $\vec{x}_0$ stetig ist, gilt sicher

$$h_1 f_{x_1}(x_{01} + \vartheta h_1, x_{02} + h_2, \ldots, x_{0n} + h_n) = h_1 f_{x_1}(\vec{x}_0) + \mathcal{O}(\|\vec{h}\|) \quad \text{für} \quad \vec{h} \to \vec{0},$$

und somit

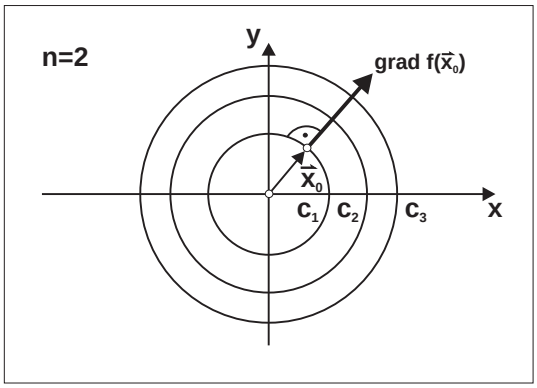
$$f(\vec{x}_0 + \vec{h}) = h_1 f_{x_1}(\vec{x}_0) + f(x_{01}, x_{02} + h_2, \ldots, x_{0n} + h_n) + \mathcal{O}(\|\vec{h}\|).$$

Behandelt man die anderen Variablen in entsprechender Weise, so erhält man nach $n - 1$ weiteren Schritten

$$f(\vec{x}_0 + \vec{h}) = f(\vec{x}_0) + \sum_{j=1}^{n} h_j f_{x_j}(\vec{x}_0) + \mathcal{O}(\|\vec{h}\|) = f(\vec{x}_0) + \langle \mathrm{grad}\, f(\vec{x}_0), \vec{h}\rangle + \mathcal{O}(\|\vec{h}\|).$$

Also ist f in $\vec{x}_0$ differenzierbar. □

Bemerkung 13.8 Die Teile (b) und (c) des soeben bewiesenen Satzes liefern die in Satz 13.12 behauptete Stetigkeitsaussage. Gleichzeitig liefert die Aussage (c) ein einfaches **Kriterium** für die Differenzierbarkeit einer Funktion f in einem Punkte $\vec{x}_0$. Mit Hilfe der Relationen (5.7) und (5.9) können nun alle Ableitungen von f in einfacher Weise berechnet werden. □



Die Äquipotentialflächen der Funktion $f(\vec{x}) := \ln \|\vec{x}\|$

BSP. (13.5.2) Es sei $r := \|\vec{x}\| = \Big(\sum_{j=1}^{n} |x_j|^2\Big)^{1/2}$ für $\vec{x} \in \mathbf{R}^n$ gesetzt. Wir betrachten die Funktion $f(\vec{x}) := \ln r$, $r > 0$. Es gilt:

$$D_j f(\vec{x}) = \frac{1}{r}\,\frac{\partial r}{\partial x_j} = \frac{x_j}{r^2} \;\; \forall\, 1 \le j \le n, \qquad \operatorname{grad} f(\vec{x}) = \frac{\vec{x}}{r^2}, \;\; r > 0.$$

Die Richtungsableitung von f in Richtung eines beliebigen Einheitsvektors $\vec{h} \in \mathbf{R}^n$ ist dann

$$\boxed{\frac{\partial f}{\partial \vec{h}}\,(\vec{x}_0) = \frac{1}{r^2}\,\langle \vec{x}_0, \vec{h}\rangle.}$$

Beachte: Die Äquipotentialflächen $f(\vec{x}) = c$ sind die konzentrischen Sphären $r = \|\vec{x}\| = e^c > 0$ mit Mittelpunkt $\vec{0}$. Da der Vektor $\operatorname{grad} f(\vec{x})$ die Richtung $\vec{x}$ hat, steht $\operatorname{grad} f(\vec{x}_0)$ **senkrecht** auf der Äquipotentialfläche durch den Punkt $\vec{x}_0$. Dass dies kein Zufall ist, werden wir weiter unten begründen. Es gilt ferner

$$\langle \vec{x}_0, \vec{h}\rangle = \|\vec{x}_0\| \cos \sphericalangle\,(\vec{x}_0, \vec{h}),$$

so dass das Skalarprodukt seinen größten Wert annimmt, wenn der Vektor $\vec{h}$ in die Richtung von $\vec{x}_0$ fällt. Mit anderen Worten, die Richtungsableitung $\frac{\partial f}{\partial \vec{h}}\,(\vec{x}_0)$ in $\vec{x}_0$ wird am **größten**, wenn $\vec{h}$ in die Richtung des Gradienten $\operatorname{grad} f(\vec{x}_0)$ fällt.

BSP. (13.5.3) Wir betrachten für festes $\vec{0} \neq \vec{b} \in \mathbf{R}^n$ die Funktion $f(\vec{x}) := e^{\langle \vec{b}, \vec{x}\rangle}$, $\vec{x} \in \mathbf{R}^n$. Es gilt hier:

$$D_j f(\vec{x}) = b_j\, e^{\langle \vec{b}, \vec{x}\rangle} \;\; \forall\, 1 \le j \le n, \qquad \operatorname{grad} f(\vec{x}) = \vec{b}\, e^{\langle \vec{b}, \vec{x}\rangle}.$$

Hieraus resultiert die Richtungsableitung in Richtung eines beliebigen Einheitsvektors $\vec{h} \in \mathbf{R}^n$

$$\boxed{\frac{\partial f}{\partial \vec{h}}\,(\vec{x}_0) = \langle \vec{b}, \vec{h}\rangle\, e^{\langle \vec{b}, \vec{x}_0\rangle}.}$$

Die Äquipotentialflächen $f(\vec{x}) = c > 0$ sind offenbar die Hyperebenen $\langle \vec{b}, \vec{x}\rangle = \ln c$. Hier steht der Normalenvektor $\vec{b}$ senkrecht auf der Hyperebene. Da der Vektor $\vec{b}$ wieder die Richtung des Gradienten $\operatorname{grad} f(\vec{x})$ hat, gilt wie im vorangegangenen Beispiel, dass $\operatorname{grad} f(\vec{x}_0)$ **senkrecht** auf der Äquipotentialfläche von f durch den Punkt $\vec{x}_0$ steht. Ebenso erkennt man, dass die Richtungsableitung $\frac{\partial f}{\partial \vec{h}}\,(\vec{x}_0)$ am größten wird, wenn $\vec{h}$ in Richtung des Vektors $\vec{b}$ und somit in Richtung von $\operatorname{grad} f(\vec{x}_0)$ fällt.

Satz 13.15 *Gegeben seien eine Funktion* $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ *sowie ein innerer Punkt* $\vec{x}_0 \in D(f)$. *Ist* f *in* $\vec{x}_0$ *differenzierbar und gilt* $\operatorname{grad} f(\vec{x}_0) \neq \vec{0}$, *so ist im Punkte* $\vec{x}_0$ *die* **Richtung des steilsten Anstiegs** *von* f *durch den Vektor* $\operatorname{grad} f(\vec{x}_0)$ *bestimmt.*

Begründung: Für eine beliebige Richtung $\vec{h} \in \mathbf{R}^n$, $\|\vec{h}\| = 1$, gilt

$$\frac{\partial f}{\partial \vec{h}}(\vec{x}_0) = \langle \operatorname{grad} f(\vec{x}_0), \vec{h} \rangle = \|\operatorname{grad} f(\vec{x}_0)\| \cos \sphericalangle (\operatorname{grad} f(\vec{x}_0), \vec{h}).$$

Hieraus folgt

$$-\|\operatorname{grad} f(\vec{x}_0)\| \leq \frac{\partial f}{\partial \vec{h}}(\vec{x}_0) \leq \|\operatorname{grad} f(\vec{x}_0)\|,$$

und für $\vec{h} := \frac{\operatorname{grad} f(\vec{x}_0)}{\|\operatorname{grad} f(\vec{x}_0)\|}$ gilt genau $\frac{\partial f}{\partial \vec{h}}(\vec{x}_0) = \|\operatorname{grad} f(\vec{x}_0)\|$. □

BSP. (13.5.4) Es seien ein räumliches Temperaturfeld $T(x, y, z) := xz - y^2 + 1$ und dazu ein Punkt P_0 mit Ortsvektor $\vec{x}_0 := (-1, -1, 1)^T \in \mathbf{R}^3$ gegeben. Gesucht ist diejenige Richtung $\vec{h}$, in welcher sich die Temperatur am stärksten ändert, wenn man von P_0 ausgeht. Sodann ist der maximale Temperaturanstieg in P_0 zu bestimmen.

Lösung: Gemäß Satz 13.15 gilt

$$\vec{h} = \frac{\operatorname{grad} T(\vec{x}_0)}{\|\operatorname{grad} T(\vec{x}_0)\|}, \qquad \operatorname{grad} T(\vec{x}_0) = \begin{bmatrix} z \\ -2y \\ x \end{bmatrix}\Big|_{\vec{x}_0} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix},$$

und hiermit

$$\vec{h} = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} \quad \text{sowie} \quad \frac{\partial T}{\partial \vec{h}}(\vec{x}_0) = \|\operatorname{grad} T(\vec{x}_0)\| = \sqrt{6}$$

für die maximale Temperaturzunahme.

Der Ableitungsoperator "grad" ist den Rechenregeln der gewöhnlichen Differentiation unterworfen. Er genügt also der Summen–, Produkt– und Kettenregel, sofern dafür geeignete Funktionen vorgelegt sind. Die folgenden Beziehungen können sehr einfach komponentenweise nachgeprüft werden:

Satz 13.16 *Gegeben seien Funktionen* $f, g \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ *sowie ein innerer Punkt* $\vec{x}_0 \in D(f) \cap D(g)$. *Es seien* f *und* g *in* $\vec{x}_0$ *differenzierbar. Ferner seien* $\vec{x} \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R}^n)$ *und* $h \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R})$ *differenzierbar in den Punkten* t_0 *bzw.* $f(\vec{x}_0)$, *und es gelte* $\vec{x}(t_0) = \vec{x}_0$. *Dann sind die folgenden Funktionen in* $\vec{x}_0$ *bzw. in* t_0 *differenzierbar:*

$$\lambda f + \mu g \;\forall\, \lambda, \mu \in \mathbf{R}, \qquad f \cdot g, \qquad h \circ f, \qquad f \circ \vec{x}.$$

Es gelten die folgenden

Ableitungsregeln		
(a)	$\operatorname{grad}(\lambda f + \mu g)(\vec{x}_0) = \lambda \operatorname{grad} f(\vec{x}_0) + \mu \operatorname{grad} g(\vec{x}_0)$	Linearität
(b)	$\operatorname{grad}(f \cdot g)(\vec{x}_0) = g(\vec{x}_0) \operatorname{grad} f(\vec{x}_0) + f(\vec{x}_0) \operatorname{grad} g(\vec{x}_0)$	Produktregel
(c)	$\operatorname{grad} h(f(\vec{x}_0)) = h'(f(\vec{x}_0)) \cdot \operatorname{grad} f(\vec{x}_0)$	1.Kettenregel
(d)	$\frac{d}{dt} f(\vec{x}(t_0)) = \langle \operatorname{grad} f(\vec{x}_0), \dot{\vec{x}}(t_0) \rangle$	2.Kettenregel

(5.10)

BSP. (13.5.5) Eine im Ursprung O des $\mathbf{R}^3$ angebrachte punktförmige **elektrische Ladung** Q erzeugt das **elektrische Potential**

$$\varphi(\vec{x}) := \frac{Q}{4\pi\epsilon r}, \qquad r := \|\vec{x}\| > 0.$$

Hierin sind Q und ϵ Konstanten. Die durch Q erzeugte **elektrische Feldstärke** $\vec{E}$ ist für $\vec{x} \neq \vec{0}$ wie folgt erklärt:

$$\vec{E} := -\operatorname{grad}\varphi(\vec{x}) \overset{(5.10(c))}{=} -\varphi'(r)\cdot\operatorname{grad} r = \frac{Q}{4\pi\epsilon r^2}\,\frac{\vec{x}}{r} = \frac{Q}{4\pi\epsilon r^3}\,\vec{x}.$$

Hier ist die Vektorfunktion $\vec{E}$ ganz offensichtlich eine Funktion von mehreren reellen Veränderlichen $\vec{x}$: $\vec{E} \in \operatorname{Abb}(\mathbf{R}^3, \mathbf{R}^3)$.

Definition 13.18 (a) *Physikalisch relevante Abbildungen* $\varphi \in \operatorname{Abb}(\mathbf{R}^n, \mathbf{R})$ *heißen* **Skalarfelder**; *physikalisch relevante Abbildungen* $\vec{E} \in \operatorname{Abb}(\mathbf{R}^n, \mathbf{R}^m)$, $m > 1$, *heißen* **Vektorfelder**.

(b) *Gibt es zu einem Vektorfeld* $\vec{E}$ *ein Skalarfeld* φ *mit*

$$\boxed{\vec{E} = -\operatorname{grad}\varphi,}$$

so heiße φ *ein* **Potential** *von* $\vec{E}$.

Die Frage, welche Vektorfelder $\vec{E}$ ein Potential φ besitzen, soll hier nicht erörtert werden. Dies ist ein Thema der **Vektoranalysis**.

BSP. (13.5.6) Es sei $f \in \operatorname{Abb}(\mathbf{R}^2, \mathbf{R})$ die Funktion $f(x,y) := x^3 - xy + y^3$, und es sei $\vec{x} \in \operatorname{Abb}(\mathbf{R}, \mathbf{R}^2)$ die Einheitskreislinie $\vec{x}(t) := (\cos t, \sin t)^T$, $t \in \mathbf{R}$. Dann gelten

$$f(\vec{x}(t)) = \cos^3 t - \cos t \sin t + \sin^3 t, \quad \frac{d}{dt} f(\vec{x}(t)) = (3\cos^2 t - \sin t)(-\sin t) + (3\sin^2 t - \cos t)\cos t.$$

Unter Verwendung der Ableitungsregel (5.10(d)) gelangt man zum gleichen Resultat:

$$\begin{aligned} \frac{d}{dt} f(\vec{x}(t)) &= \langle \operatorname{grad} f(\vec{x}), \dot{\vec{x}}(t)\rangle = \left\langle \left[\begin{array}{c} 3x^2 - y \\ 3y^2 - x \end{array}\right]_{|_{\vec{x}(t)}}, \left[\begin{array}{c} -\sin t \\ \cos t \end{array}\right] \right\rangle \\ &= (3\cos^2 t - \sin t)(-\sin t) + (3\sin^2 t - \cos t)\cos t. \end{aligned}$$

Bemerkung 13.9 (a) Unter Verwendung des Nabla–Operators $\vec{\nabla}$ können die Ableitungsregeln (5.10) in der folgenden Weise formalisiert werden:

$$\boxed{\begin{array}{lcl} \vec{\nabla}(\lambda f + \mu g) &=& \lambda\vec{\nabla} f + \mu\vec{\nabla} g, \\ \vec{\nabla}(f \cdot g) &=& g\vec{\nabla} f + f\vec{\nabla} g, \\ \vec{\nabla}(h \circ f) &=& \dfrac{dh}{df}\,\vec{\nabla} f. \end{array}}$$

Wegen $\frac{\partial f}{\partial \vec{h}} = \Big(\sum\limits_{j=1}^{n} h_j D_j\Big) f$ folgt noch

$$\boxed{\frac{\partial f}{\partial \vec{h}} = \langle \vec{h}, \vec{\nabla}\cdot\rangle\, f.}$$

(b) Man nennt häufig den Ausdruck

$$\boxed{\frac{d}{dt} f(\vec{x}(t)) = \langle \operatorname{grad} f(\vec{x}(t)), \dot{\vec{x}}(t) \rangle}$$

die **Wegableitung** von f **entlang des Weges** $\vec{x}(t)$. Offensichtlich ist die Wegableitung nichts anderes als die Richtungsableitung von f im Punkte $\vec{x}(t)$ in Richtung $\dot{\vec{x}}(t)$, sofern $\|\dot{\vec{x}}(t)\| = 1$ angenommen wird. Umgekehrt ist eine Richtungsableitung von f im Punkte $\vec{x}_0$ in Richtung $\vec{h}$ eine spezielle Wegableitung entlang des (geradlinigen) Weges $\vec{x}(t) := \vec{x}_0 + t\vec{h}$. Das Konzept der Wegableitungen ist jedoch allgemeiner: Zum Beispiel muss eine *gekrümmte* Fläche F neben dem Punkt $\vec{x}_0 \in F$ keine weiteren Punkte der Geraden $\vec{x}(t) = \vec{x}_0 + t\vec{h}$ enthalten. Gilt $D(f) = F$, so kann in diesem Fall die Richtungsableitung $\frac{\partial f}{\partial \vec{h}}(\vec{x}_0)$ nicht gebildet werden. Hingegen existieren – unter geeigneten Differenzierbarkeitsvoraussetzungen – die Wegableitungen der Funktion f für jeden Weg $\vec{x}(t) \in F$. □

Wir zeigen nun unter Verwendung dieser Bemerkung:

Satz 13.17 *Gegeben seien eine Funktion* $f \in \operatorname{Abb}(\mathbf{R}^n, \mathbf{R})$ *und ein innerer Punkt* $\vec{x}_0 \in D(f)$. *Ist* f *differenzierbar in* $\vec{x}_0$, *so steht der Vektor* $\operatorname{grad} f(\vec{x}_0)$ **senkrecht** *auf der Äquipotentialfläche von* f *durch den Punkt* $\vec{x}_0$.

Begründung: Die ÄP der Funktion f durch den Punkt $\vec{x}_0$ sei implizit durch die Gleichung $f(\vec{x}) = c_0 = const$ gegeben. Wir betrachten einen beliebigen differenzierbaren Weg $\vec{x}(t)$ auf dieser ÄP mit $\vec{x}(t_0) = \vec{x}_0$. Wegen $c_0 = f(\vec{x}(t))$ erhalten wir aus (5.10(d)):

$$0 = \frac{d}{dt} f(\vec{x}(t)) = \langle \operatorname{grad} f(\vec{x}_0), \dot{\vec{x}}(t_0) \rangle.$$

Der Vektor $\operatorname{grad} f(\vec{x}_0)$ steht also senkrecht auf dem Tangentenvektor $\dot{\vec{x}}(t_0)$ und somit senkrecht auf der ÄP durch den Punkt $\vec{x}_0$. □

BSP. (13.5.7) Die Funktion $f \in \operatorname{Abb}(\mathbf{R}^3, \mathbf{R})$ sei für Konstanten $a, b, c > 0$ wie folgt definiert:

$$\boxed{f(x,y,z) := \left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 + \left(\frac{z}{c}\right)^2 - 1.}$$

Die Äquipotentialfläche $f(x,y,z) = 0$ ist ein **Ellipsoid** E, und es gilt $\vec{x}_0 = (x_0, y_0, z_0) \in E$ genau für

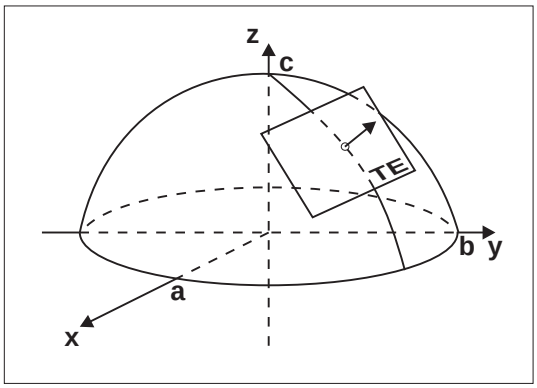
$$\left(\frac{x_0}{a}\right)^2 + \left(\frac{y_0}{b}\right)^2 + \left(\frac{z_0}{c}\right)^2 = 1.$$

Gemäß Satz 13.17 ist der **Normalenvektor** $\vec{n}$ in einem Punkt $\vec{x}_0$ an die Fläche E gerade der Gradient $\operatorname{grad} f(\vec{x}_0)$:

$$\vec{n} = \operatorname{grad} f(\vec{x}_0) = \left(\frac{2x_0}{a^2}, \frac{2y_0}{b^2}, \frac{2z_0}{c^2}\right)^T.$$

Die Normale $\vec{n}$ legt genau eine **Ebene** TE durch den Punkt $\vec{x}_0$ fest. In dieser Ebene liegen alle Vektoren senkrecht zu $\vec{n}$, die in $\vec{x}_0$ angeheftet sind. Insbesondere liegt der Tangentenvektor $\dot{\vec{x}}(t_0)$ eines beliebigen differenzierbaren Weges $\vec{x}(t) \in E$ in der Ebene TE, sofern $\vec{x}(t_0) = \vec{x}_0$ gilt. Die HESSE–Normalform von TE lautet $\langle \operatorname{grad} f(\vec{x}_0), \vec{x} - \vec{x}_0 \rangle = 0$, wobei natürlich die Nebenbedingung $f(\vec{x}_0) = 0$ erfüllt sei muss. Dies führt auf die allgemeine Ebenengleichung für TE, nämlich

$$\frac{2x_0}{a^2} x + \frac{2y_0}{b^2} y + \frac{2z_0}{c^2} z = \frac{2x_0^2}{a^2} + \frac{2y_0^2}{b^2} + \frac{2z_0^2}{c^2} = 2.$$



Normale und Tangentialebene in einem Punkt des Ellipsoids

Wegen der oben erklärten Eigenschaften der Ebene TE definiert man:

Definition 13.19 *Gegeben seien eine Funktion* $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ *und ein innerer Punkt* $\vec{x}_0 \in D(f)$. *Ist* f *differenzierbar in* $\vec{x}_0$ *und gilt* $\operatorname{grad} f(\vec{x}_0) \neq \vec{0}$, *so heiße die Hyperebene*

$$\boxed{TE := \{\vec{x} \in \mathbf{R}^n \,:\, \langle \operatorname{grad} f(\vec{x}_0), \vec{x} - \vec{x}_0 \rangle = 0\}} \tag{5.11}$$

die **Tangentialhyperebene** *an die Äquipotentialfläche von* f *durch den Punkt* $\vec{x}_0$. *In den Fällen* $f \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R})$ *und* $f \in \mathrm{Abb}\,(\mathbf{R}^3, \mathbf{R})$ *spricht man einfacher von der* **Tangentengerade** *bzw. der* **Tangentialebene** *im Punkte* $\vec{x}_0$.

Ist die Fläche $G \subset \mathbf{R}^3$ in der **expliziten** Darstellung $z = f(x, y)$ gegeben, so ist G die Äquipotentialfläche der Funktion $g(x, y, z) := f(x, y) - z$ mit der Gleichung $g = 0$. Die **Tangentialebene** an G in einem Punkt $\vec{x}_0 := (x_0, y_0, z_0)$ mit $z_0 := f(x_0, y_0)$ hat gemäß (5.11) die Form

$$\boxed{TE: \quad z - z_0 = f_x(x_0, y_0) \cdot (x - x_0) + f_y(x_0, y_0) \cdot (y - y_0), \qquad z_0 = f(x_0, y_0).} \tag{5.12}$$

Demgemäß ist der **Normalenvektor** $\vec{n} \perp G$ im Punkte $\vec{x}_0$ an die ÄP $0 = g(x, y, z) := f(x, y) - z$ der Vektor

$$\boxed{\vec{n} = \operatorname{grad} g(\vec{x}_0) = \begin{bmatrix} f_x(x_0, y_0) \\ f_y(x_0, y_0) \\ -1 \end{bmatrix}.} \tag{5.13}$$

BSP. (13.5.8) Gegeben seien die Funktion $f \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R})$ und der Punkt $(x_0, y_0) \in D(f) := \mathbf{R}^2$ gemäß

$$f(x, y) := (x^2 - 3x)\cos(\pi y) + (y - 5)\sin\Big(\frac{\pi x}{2}\Big), \qquad (x_0, y_0) := (3, 5).$$

Man bestimme die HESSE–Normalform der Tangentialebene an f in (x_0, y_0).

Lösung: Die durch den Graph $G(f) \subset \mathbf{R}^3$ definierte räumliche Fläche G ist die Äquipotentialfläche $g(x, y, z) = 0$ der Funktion $g(x, y, z) := f(x, y) - z$. Wird $z_0 := f(x_0, y_0) = f(3, 5) = 0$ gesetzt, so ist $\vec{x}_0 := (x_0, y_0, z_0) = (3, 5, 0)$ ein Punkt dieser ÄP, und der Normalenvektor $\vec{n}$ der Tangentialebene an die ÄP in diesem Punkt $\vec{x}_0$ ist gemäß (5.13) der Vektor $\vec{n} = (f_x(x_0, y_0), f_y(x_0, y_0), -1)^T$ mit

$$f_x(x, y) = (2x - 3)\cos(\pi y) + \frac{\pi(y - 5)}{2}\cos\Big(\frac{\pi x}{2}\Big), \quad f_y(x, y) = -\pi(x^2 - 3x)\sin(\pi y) + \sin\Big(\frac{\pi x}{2}\Big).$$

Hieraus resultieren $f_x(3,5) = -3$, $f_y(3,5) = -1$, und somit bei richtiger Vorzeichenwahl der Einheitsnormalenvektor

$$\vec{n}_0 = \frac{1}{\sqrt{11}}\,(3,1,1)^T.$$

Die gesuchte HNF der Tangentialebene lautet nun

$$\langle \vec{x}, \vec{n}_0 \rangle = \langle \vec{x}_0, \vec{n}_0 \rangle = \frac{14}{\sqrt{11}} = d(\vec{0}, TE).$$

Bemerkung 13.10 Offenbar haben Flächen $G \subset \mathbf{R}^3$ mit **expliziter** Darstellung $z = f(x,y)$ unter hinreichenden Differenzierbarkeitsvoraussetzungen in jedem Flächenpunkt $(x_0, y_0, z_0) = \vec{x}_0 \in G$ einen Normalenvektor $\vec{n}$, da der Vektor $\operatorname{grad}(f(x_0,y_0) - z_0)$ nirgends verschwinden kann. Im allgemeinen sind aber Flächen $G \subset \mathbf{R}^3$ als Äquipotentialflächen einer Funktion $g \in \operatorname{Abb}(\mathbf{R}^3, \mathbf{R})$ in **impliziter** Darstellung $g(x,y,z) = c$ vorgelegt. Die Frage, ob aus dieser Gleichung eine explizite Darstellung $z = f(x,y)$ gewonnen werden kann, wird uns noch im Zusammenhang mit dem Problem der impliziten Funktionen beschäftigen. □

13.6 Das vollständige oder totale Differential

Für eine gegebene Funktion $f \in \operatorname{Abb}(\mathbf{R}^n, \mathbf{R})$, die in einem inneren Punkt $\vec{x}_0 \in D(f)$ differenzierbar sei, sollen Aussagen über die **Veränderung** der Funktionswerte von f in einer Umgebung des Punktes $\vec{x}_0$ getroffen werden.

Bei **skalaren** Funktionen $f \in \operatorname{Abb}(\mathbf{R}, \mathbf{R})$ wird häufig mit den Differentialen $\frac{df}{dx} = f'(x_0)$ wie mit Zahlen des Körpers $\mathbf{R}$ gerechnet. Das Differential

$$df = f'(x_0)\,dx \tag{6.1}$$

gestattet die folgende geometrische Interpretation. Wird die Tangente $T(x)$ im Punkt $x_0 \in D(f)$ betrachtet,

$$T(x) = f(x_0) + f'(x_0)\,(x - x_0), \quad x \in \mathbf{R},$$

so ist der **Tangentenzuwachs** bei Änderung von x um den Wert h **relativ zu** h eine Konstante:

$$\frac{T(x+h) - T(x)}{h} = f'(x_0) = \frac{df}{dx}.$$

Wir schreiben diese Relation in der folgenden Form:

$$\boxed{T(x+h) - T(x) = \lambda\, df \quad \text{und} \quad h = \lambda\, dx.} \tag{6.2}$$

Bei Funktionen $f \in \operatorname{Abb}(\mathbf{R}^n, \mathbf{R})$ gilt als Analogon zur Tangente im Punkte $\vec{x}_0$ die affine Funktion

$$T(\vec{x}) = f(\vec{x}_0) + \langle \operatorname{grad} f(\vec{x}_0), \vec{x} - \vec{x}_0 \rangle,$$

und in Anlehnung an (6.2) setzen wir

$$\lambda\, df := T(\vec{x} + \vec{h}) - T(\vec{x}) = \langle \operatorname{grad} f(\vec{x}_0), \vec{h} \rangle, \qquad \vec{h} \neq \vec{0}. \tag{6.3}$$

Um die Konsistenz der Bezeichnungen zu gewährleisten, muss der Spezialfall $f(\vec{x}) := x_j$, $j = 1, 2, \ldots, n$, wiederum auf das Differential $df = dx_j$ führen. Da in diesem Fall $\operatorname{grad} f(\vec{x}_0) =$

$\vec{e}_j = const$ gilt, erhalten wir aus dem Ansatz (6.3) die Beziehung $\lambda dx_j = h_j$, $1 \le j \le n$, und somit nach Kürzung durch λ den Ausdruck

$$\boxed{df = \langle \operatorname{grad} f(\vec{x}_0), d\vec{x} \rangle, \quad d\vec{x} := (dx_1, dx_2, \ldots, dx_n)^T \in \mathbf{R}^n.} \tag{6.4}$$

Definition 13.20 *Der durch die Relation* (6.4) *definierte Ausdruck df heiße das zum* **Zuwachs** $d\vec{x} \in \mathbf{R}^n$ *gehörende* **vollständige** *oder* **totale Differential** *der Funktion f im Punkte* $\vec{x}_0$. *Es ist ein Maß für die Änderung des Funktionswertes der affinen Funktion* $T(\vec{x})$ *bei Übergang von einem Punkte* $\vec{x} \in \mathbf{R}^n$ *zum Punkt* $\vec{x} + d\vec{x}$.

Mit dieser Erklärung kann eine Angabe darüber gemacht werden, wie sich $T(\vec{x})$ in einer Umgebung des Punktes $\vec{x}_0 \in D(f)$ verhält, wenn wir von $\vec{x}_0$ auf $T(\vec{x})$ nach $\vec{x}_0 + d\vec{x}$ schreiten. Verwenden wir andererseits die Definition (5.5) der Ableitung von f in $\vec{x}_0$, so resultiert

$$\Delta f := f(\vec{x}_0 + d\vec{x}) - f(\vec{x}_0) = \langle \operatorname{grad} f(\vec{x}_0), d\vec{x} \rangle + \mathcal{O}(\|d\vec{x}\|) \quad \text{für} \quad d\vec{x} \to \vec{0},$$

und somit die Beziehung

$$\boxed{\Delta f = df + \mathcal{O}(\|d\vec{x}\|) \quad \text{für} \quad d\vec{x} \to \vec{0}.}$$

Das totale Differential df ist eine **lineare Näherung** für die Änderung des Funktionswertes der Funktion f in einer Umgebung des Punktes $\vec{x}_0 \in D(f)$. Deshalb wird die **Näherung** $\Delta f \approx df$ für **Fehlerrechnungen** verwendet:

$$\boxed{\Delta f \approx \langle \operatorname{grad} f(\vec{x}_0), \Delta\vec{x} \rangle = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\vec{x}_0) \cdot \Delta x_j.} \tag{6.5}$$

Ist also eine abgeleitete Größe $f = f(x_1, x_2, \ldots, x_n)$ gegeben, deren Daten $x_1, x_2, \ldots, x_n$ mit möglichen Messfehlern Δx_j behaftet sind, so ist eine Schranke des Messfehlers an f durch die Relation (6.5) bestimmt:

$$|\Delta f| \le \sum_{j=1}^{n} \Big|\frac{\partial f}{\partial x_j}(\vec{x}_0)\Big| \, |\Delta x_j|. \tag{6.6}$$

Man nennt die Größen $|\Delta f|$ und $\left|\frac{\Delta f}{f}\right|$ den **absoluten** bzw. den **relativen Fehler** von f. Die Ausdrücke

$$\sum_{j=1}^{n} \Big|\frac{\partial f}{\partial x_j}(\vec{x}_0)\Big| \, |\Delta x_j| \quad \text{und} \quad \sum_{j=1}^{n} \Big|\frac{\partial f}{\partial x_j}(\vec{x}_0)\Big| \, \Big|\frac{\Delta x_j}{f}\Big|$$

heißen **Schranken** für den **maximalen absoluten** bzw. den **maximalen relativen Fehler**.

Häufig treten Funktionen in der speziellen Form $f(\vec{x}) := a\, x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$ auf. Diese haben die folgende Schranke für den maximalen relativen Fehler:

$$\boxed{\Big|\frac{\Delta f}{f}\Big| \le \sum_{j=1}^{n} |\alpha_j| \, \Big|\frac{\Delta x_j}{x_j}\Big|.}$$

BSP. (13.6.1) Der Wert der idealen Gaskonstanten R ist aus der Zustandsgleichung eines idealen Gases der Masse 1 durch Messung von *Druck* p, *Volumen* V und *Temperatur* T zu berechnen. Gemessen werden die Werte p_0, V_0, T_0, die aber infolge von Messungenauigkeiten mit Fehlern Δp, ΔV, ΔT behaftet sind:

$$p = p_0 \pm \Delta p, \quad V = V_0 \pm \Delta V, \quad T = T_0 \pm \Delta T.$$

Man finde Schranken für den maximalen absoluten bzw. den maximalen relativen Fehler von R.

Lösung: Aus der Zustandsgleichung eines idealen Gases erhalten wir $R = f(p,V,T) = \frac{pV}{T}$, und somit gemäß (6.5)

$$\Delta R \approx \frac{V_0}{T_0}\,\Delta p + \frac{p_0}{T_0}\,\Delta V - \frac{p_0 V_0}{T_0^2}\,\Delta T, \qquad |\Delta R| \le \Big|\frac{V_0\Delta p}{T_0}\Big| + \Big|\frac{p_0\Delta V}{T_0}\Big| + \Big|\frac{p_0 V_0 \Delta T}{T_0^2}\Big|.$$

Eine Schranke für den maximalen relativen Fehler erhalten wir mit der Spezifikation $x_1 := p$, $x_2 := V$, $x_3 := T$, $\alpha_1 := 1$, $\alpha_2 := 1$, $\alpha_3 := -1$ aus obiger Formel:

$$\Big|\frac{\Delta R}{R}\Big| \le \Big|\frac{\Delta p}{p_0}\Big| + \Big|\frac{\Delta V}{V_0}\Big| + \Big|\frac{\Delta T}{T_0}\Big|.$$

BSP. (13.6.2) Die Hintereinanderschaltung zweier Kondensatoren mit den Kapazitäten C_1, C_2 hat eine Gesamtkapazität

$$C = \Big(\frac{1}{C_1} + \frac{1}{C_2}\Big)^{-1} = \frac{C_1 C_2}{C_1 + C_2}.$$

Es seien die Werte $C_1 := (200 \pm 1)\mu F$ und $C_2 := (300 \pm 1.5)\mu F$ vorgegeben. Man bestimme Schranken für die beiden maximalen Fehler von C.

Lösung: Mit den Werten $C_1^o = 200\mu F$, $C_2^o = 300\mu F$, $|\Delta C_1| = 1\mu F$, $|\Delta C_2| = 1.5\mu F$ erhalten wir aus (6.5):

$$\Delta C \approx \frac{(C_2^o)^2}{(C_1^o + C_2^o)^2}\,\Delta C_1 + \frac{(C_1^o)^2}{(C_1^o + C_2^o)^2}\,\Delta C_2, \qquad |\Delta C| \le \Big[\Big(\frac{3}{5}\Big)^2 \cdot 1 + \Big(\frac{2}{5}\Big)^2 \cdot 1.5\Big]\mu F = 0.6\mu F.$$

Ganz analog folgt für den relativen Fehler:

$$\frac{\Delta C}{C} \approx \frac{C_2^o}{C_1^o(C_1^o + C_2^o)}\,\Delta C_1 + \frac{C_1^o}{C_2^o(C_1^o + C_2^o)}\,\Delta C_2, \qquad \Big|\frac{\Delta C}{C}\Big| \le \Big(\frac{3}{2\cdot 500} + \frac{2\cdot 1.5}{3\cdot 500}\Big) = \frac{5}{1000} = 0.5\%.$$

Schließlich resultiert aus

$$C^o = \frac{C_1^o C_2^o}{C_1^o + C_2^o} = \frac{200\cdot 300}{500}\mu F = 120\mu F$$

als Ergebnis der Rechnung die Gesamtkapazität $C = (120 \pm 0.6)\mu F$.

13.7 Mittelwertsatz und TAYLORsche Formel

Der TAYLORsche Satz 7.19 für Funktionen einer reellen Veränderlichen $f \in \mathrm{Abb}\,(\mathbf{R},\mathbf{R})$ besagt, dass sich die Funktion f und das Polynom

$$T_m(x) := \sum_{k=0}^{m} \frac{1}{k!}\, f^{(k)}(x_0)\cdot(x - x_0)^k, \qquad x \in \mathbf{R}, \tag{7.1}$$

in einem inneren Punkt $x_0 \in D(f)$ **von der Ordnung** $m \in \mathbf{N}$ **berühren**, sofern f in x_0 stetige Ableitungen bis zur Ordnung $m+1$ besitzt. Das heißt, es gilt in einer Umgebung des Punktes x_0 die Gleichung $f(x) = T_m(x) + R_m(x;x_0)$, worin R_m das LAGRANGEsche Restglied bezeichnet:

$$R_m(x;x_0) := \frac{f^{(m+1)}(\xi)}{(m+1)!}\,(x - x_0)^{m+1}, \qquad \xi := x_0 + \vartheta(x - x_0), \quad \vartheta \in (0,1). \tag{7.2}$$

Der Sonderfall $m = 0$ führt auf den **Mittelwertsatz**:

$$f(x) - f(x_0) = f'(\xi) \cdot (x - x_0), \qquad \xi := x_0 + \vartheta(x - x_0), \;\; \vartheta \in (0,1). \tag{7.3}$$

Nachfolgend zeigen wir, in welcher Weise diese Aussagen auf Funktionen $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ übertragbar sind. Wir beginnen mit einem Mittelwertsatz (MWS) für Funktionen mehrerer Veränderlicher.

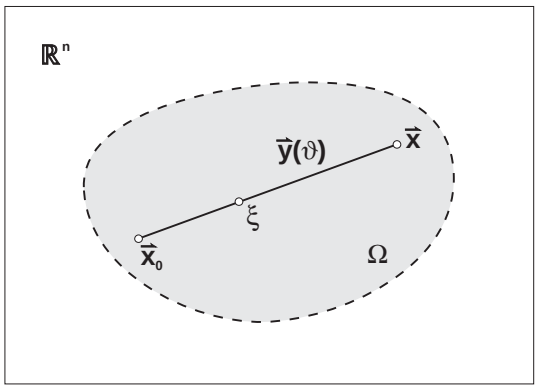
Satz 13.18 (Mehrdimensionaler MWS)
Gegeben seien ein Gebiet $\Omega \subset \mathbf{R}^n$ und eine Funktion $f \in C^1(\Omega)$ sowie zwei Punkte $\vec{x}, \vec{x}_0 \in \Omega$ mit der Eigenschaft: Die **Verbindungsstrecke** $\vec{y}(\vartheta) := \vec{x}_0 + \vartheta(\vec{x} - \vec{x}_0)$, $\vartheta \in [0,1]$, *liegt ganz in Ω. Dann gilt für einen Zwischenwert $\vec{\xi} := \vec{x}_0 + \vartheta(\vec{x} - \vec{x}_0)$, $\vartheta \in (0,1)$, die Gleichung*

$$\boxed{f(\vec{x}) - f(\vec{x}_0) = \langle \mathrm{grad}\, f(\vec{\xi}), \vec{x} - \vec{x}_0 \rangle.} \tag{7.4}$$

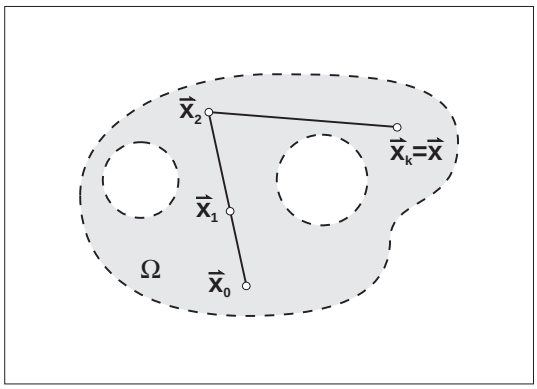
Begründung: Wir betrachten die durch $g(t) := f(\vec{x}_0 + t(\vec{x} - \vec{x}_0))$ definierte Funktion $g \in C^1([0,1])$. Wegen (7.3) gilt nun für ein $\vartheta \in (0,1)$

$$g(1) - g(0) = g'(\vartheta),$$

und daraus folgt unter Verwendung der 2.Kettenregel (5.10(d)) die behauptete Gleichung (7.4). □



Zum Mittelwertsatz 13.18 für Funktionen mehrerer Veränderlicher



Skizze zum Beweis von Satz 13.19

Wie bei Funktionen einer reellen Veränderlichen $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R})$ können aus dem Mittelwertsatz Kriterien für die Konstanz von Funktionen $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ hergeleitet werden.

Satz 13.19 *Gegeben seien ein Gebiet $\Omega \subset \mathbf{R}^n$ und eine Funktion $f \in C^1(\Omega)$. Dann gilt*

$$\boxed{\mathrm{grad}\, f(\vec{x}) = \vec{0} \;\; \forall\, \vec{x} \in \Omega \quad \Leftrightarrow \quad f = const \text{ in } \Omega.} \tag{7.5}$$

Begründung: Da die Implikation $''\Leftarrow''$ trivial ist, beweisen wir nur die Richtung $''\Rightarrow''$. Gelte also $\mathrm{grad}\, f(\vec{x}) = \vec{0} \;\; \forall\, \vec{x} \in \Omega$. Je zwei Punkte $\vec{x}_0, \vec{x} \in \Omega$ können in Ω wegen der Gebietseigenschaft durch einen Polygonzug mit Eckpunkten $\vec{x}_0, \vec{x}_1, \ldots, \vec{x}_k = \vec{x}$ stetig verbunden werden. Es gilt wegen (7.4) auf jeder Kante des Polygonzugs:

$$f(\vec{x}_j) - f(\vec{x}_{j-1}) = \langle \mathrm{grad}\, f(\vec{\xi}_j), \vec{x}_j - \vec{x}_{j-1} \rangle = 0, \quad 1 \le j \le k.$$

Also folgt $f(\vec{x}) = f(\vec{x}_0) = const$ für alle Punkte $\vec{x} \in \Omega$. □

BSP. (13.7.1) Auf dem *positiven Quadranten* $\Omega := \{(x,y) : x > 0, \quad y > 0\} \subset \mathbf{R}^2$ betrachten wir die Funktion

$$f(x,y) := \operatorname{arc\,tan}_H \frac{y}{x} + \operatorname{arc\,tan}_H \frac{x}{y}, \quad (x,y) \in \Omega.$$

Man berechnet auf der Menge Ω die partiellen Ableitungen

$$f_x(x,y) = \frac{1}{1+(y/x)^2}\left(-\frac{y}{x^2}\right) + \frac{1}{1+(x/y)^2}\,\frac{1}{y} = 0, \quad f_y(x,y) = \frac{1}{1+(y/x)^2}\,\frac{1}{x} + \frac{1}{1+(x/y)^2}\left(-\frac{x}{y^2}\right) = 0,$$

und somit $\operatorname{grad} f(x,y) = \vec{0}$ für alle $(x,y) \in \Omega$. Aus Satz 13.19 erschließen wir deshalb $f(x,y) = const = f(1,1) = 2\operatorname{arc\,tan}_H 1 = \frac{\pi}{2}$.

Wird in (7.4) $\vec{x} := \vec{x}_0 + \vec{h}$ gesetzt, so ergibt sich im Fall $n = 2$ als einfachste Verallgemeinerung der TAYLOR–Formel für ein geeignetes $\vartheta \in (0,1)$ und für $\vec{h} := (h,k)^T$ der Ausdruck

$$\boxed{f(x_0+h, y_0+k) = f(x_0,y_0) + f_x(x_0+\vartheta h, y_0+\vartheta k)\cdot h + f_y(x_0+\vartheta h, y_0+\vartheta k)\cdot k.}$$

Es lässt sich hier bereits erahnen, dass der schreibtechnische Aufwand sehr groß wird, wenn weitere Variable und höhere Ableitungen hinzukommen. Man gewinnt mehr Ökonomie in der Bezeichnungsweise, wenn **Multiindizes** verwendet werden. Im weiteren Verlauf unserer Überlegungen wird es erforderlich sein, die k–fache Ableitung einer Funktion $f \in C^k(\Omega)$ in Richtung eines Vektors $\vec{h} \neq \vec{0}$ zu bilden, worin $\vec{h} \in \mathbf{R}^n$ nicht notwendig ein Einheitsvektor ist.

Wir bedienen uns dabei des formalen Differentialoperators $\langle \vec{h}, \vec{\nabla}\cdot\rangle$:

$$\boxed{\langle \vec{h}, \vec{\nabla}\cdot\rangle\, f(\vec{x}) := \langle \vec{h}, \operatorname{grad} f(\vec{x})\rangle = \sum_{j=1}^{n} h_j D_j f(\vec{x}) = \sum_{|\rho|=1} \vec{h}^{\rho} D^{\rho} f(\vec{x}),}$$

für den wir das folgende Hilfsresultat herleiten:

Satz 13.20 *Gegeben seien ein Gebiet $\Omega \subset \mathbf{R}^n$ und eine Funktion $f \in C^m(\Omega)$. Dann gilt für jede Zahl $k \in \mathbf{N}_0$ mit $0 \le k \le m$:*

$$\boxed{\langle \vec{h}, \vec{\nabla}\cdot\rangle^k f(\vec{x}) = \sum_{|\rho|=k} \binom{k}{\rho} \vec{h}^{\rho} D^{\rho} f(\vec{x}), \quad \vec{x} \in \Omega.} \tag{7.6}$$

Hierin ist im Fall $|\rho| = k$ der **Polynomialkoeffizient** $\binom{k}{\rho}$ *gemäß folgender Vorschrift zu bilden:*

$$\binom{k}{\rho} := \frac{k!}{\rho_1!\rho_2!\cdots\rho_n!}, \quad |\rho| = k, \quad \rho = (\rho_1, \rho_2, \ldots, \rho_n) \in \mathbf{N}_0^n.$$

Begründung: Wir beweisen die Formel (7.6) für $k \ge 1$ durch vollständige Induktion nach der Raumdimension n und beginnen mit der

Induktionsverankerung $n = 2$: Es gilt für jedes feste $k \in \mathbf{N}$ unter Verwendung des binomischen Lehrsatzes:

$$\langle \vec{h}, \vec{\nabla}\cdot\rangle^k = (h_1 D_1 + h_2 D_2)^k = \sum_{j=0}^{k} \binom{k}{j} h_1^j h_2^{k-j} D_1^j D_2^{k-j}.$$

Mit $\rho_1 := j$ und $\rho_2 := k - j$ gelten nun $|\rho| = k$ und $\binom{k}{j} = \frac{k!}{\rho_1!\rho_2!} = \binom{k}{\rho}$ sowie

$$\langle \vec{h}, \vec{\nabla}\cdot\rangle^k = \sum_{|\rho|=k} \binom{k}{\rho} \vec{h}^{\rho} D^{\rho}. \tag{7.7}$$

Vererbung: Es gelte (7.6) nun bereits für alle Raumdimensionen $\leq n-1$. Wir vollziehen den Induktionsschluss auf die Raumdimension n und setzen dazu $\rho' := (\rho_1, \rho_2, \ldots, \rho_{n-1})$ sowie $S := h_1 D_1 + \cdots + h_{n-1} D_{n-1}$. Dann folgt aus der Gleichung (7.7)

$$\langle \vec{h}, \vec{\nabla} \cdot \rangle^k = (S + h_n D_n)^k = \sum_{j+\rho_n=k} \frac{k!}{j!\rho_n!} S^j h_n^{\rho_n} D_n^{\rho_n},$$

während die Induktionsannahme schon

$$S^j = \sum_{|\rho'|=j} \binom{j}{\rho'} \vec{h}'^{\rho'} D'^{\rho'}, \qquad \vec{h}' := (h_1, h_2, \ldots, h_{n-1}), \quad D'^{\rho'} := D_1^{\rho_1} D_2^{\rho_2} \cdots D_{n-1}^{\rho_{n-1}},$$

liefert. Setzt man dies ein, so resultiert die behauptete Relation (7.6):

$$\langle \vec{h}, \vec{\nabla} \cdot \rangle^k = \sum_{j+\rho_n=k} \sum_{|\rho'|=j} \frac{k!}{j!\rho_n!} \frac{j!}{\rho_1!\rho_2! \cdots \rho_{n-1}!} \vec{h}'^{\rho'} h_n^{\rho_n} D'^{\rho'} D_n^{\rho_n} = \sum_{|\rho|=k} \binom{k}{\rho} \vec{h}^{\rho} D^{\rho}.$$

Nach diesen Vorbetrachtungen sind wir jetzt in der Lage, die TAYLORsche Formel für Funktionen mehrerer Variabler zu beweisen.

Satz 13.21 (von der TAYLORschen Formel)
Gegeben seien ein Gebiet $\Omega \subset \mathbf{R}^n$ und eine Funktion $f \in C^{m+1}(\Omega)$ sowie zwei Punkte $\vec{x}, \vec{x}_0 \in \Omega$ mit der Eigenschaft: Die **Verbindungsstrecke** *$\vec{y}(\vartheta) := \vec{x}_0 + \vartheta(\vec{x} - \vec{x}_0)$, $\vartheta \in [0,1]$, liegt ganz in Ω. Dann gilt mit $\vec{h} := \vec{x} - \vec{x}_0$ die* TAYLOR**–Formel**

$$\boxed{f(\vec{x}) = f(\vec{x}_0) + \sum_{k=1}^{m} \frac{1}{k!} \langle \vec{h}, \vec{\nabla} \cdot \rangle^k f(\vec{x}_0) + R_m(\vec{x}; \vec{x}_0),} \tag{7.8}$$

worin $R_m(\vec{x}; \vec{x}_0)$ das LAGRANGE**sche Restglied** *bezeichnet:*

$$\boxed{R_m(\vec{x}; \vec{x}_0) := \frac{1}{(m+1)!} \langle \vec{h}, \vec{\nabla} \cdot \rangle^{m+1} f(\vec{x}_0 + \vartheta \vec{h}), \quad \vartheta \in (0,1).} \tag{7.9}$$

Begründung: Wir betrachten für $t \in [0,1]$ die Funktion $\varphi(t) := f(\vec{x}_0 + t\vec{h})$. Dann ist auf $\varphi(t)$ der TAYLORsche Satz 7.19 anwendbar. Mit den Spezifikationen $t := 1$ und $t_0 := 0$ gilt also

$$\varphi(1) = \varphi(0) + \sum_{k=1}^{m} \frac{\varphi^{(k)}(0)}{k!} 1^k + \frac{\varphi^{(m+1)}(\vartheta)}{(m+1)!}, \quad \vartheta \in (0,1).$$

Unter Verwendung der 2.Kettenregel (5.10(d)) berechnet man nun:

$$\begin{aligned}
\frac{d}{dt}\varphi(t) &= \langle \operatorname{grad} f(\vec{x}_0 + t\vec{h}), \vec{h} \rangle = \langle \vec{h}, \vec{\nabla} \cdot \rangle\, f(\vec{x}_0 + t\vec{h}), \\
\frac{d^2}{dt^2}\varphi(t) &= \langle \vec{h}, \vec{\nabla} \cdot \rangle \frac{d}{dt} f(\vec{x}_0 + t\vec{h}) = \langle \vec{h}, \vec{\nabla} \cdot \rangle^2 f(\vec{x}_0 + t\vec{h}), \\
&\vdots \\
\frac{d^k}{dt^k}\varphi(t) &= \langle \vec{h}, \vec{\nabla} \cdot \rangle^k f(\vec{x}_0 + t\vec{h}).
\end{aligned}$$

Werden hier $t = 0$ bzw. $t = \vartheta$ eingesetzt, so gewinnt man aus der obigen TAYLOR–Formel die behauptete Form

$$f(\vec{x}_0 + \vec{h}) = f(\vec{x}_0) + \sum_{k=1}^{m} \frac{1}{k!} \langle \vec{h}, \vec{\nabla} \cdot \rangle^k f(\vec{x}_0) + \frac{1}{(m+1)!} \langle \vec{h}, \vec{\nabla} \cdot \rangle^{m+1} f(\vec{x}_0 + \vartheta \vec{h}).$$

Sonderfall: Funktionen in 2 Veränderlichen

Für Funktionen $f = f(x,y)$ in **zwei** Veränderlichen hat das TAYLOR–Polynom vom Grade $m \in \mathbf{N}$ im Entwicklungspunkt (x_0, y_0) gemäß (7.8) die Form

$$\boxed{T_m(x,y) = f(x_0,y_0) + \sum_{j=1}^{m} \frac{1}{j!} \langle \vec{h}, \vec{\nabla} \cdot \rangle^j f(x_0,y_0).}$$

Dabei ist $\vec{h} \in \mathbf{R}^2$ der Vektor

$$\boxed{\vec{h} := (h,k)^T \quad \text{mit} \quad h := x - x_0, \;\; k := y - y_0.}$$

Die Koeffizienten $\langle \vec{h}, \vec{\nabla} \cdot \rangle^j f(x_0,y_0)$, $1 \le j \le m$, des TAYLOR–Polynoms gestatten nun die Darstellung

$$\boxed{\langle \vec{h}, \vec{\nabla} \cdot \rangle^j f(x_0,y_0) = \Big(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\Big)^j f(x_0,y_0) = \sum_{r=0}^{j} \binom{j}{r} h^r k^{j-r} D_x^r D_y^{j-r} f(x_0,y_0).}$$

Wir schreiben diese für $j = 1,2,3$ in expliziten Formeln auf:

$$\boxed{\begin{aligned}
\langle \vec{h}, \vec{\nabla} \cdot \rangle f(x_0,y_0) &= h\, f_x(x_0,y_0) + k f_y(x_0,y_0), \\
\langle \vec{h}, \vec{\nabla} \cdot \rangle^2 f(x_0,y_0) &= h^2 f_{xx}(x_0,y_0) + 2hk f_{xy}(x_0,y_0) + k^2 f_{yy}(x_0,y_0), \\
\langle \vec{h}, \vec{\nabla} \cdot \rangle^3 f(x_0,y_0) &= h^3 f_{xxx}(x_0,y_0) + 3h^2 k f_{xxy}(x_0,y_0) + 3hk^2 f_{xyy}(x_0,y_0) + k^3 f_{yyy}(x_0,y_0).
\end{aligned}}$$

BSP. (13.7.2) Man bestimme das TAYLOR–Polynom 3.Grades der Funktion $f(x,y) := x^3 + xy^2 + y^3$ im Entwicklungspunkt $(x_0,y_0) := (1,2)$.

Lösung: Wir setzen $h := x - 1$ und $k := y - 2$. Es gelten hier

$$\begin{aligned}
&f(x_0,y_0) = 13, \\
&f_x(x_0,y_0) = 3x_0^2 + y_0^2 = 7, \quad f_y(x_0,y_0) = 2x_0y_0 + 3y_0^2 = 16, \\
&f_{xx}(x_0,y_0) = 6x_0 = 6, \quad f_{xy}(x_0,y_0) = 2y_0 = 4, \quad f_{yy}(x_0,y_0) = 2x_0 + 6y_0 = 14, \\
&f_{xxx}(x_0,y_0) = 6, \quad f_{xxy}(x_0,y_0) = 0, \quad f_{xyy}(x_0,y_0) = 2, \;\; f_{yyy}(x_0,y_0) = 6.
\end{aligned}$$

Es resultiert nun aus den obigen Formeln:

$$\begin{aligned}
\langle \vec{h}, \vec{\nabla} \cdot \rangle f(x_0,y_0) &= h\, f_x(x_0,y_0) + k f_y(x_0,y_0) = 7h + 16k, \\
\langle \vec{h}, \vec{\nabla} \cdot \rangle^2 f(x_0,y_0) &= h^2 f_{xx}(x_0,y_0) + 2hk f_{xy}(x_0,y_0) + k^2 f_{yy}(x_0,y_0) = 6h^2 + 8hk + 14k^2, \\
\langle \vec{h}, \vec{\nabla} \cdot \rangle^3 f(x_0,y_0) &= h^3 f_{xxx}(x_0,y_0) + 3h^2 k f_{xxy}(x_0,y_0) + 3hk^2 f_{xyy}(x_0,y_0) + k^3 f_{yyy}(x_0,y_0) \\
&= 6h^3 + 6hk^2 + 6k^3,
\end{aligned}$$

und somit das gesuchte TAYLOR–Polynom

$$T_3(x,y) = 13 + \frac{1}{1!}(7h + 16k) + \frac{1}{2!}(6h^2 + 8hk + 14k^2) + \frac{1}{3!}(6h^3 + 6hk^2 + 6k^3).$$

Bemerkung 13.11 (a) Unter Verwendung der Cauchy–Schwarz–Ungleichung erhält man aus der Darstellung (7.6) die folgende Abschätzung für die Koeffizienten des Taylor–Polynoms:

$$\begin{aligned} |\langle \vec{h}, \vec{\nabla} \cdot \rangle^k f(\vec{x})| &= \Big| \sum_{|\rho|=k} \binom{k}{\rho} \vec{h}^\rho D^\rho f(\vec{x}) \Big| \le \Big(\sum_{|\rho|=k} \binom{k}{\rho} |D^\rho f(\vec{x})|^2 \Big)^{1/2} \Big(\sum_{|\rho|=k} \binom{k}{\rho} \vec{h}^{2\rho} \Big)^{1/2} \\ &\le \Big(\sum_{|\rho|=k} \binom{k}{\rho} |D^\rho f(\vec{x})|^2 \Big)^{1/2} \|\vec{h}\|^k. \end{aligned}$$

Dabei haben wir die folgende Identität benutzt: Wird in (7.6) an die Stelle des Vektors $\vec{\nabla} f(\vec{x})$ der Vektor $\vec{h}$ gesetzt, so ergibt sich

$$\langle \vec{h}, \vec{h} \rangle^k = \|\vec{h}\|^{2k} = \sum_{|\rho|=k} \binom{k}{\rho} \vec{h}^\rho \vec{h}^\rho = \sum_{|\rho|=k} \binom{k}{\rho} \vec{h}^{2\rho}.$$

Mit der obigen Ungleichung können wir nun das Lagrange–Restglied (7.9) in der folgenden Weise abschätzen.

$$\boxed{|R_m(\vec{x}; \vec{x}_0)| \le \max_{0 \le \vartheta \le 1} \Big(\sum_{|\rho|=m+1} \binom{m+1}{\rho} |D^\rho f(\vec{x}_0 + \vartheta \vec{h})|^2 \Big)^{1/2} \frac{\|\vec{h}\|^{m+1}}{(m+1)!}.} \tag{7.10}$$

Liegt die Vollkugel $\overline{B_R}(\vec{x}_0)$ noch ganz in dem Gebiet Ω, so existiert die Konstante

$$C_m := \frac{1}{(m+1)!} \max_{\vec{x} \in \overline{B_R}(\vec{x}_0)} \Big(\sum_{|\rho|=m+1} \binom{m+1}{\rho} |D^\rho f(\vec{x})|^2 \Big)^{1/2},$$

und es gilt also die Restgliedabschätzung

$$|R_m(\vec{x}; \vec{x}_0)| \le C_m \, \|\vec{h}\|^{m+1} \quad \forall \, \|\vec{h}\| \le R.$$

Das heißt, das Taylor–Polynom vom Grade m

$$T_m(\vec{x}) = f(\vec{x}_0) + \sum_{k=1}^{m} \frac{1}{k!} \langle \vec{h}, \vec{\nabla} \cdot \rangle^k f(\vec{x}_0), \quad \vec{h} := \vec{x} - \vec{x}_0,$$

approximiert in einer Umgebung des Entwicklungspunktes $\vec{x}_0$ den Funktionswert $f(\vec{x})$ von der Ordnung $\|\vec{h}\|^{m+1}$:

$$|f(\vec{x}) - T_m(\vec{x})| = \mathcal{O}(\|\vec{h}\|^{m+1}) \quad \text{für} \quad \|\vec{h}\| \to 0.$$

(b) Hat die Funktion f die Regularität $f \in C^\infty(\Omega)$ und gilt $\lim\limits_{m \to \infty} R_m(\vec{x}; \vec{x}_0)$ in einer Umgebung des Entwicklungspunktes $\vec{x}_0$, so besitzt f im Punkte $\vec{x}_0$ die Taylor–Reihe

$$f(\vec{x}) = f(\vec{x}_0) + \sum_{k=1}^{\infty} \frac{1}{k!} \langle \vec{h}, \vec{\nabla} \cdot \rangle^k f(\vec{x}_0). \tag{7.11}$$

Ganz analog zu Satz 9.12 gilt als hinreichendes Kriterium für die Existenz der Taylor–Reihe (7.11) in der Vollkugel $\overline{B_R}(\vec{x}_0)$ die gleichmäßige Beschränktheit der Ableitungen: □

$$\exists \, M \, : \, \max_{\vec{x} \in \overline{B_R}(\vec{x}_0)} \Big(\sum_{|\rho|=k} \binom{k}{\rho} |D^\rho f(\vec{x})|^2 \Big)^{1/2} \le M \quad \forall \, k \in \mathbf{N}_0.$$

BSP. (13.7.3) Für reelle Zahlen $\alpha \neq 0 \neq \beta$ betrachten wir die Funktion $f(x,y) := \cos(\alpha x) \cdot \sin(\beta y)$. Gesucht ist ihre TAYLOR–Reihe im Entwicklungspunkt $(x_0, y_0) := (0,0)$.

Lösung: Natürlich wird man hier **nicht** die TAYLORsche Formel anwenden, sondern das CAUCHY–Produkt von Cosinus– und Sinus–Reihe bilden:

$$f(x,y) = \Big(\sum_{k=0}^{\infty} \frac{(-1)^k(\alpha x)^{2k}}{(2k)!}\Big)\Big(\sum_{k=0}^{\infty} \frac{(-1)^k(\beta y)^{2k+1}}{(2k+1)!}\Big) = \sum_{k=0}^{\infty}(-1)^k \sum_{j=0}^{k} \frac{(\alpha x)^{2j}}{(2j)!}\,\frac{(\beta y)^{2(k-j)+1}}{(2(k-j)+1)!}.$$

Dabei ist die absolute Konvergenz für alle $x, y \in \mathbf{R}$ gewährleistet.

13.8 Extremwertaufgaben für Funktionen in mehreren Veränderlichen

In Abschnitt 7.7 wurde die TAYLOR–Formel dazu verwendet, um Aussagen über die Extremwerte von Funktionen einer reellen Veränderlichen $f \in \mathrm{Abb}(\mathbf{R}, \mathbf{R})$ herzuleiten. Es liegt jetzt nahe, auch im Fall von Funktionen mehrerer Veränderlicher $f \in \mathrm{Abb}(\mathbf{R}^n, \mathbf{R})$ Kriterien für Extremwerte unter Zuhilfenahme der TAYLOR–Formel zu gewinnen.

BSP. (13.8.1) Das **Ellipsoid** in $\mathbf{R}^3$ ist *implizit* durch die Gleichung

$$g(x,y,z) := \Big(\frac{x}{a}\Big)^2 + \Big(\frac{y}{b}\Big)^2 + \Big(\frac{z}{c}\Big)^2 - 1 = 0, \quad a>0,\, b>0,\, c>0,$$

definiert. Diese Gleichung lässt sich über dem Definitionsbereich

$$D(f) := \{(x,y) \in \mathbf{R}^2 \,:\, (\frac{x}{a})^2 + (\frac{y}{b})^2 \leq 1\}$$

explizit nach z auflösen: Man erhält für das obere und das untere Halbellipsoid jeweils eine **lokale** Darstellung

$$z = f^{\pm}(x,y) := \pm c\sqrt{1 - \Big((\frac{x}{a})^2 + (\frac{y}{b})^2\Big)}, \quad (x,y) \in D(f).$$

Offensichtlich sind die Scheitelpunkte $(0,0,\pm c)$ Extremalpunkte von $f^{\pm}$:

$$z = f^+(x,y) \leq f^+(0,0) = c \;\forall\, (x,y) \in D(f) \quad \Rightarrow \quad (0,0) \text{ ist ein Maximum für } f^+,$$
$$z = f^-(x,y) \geq f^-(0,0) = -c \;\forall\, (x,y) \in D(f) \quad \Rightarrow \quad (0,0) \text{ ist ein Minimum für } f^-.$$

In den Extremalpunkten liegen die Tangentialebenen an die Flächen $z = f^{\pm}(x,y)$ parallel zur (x,y)–Ebene, was aus der Anschauung folgt. Der Normalenvektor $\vec{n}$ der Tangentialebenen fällt also in Richtung des Standardbasisvektors $\vec{e}_z$:

$$\vec{n} = \begin{bmatrix} f_x^{\pm}(0,0) \\ f_y^{\pm}(0,0) \\ -1 \end{bmatrix} = -\vec{e}_z = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}.$$

Wir haben somit in den Extremalpunkten der Funktionen $f^{\pm}$ die Bedingung $\mathrm{grad}\, f^{\pm}(0,0) = \vec{0}$ vorliegen. Dass dies kein Zufall ist, werden wir im folgenden aufzeigen.

Definition 13.21 *Gegeben seien eine Funktion $f \in \mathrm{Abb}(\mathbf{R}^n, \mathbf{R})$ und ein Punkt $\vec{x}_0 \in D(f)$. Gibt es zu $\vec{x}_0$ eine offene ϵ–Kugel $B_\epsilon(\vec{x}_0)$ derart, dass die Bedingung*

$$\boxed{f(\vec{x}) - f(\vec{x}_0) \leq 0 \quad (\text{bzw. } \geq 0) \quad \forall\, \vec{x} \in B_\epsilon(\vec{x}_0) \cap D(f)} \tag{8.1}$$

erfüllt ist, so besitze f im Punkte $\vec{x}_0$ ein **relatives Maximum** *(bzw. ein* **relatives Minimum***). Man fasst beide Begriffe zusammen zum Begriff des* **relativen Extremums***.*

Die Erkenntnis über das Verschwinden des Gradienten $\operatorname{grad} f(\vec{x}_0)$ in Extremalpunkten $\vec{x}_0 \in D(f)$ wird durch folgenden Satz bestätigt:

Satz 13.22 *Gegeben seien eine Funktion $f \in \operatorname{Abb}(\mathbf{R}^n, \mathbf{R})$ und ein innerer Punkt $\vec{x}_0 \in D(f)$. Ist f in $\vec{x}_0$ differenzierbar und besitzt die Funktion f dort ein relatives Extremum, so gilt notwendig* $\operatorname{grad} f(\vec{x}_0) = \vec{0}$.

Begründung: In $\vec{x}_0$ existiert die Richtungsableitung $\frac{\partial f}{\partial \vec{h}}(\vec{x}_0) = \langle \operatorname{grad} f(\vec{x}_0), \vec{h} \rangle$ für alle Richtungen $\vec{h} \in \mathbf{R}^n$, $\|\vec{h}\| = 1$. Liegt die ϵ–Kugel $B_\epsilon(\vec{x}_0)$ noch ganz in $D(f)$, so setzen wir $\vec{x} := \vec{x}_0 + t\vec{h}$ mit $|t| < \epsilon$. Dann folgt aus der Definition der Ableitung

$$f(\vec{x}) - f(\vec{x}_0) = t\langle \operatorname{grad} f(\vec{x}_0), \vec{h} \rangle + \mathcal{O}(|t|) \quad \text{für} \quad t \to 0.$$

Wäre $\operatorname{grad} f(\vec{x}_0) \neq \vec{0}$, so könnten wir $\vec{h} := \operatorname{grad} f(\vec{x}_0)/\|\operatorname{grad} f(\vec{x}_0)\|$ wählen und erhielten

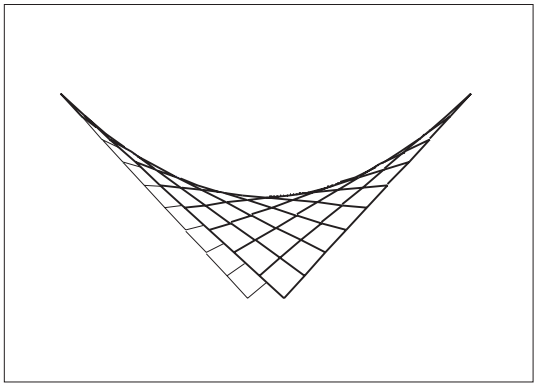
$$f(\vec{x}) - f(\vec{x}_0) = t\|\operatorname{grad} f(\vec{x}_0)\| + \mathcal{O}(|t|).$$

Die rechte Seite hat aber je nach Wahl von $t \gtrless 0$ positive und negative Werte, was der Definition eines relativen Extremums bei $\vec{x}_0$ widerspricht. □

BSP. (13.8.2) Es sei $f \in \operatorname{Abb}(\mathbf{R}^2, \mathbf{R})$ die durch $f(x, y) := xy$ definierte Funktion. (Die Fläche $z = f(x, y)$ ist ein sogenannter *Affensattel.*) Ganz offensichtlich gilt

$$\operatorname{grad} f(\vec{x}_0) = (y_0, x_0)^T \stackrel{!}{=} \vec{0} \quad \Leftrightarrow \quad (x_0, y_0) = (0, 0).$$

Das heißt, ein relatives Extremum von f kann höchstens bei $\vec{x}_0 = \vec{0}$ liegen. Tatsächlich liefert aber eine Skizze der Fläche die Einsicht, dass im Punkt $\vec{0}$ **kein** Extremum liegt, sondern ein **Sattelpunkt**. Die Bedingung $\operatorname{grad} f(\vec{x}_0) = \vec{0}$ ist in der Tat nur **notwendig**, nicht aber hinreichend für die Existenz relativer Extrema!



Die Fläche der Funktion $f(x, y) := xy$

Definition 13.22 *Jeder Punkt $\vec{x}_0 \in D(f)$ einer Funktion $f \in \operatorname{Abb}(\mathbf{R}^n, \mathbf{R})$ mit $\operatorname{grad} f(\vec{x}_0) = \vec{0}$ heiße ein* **kritischer Punkt** *von f. Jeder kritische Punkt von f, der nicht gleichzeitig ein relatives Extremum ist, heiße ein* **Sattelpunkt** *von f.*

BSP. (13.8.3) Es sei $f \in \operatorname{Abb}(\mathbf{R}^2, \mathbf{R})$ die durch $f(x, y) := -x^4 + x^2y^2 + y^3 + y^2$ definierte Funktion. Wir berechnen ihre kritischen Punkte.

$$\operatorname{grad} f(x, y) = \begin{bmatrix} -4x^3 + 2xy^2 \\ 2x^2y + 3y^2 + 2y \end{bmatrix} = \begin{bmatrix} 2x(y^2 - 2x^2) \\ 2y(x^2 + 1 + 1.5y) \end{bmatrix} \stackrel{!}{=} \vec{0}.$$

Aus der ersten Gleichung resultiert als Lösung $x_0 = 0$ und/oder $y_0^2 = 2x_0^2$. Setzen wir diese Werte in die zweite Gleichung ein, so sind die folgenden Fallunterscheidungen zu treffen:

- $x_0 = 0$ liefert die zwei kritischen Punkte

$$\vec{x}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \vec{x}_2 = \begin{bmatrix} 0 \\ -\frac{2}{3} \end{bmatrix},$$

- $x_0 \neq 0$ und $y_0 = \pm\sqrt{2}\, x_0$ liefert vier weitere kritische Punkte, nämlich

$$\vec{x}_3 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -1 \end{bmatrix}, \quad \vec{x}_4 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ -1 \end{bmatrix}, \quad \vec{x}_5 = \begin{bmatrix} \sqrt{2} \\ -2 \end{bmatrix}, \quad \vec{x}_6 = \begin{bmatrix} -\sqrt{2} \\ -2 \end{bmatrix}.$$

Um herauszufinden, welche der kritischen Punkte relative Extrema bzw. Sattelpunkte sind, untersuchen wir wie bei Funktionen einer Veränderlichen die zweiten Ableitungen. Dazu nehmen wir an, f sei in einem kritischen Punkt hinreichend oft differenzierbar. Gelte nun $\operatorname{grad} f(\vec{x}_0) = \vec{0}$ in einem inneren Punkt $\vec{x}_0 \in D(f)$. Aus dem Satz 13.21 von der TAYLORschen Formel erhalten wir

$$f(\vec{x}_0 + \vec{h}) - f(\vec{x}_0) = \frac{1}{2!} \langle \vec{h}, \vec{\nabla} \cdot \rangle^2 f(\vec{x}_0) + \mathcal{O}(\|\vec{h}\|^2) \quad \text{für} \quad \|\vec{h}\| \to 0. \tag{8.2}$$

Es gilt hier

$$\langle \vec{h}, \vec{\nabla} \cdot \rangle^2 f(\vec{x}_0) = \sum_{j=1}^{n} \sum_{k=1}^{n} h_j h_k D_j D_k f(\vec{x}_0) =: \langle H(\vec{x}_0)\vec{h}, \vec{h} \rangle,$$

worin die Matrix $H(\vec{x}_0) \in \mathbf{R}^{(n,n)}$ wie folgt definiert ist:

$$H = H(\vec{x}_0) = (H_{jk}(\vec{x}_0)) := \Big(\frac{\partial^2 f}{\partial x_j \partial x_k} (\vec{x}_0) \Big).$$

Für Funktionen $f \in C^2(\Omega)$ gilt in jedem Punkt $\vec{x}_0 \in \Omega$ die SCHWARZsche Vertauschungsregel $H_{jk}(\vec{x}_0) = H_{kj}(\vec{x}_0)$; in diesem Fall ist H eine **symmetrische** Matrix: $H = H^T$.

Definition 13.23 *Für eine gegebene Funktion $f \in C^2(\Omega)$ heiße die symmetrische Matrix*

$$\boxed{H = H(\vec{x}_0) := \Big(\frac{\partial^2 f}{\partial x_j \partial x_k} (\vec{x}_0) \Big)}$$

die HESSE–**Matrix** *von f im Punkte $\vec{x}_0 \in D(f) = \Omega$.*

Wir setzen jetzt

$$Q(\vec{h}) := \langle H(\vec{x}_0)\vec{h}, \vec{h} \rangle, \quad \vec{h} \in \mathbf{R}^n,$$

in (8.2) ein. Dann resultiert

$$f(\vec{x}_0 + \vec{h}) - f(\vec{x}_0) = \frac{1}{2} Q(\vec{h}) + \mathcal{O}(\|\vec{h}\|^2) \quad \text{für} \quad \|\vec{h}\| \to 0. \tag{8.3}$$

Das heißt, der Ausdruck $Q(\vec{h})$ ist ein *Indikator* für die Existenz eines relativen Extremums in $\vec{x}_0$.

Definition 13.24 *Für eine gegebene symmetrische Matrix $A \in \mathbf{R}^{(n,n)}$ heiße die Funktion $Q \in \operatorname{Abb}(\mathbf{R}^n, \mathbf{R})$ mit*

$$\boxed{Q(\vec{h}) := \langle A\vec{h}, \vec{h} \rangle, \quad \vec{h} \in \mathbf{R}^n,}$$

die zu A gehörige **quadratische Form** *. Beide Größen A und Q heißen*

- **positiv** (*negativ*) **semidefinit** $:\Leftrightarrow \quad Q(\vec{h}) \geq 0 \quad (\text{bzw. } \leq 0) \quad \forall\, \vec{h} \in \mathbf{R}^n$,
- **positiv** (*negativ*) **definit** $:\Leftrightarrow \quad Q(\vec{h}) > 0 \quad (\text{bzw. } < 0) \quad \forall\, \vec{0} \neq \vec{h} \in \mathbf{R}^n$,
- **indefinit** $:\Leftrightarrow$ *Q ist nicht semidefinit.*

Bemerkung 13.12 Die hier gegebene Definition der positiven Definitheit einer Matrix $A \in \mathbf{R}^{(n,n)}$ ist konsistent mit der bereits früher formulierten Definition 5.26; man vergleiche auch Satz 11.13. □

BSP. (13.8.4) Es sei $f(x,y) := xy$ die Funktion aus BSP. (13.8.2). Hier gilt

$$H(\vec{x}_0) = \begin{bmatrix} f_{xx}(x_0,y_0) & f_{xy}(x_0,y_0) \\ f_{xy}(x_0,y_0) & f_{yy}(x_0,y_0) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

und somit

$$Q(\vec{h}) = \langle H(\vec{x}_0)\vec{h}, \vec{h}\rangle = \left\langle \begin{bmatrix} h_2 \\ h_1 \end{bmatrix}, \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \right\rangle = 2h_1h_2 \gtreqless 0.$$

Der im Punkt $\vec{x}_0 := \vec{0}$ vorhandene Sattelpunkt zeichnet sich offenbar durch eine indefinite HESSE–Matrix $H(\vec{x}_0)$ aus.

Auf Grund der Beziehung (8.3) und der obigen Definition 13.24 erhält man nun die folgende Charakterisierung eines kritischen Punktes durch die HESSE–Matrix.

Satz 13.23 *Gegeben seien ein Gebiet $\Omega \subset \mathbf{R}^n$ und darauf eine skalare Funktion $f \in C^2(\Omega)$. Es sei ferner $\vec{x}_0 \in \Omega$ ein kritischer Punkt von f und $H(\vec{x}_0)$ die zugeordnete* HESSE*–Matrix von f in $\vec{x}_0$. Dann gelten die folgenden Implikationen:*

$H(\vec{x}_0)$ *ist* **positiv definit**	$\Rightarrow$	*f hat in $\vec{x}_0$ ein relatives* **Minimum**,
$H(\vec{x}_0)$ *ist* **negativ definit**	$\Rightarrow$	*f hat in $\vec{x}_0$ ein relatives* **Maximum**,
$H(\vec{x}_0)$ *ist* **indefinit**	$\Rightarrow$	*f hat in $\vec{x}_0$ einen* **Sattelpunkt**.

Ist $H(\vec{x}_0)$ **semidefinit** *oder* **Null**, *nicht aber definit, so ist eine Charakterisierung nur mit Hilfe höherer Ableitungen möglich.*

Eine Aussage darüber, wann eine symmetrische Matrix $H = H^T \in \mathbf{R}^{(n,n)}$ positiv definit ist, wurde bereits in Satz 11.13 mit Hilfe der Eigenwerte λ_j der Matrix H getroffen: Eine symmetrische Matrix ist stets **normal**, und sie besitzt daher ein vollständiges System von Eigenvektoren $\vec{v}_j$, $j = 1, 2, \ldots, n$, die eine ON–Basis des $\mathbf{R}^n$ aufspannen. Die Matrix H gestattet die *Spektralzerlegung*

$$H = \sum_{j=1}^{n} \lambda_j (\vec{v}_j \otimes \vec{v}_j),$$

und aus dieser Spektralzerlegung resultiert die folgende Darstellung der zu H gehörigen quadratischen Form

$$Q(\vec{h}) = \langle H\vec{h}, \vec{h}\rangle = \sum_{j=1}^{n} \lambda_j \, |\langle \vec{v}_j, \vec{h}\rangle|^2, \qquad \vec{h} \in \mathbf{R}^n.$$

Man liest an dieser Darstellung die folgenden Definitheitseigenschaften direkt ab:

Satz 13.24 *Es sei eine symmetrische Matrix $H = H^T \in \mathbf{R}^{(n,n)}$ vorgelegt. Dann gelten die folgenden Äquivalenzen:*

H *ist positiv (negativ) definit* $\Leftrightarrow$ *alle Ew* λ_j *von* H *sind positiv (negativ),*
H *ist positiv (negativ) semidefinit* $\Leftrightarrow$ *alle Ew* λ_j *von* H *sind nichtnegativ (nichtpositiv),*
H *ist indefinit* $\Leftrightarrow$ *es treten Ew* $\lambda_j > 0$ **und** *Ew* $\lambda_k < 0$ *auf.*

BSP. (13.8.5) Es sei $f(x,y) := -x^4 + x^2y^2 + y^3 + y^2$ die Funktion aus BSP. (13.8.3). Wir untersuchen die HESSE–Matrix $H(\vec{x}_0)$ in dem kritischen Punkt $\vec{x}_0 := (\sqrt{2}, -2)^T$. Zunächst bestimmen wir $H(\vec{x})$ in einem beliebigen Punkt:

$$H(\vec{x}) = \begin{bmatrix} f_{xx}(x,y) & f_{xy}(x,y) \\ f_{xy}(x,y) & f_{yy}(x,y) \end{bmatrix} = \begin{bmatrix} -12x^2 + 2y^2 & 4xy \\ 4xy & 2x^2 + 6y + 2 \end{bmatrix}.$$

Hieraus resultieren

$$H(\vec{x}_0) =: H = \begin{bmatrix} -16 & -8\sqrt{2} \\ -8\sqrt{2} & -6 \end{bmatrix}, \quad \det(H - \lambda\, Id) = \begin{vmatrix} -16-\lambda & -8\sqrt{2} \\ -8\sqrt{2} & -6-\lambda \end{vmatrix} = \lambda^2 + 22\lambda - 32,$$

mit den zwei Eigenwerten $\lambda_\pm = -11 \pm \sqrt{153}$. Da nun $\lambda_+ > 0$ und $\lambda_- < 0$ gelten, liegt der indefinite Fall vor: $\vec{x}_0$ ist ein **Sattelpunkt**.

Bemerkung 13.13 Es wird wegen der Kompliziertheit davon abgeraten, im unentschiedenen Fall der semidefiniten HESSE–Matrix $H(\vec{x}_0)$ die Diskussion mit höheren Ableitungen durchzuführen. In der Regel kann man durch eine Analyse der Umgebung des kritischen Punktes $\vec{x}_0$ auf direktem Wege eine Aussage über den Charakter von $\vec{x}_0$ treffen. □

Sonderfall: Funktionen in 2 Veränderlichen

Im praktisch wichtigen Fall einer Funktion $f = f(x,y)$ in **zwei** Veränderlichen kann das Definitheitsproblem für die HESSE–Matrix $H = H(\vec{x}_0)$ ohne Berechnung der Eigenwerte durch Inspektion der Koeffizienten von H direkt gelöst werden. Im Sinne einer Analyse setzen wir

$$a := f_{xx}(\vec{x}_0),\ b := f_{xy}(\vec{x}_0),\ c := f_{yy}(\vec{x}_0), \quad H = \begin{bmatrix} a & b \\ b & c \end{bmatrix}.$$

Nun gilt

$$\det(H - \lambda\, Id) = (a-\lambda)(c-\lambda) - b^2 = \lambda^2 - (a+c)\lambda + ac - b^2,$$

und daraus resultieren $\lambda_1 \cdot \lambda_2 = ac - b^2 = \det H$ sowie $\lambda_1 + \lambda_2 = a + c = \mathrm{Sp}\,(H)$. Diese Relationen ermöglichen nun die folgende Klassifizierung der HESSE–Matrix $H \in \mathbf{R}^{(2,2)}$:

$$\det H < 0 \quad \Leftrightarrow \quad \lambda_1 \cdot \lambda_2 < 0 \quad \Leftrightarrow \quad H \text{ ist } \textbf{indefinit},$$

$$\det H = 0 \quad \Leftrightarrow \quad \lambda_1 \cdot \lambda_2 = 0 \quad \Leftrightarrow \quad H \text{ ist } \textbf{semidefinit},$$

$$\det H > 0 \quad \Leftrightarrow \quad \lambda_1 \cdot \lambda_2 > 0 \quad \Leftrightarrow \quad H \text{ ist } \textbf{definit}, \text{ und zwar } \begin{cases} \textbf{positiv} & : a > 0, \\ \textbf{negativ} & : a < 0. \end{cases}$$

Die Umsetzung dieser Klassifizierung in Aussagen über den Charakter von kritischen Punkten führt unmittelbar auf den folgenden

Satz 13.25 *Gegeben seien ein Gebiet* $\Omega \subset \mathbf{R}^2$ *und eine Funktion* $f \in C^2(\Omega)$. *In einem Punkt* $\vec{x}_0 \in \Omega$ *gelte* $f_x(\vec{x}_0) = 0 = f_y(\vec{x}_0)$. *Dann entscheidet das Vorzeichen der Diskriminante*

$$\boxed{\Delta(\vec{x}_0) := f_{xx}(\vec{x}_0) \cdot f_{yy}(\vec{x}_0) - f_{xy}^2(\vec{x}_0)}$$

in der folgenden Weise über die Art des kritischen Punktes $\vec{x}_0$:

$\Delta(\vec{x}_0) > 0$		$\Delta(\vec{x}_0) < 0$	$\Delta(\vec{x}_0) = 0$
$f_{xx}(\vec{x}_0) > 0$	$f_{xx}(\vec{x}_0) < 0$	**Sattelpunkt**	**keine Aussage**
Minimum	**Maximum**		

BSP. (13.8.6) Es sei $f(x,y) := -x^4 + x^2y^2 + y^3 + y^2$ die Funktion aus BSP. (13.8.3). Wir berechnen

$$\Delta(x,y) := f_{xx}(x,y) \cdot f_{yy}(x,y) - f_{xy}^2(x,y) = (-12x^2 + 2y^2)(2x^2 + 6y + 2) - 16x^2y^2.$$

Daraus resultieren:

- $\Delta(\vec{x}_1) := \Delta(0,0) = 0$: Es liegt der unentscheidbare Fall vor.
- $\Delta(\vec{x}_2) := \Delta(0, -\frac{2}{3}) = -\frac{16}{9} < 0$: Der Punkt $\vec{x}_2$ ist ein **Sattelpunkt** von f.
- $\Delta(\vec{x}_{3,4}) := \Delta(\pm\frac{1}{\sqrt{2}}, -1) = 4 > 0$ und $f_{xx}(\vec{x}_{3,4}) = -4 < 0$: Die Punkte $\vec{x}_{3,4}$ sind relative **Maxima** von f.
- $\Delta(\vec{x}_{5,6}) := \Delta(\pm\sqrt{2}, -2) = -32 < 0$: Die Punkte $\vec{x}_{5,6}$ sind **Sattelpunkte** von f.

Im Punkt $\vec{x}_1 := \vec{0}$ kann man sich noch überlegen, dass wegen

$$f(x,0) = -x^4 < 0 \ \forall\, x \neq 0, \qquad f(0,y) = y^2(1+y) > 0 \ \forall\, 0 < |y| < 1$$

ein **Sattelpunkt** liegen muss.

13.9 Extremwertaufgaben mit Nebenbedingungen

In vielen anwendungsorientierten Aufgabenstellungen sind bei der Bestimmung der Extremwerte einer Funktion $f \in \mathrm{Abb}(\mathbf{R}^n, \mathbf{R})$ gewisse **Nebenbedingungen** zu beachten, durch die der zur Extremwertbildung zugelassene Bereich $D(f)$ auf eine (meistens abgeschlossene) Teilmenge $G \subset D(f)$ eingeschränkt wird.

BSP. (13.9.1) Es ist die kürzeste (euklidische) Entfernung des Punktes $(a,b,c)^T \in \mathbf{R}^3$ von der Ebene

$$G := \{\vec{x} \in \mathbf{R}^3 \,:\, g(x,y,z) := Ax + By + Cz - D = 0\}$$

zu bestimmen. Die Ebene G ist hier eine Äquipotentialfläche der Funktion g. Wir setzen zunächst $\vec{x}_0 := (a,b,c)^T$ und bestimmen den (euklidischen) Abstand zu einem beliebigen Punkt $\vec{x} \in \mathbf{R}^3$

$$d^2(\vec{x}, \vec{x}_0) = \|\vec{x} - \vec{x}_0\|^2 = (x-a)^2 + (y-b)^2 + (z-c)^2 =: f(x,y,z) = f(\vec{x}).$$

Nun kann die gestellte Aufgabe folgendermaßen formuliert werden:

> Zu bestimmen ist das Minimum der Funktion $f(\vec{x})$ für $\vec{x} \in \mathbf{R}^3$ unter der Nebenbedingung $g(\vec{x}) = 0$.
>
> Oder äquivalent: Zu bestimmen ist das Minimum von $f(\vec{x})$ für $\vec{x} \in G$.

Wir werden die Lösung dieser Extremwertaufgabe mit Nebenbedingung im nächsten Beispiel bestimmen.

Eine allgemeinere Formulierung von **Extremwertaufgaben mit Nebenbedingungen** lautet:

Gegeben seien ein Gebiet $\Omega \subset \mathbf{R}^n$ und darauf Funktionen $f, g : \Omega \to \mathbf{R}$. Zu bestimmen sind die relativen Extrema $\vec{x}_0 \in \Omega$ der Funktion f unter der Nebenbedingung $g(\vec{x}_0) = 0$.

Oder äquivalent: Zu bestimmen sind die relativen Extrema $\vec{x}_0$ der Funktion $f : G \to \mathbf{R}$ auf der Äquipotentialfläche

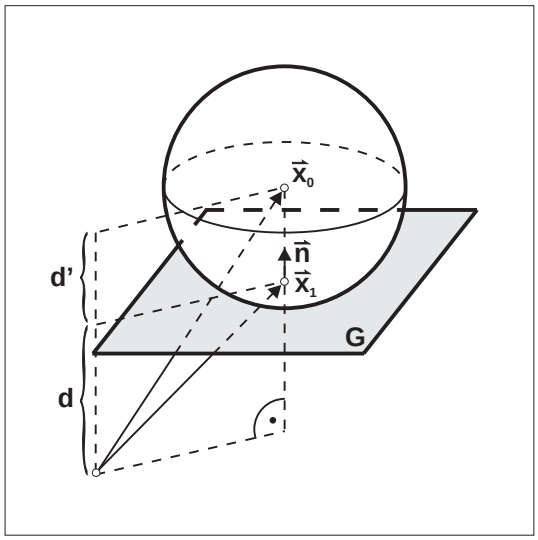
$$G := \{\vec{x} \in \Omega \, : \, g(\vec{x}) = 0\}.$$

Bemerkung 13.14 Die in Abschnitt 13.8 diskutierten Verfahren zur Extremwertbestimmung können hier nicht kritiklos auf die soeben formulierte Extremwertaufgabe mit Nebenbedingungen übertragen werden, da die Äquipotentialfläche $G \subset \mathbf{R}^n$ im allgemeinen *nicht offen* ist. Nicht alle Punkte $\vec{x}_0 \in G$ sind innere Punkte, so dass die Regeln der Differentiation meistens nicht anwendbar sind. □

BSP. (13.9.2) Die Funktionen f und g seien wie in BSP. (13.9.1) vorgegeben. Die Äquipotentialfläche G der Funktion g mit der Gleichung $g(\vec{x}) = 0$ ist eine Ebene, deren HESSEsche Normalform $\langle \vec{x}, \vec{n} \rangle = d(\vec{0}, G)$ durch die Vorgaben

$$\vec{n} = \frac{\pm 1}{\sqrt{A^2 + B^2 + C^2}} \begin{bmatrix} A \\ B \\ C \end{bmatrix}, \qquad d(\vec{0}, G) = \frac{\pm D}{\sqrt{A^2 + B^2 + C^2}} \geq 0$$

festgelegt ist. Die Äquipotentialflächen der Funktion f hingegen sind konzentrische Sphären um den Mittelpunkt $\vec{x}_0 = (a, b, c)^T$.



Die Äquipotentialflächen
$f(\vec{x}) = const > 0$

Aus der geometrischen Betrachtung wird ersichtlich, dass die Funktion f auf der Menge G im gemeinsamen Berührungspunkt $\vec{x}_1$ beider Äquipotentialflächen minimal wird. Unter der Annahme $\operatorname{grad} g(\vec{x}_1) \neq \vec{0}$ muss also $\operatorname{grad} f(\vec{x}_1) \parallel \operatorname{grad} g(\vec{x}_1)$ gelten, oder äquivalent

$$\operatorname{grad} f(\vec{x}_1) = \lambda \operatorname{grad} g(\vec{x}_1).$$

Mit den analytischen Ausdrücken $g(\vec{x}) = Ax + By + Cz - D$ und $f(\vec{x}) = (x-a)^2 + (y-b)^2 + (z-c)^2$ resultiert daraus der folgende Satz von Bestimmungsgleichungen für den Punkt $\vec{x}_1$:

$$\left.\begin{array}{rcl|l} f_x(\vec{x}_1) = 2(x-a) & \stackrel{!}{=} & \lambda g_x(\vec{x}_1) = \lambda A & \cdot A \\ f_y(\vec{x}_1) = 2(y-b) & \stackrel{!}{=} & \lambda g_y(\vec{x}_1) = \lambda B & \cdot B \\ f_z(\vec{x}_1) = 2(z-c) & \stackrel{!}{=} & \lambda g_z(\vec{x}_1) = \lambda C & \cdot C \end{array}\right\} (+) \tag{9.1}$$

Bildet man die obige Summe, so ergibt sich

$$(x-a)A + (y-b)B + (z-c)C \stackrel{g(\vec{x})=0}{=} -(Aa + Bb + Cc - D) \stackrel{!}{=} \frac{\lambda}{2}(A^2 + B^2 + C^2),$$

und daraus erhält man den Multiplikator λ gemäß

$$\lambda = -2\,\frac{Aa + Bb + Cc - D}{A^2 + B^2 + C^2}.$$

Wird dieser in (9.1) eingesetzt, so resultiert

$$\vec{x}_1 = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} - \frac{Aa + Bb + Cc - D}{\sqrt{A^2 + B^2 + C^2}} \cdot \frac{1}{\sqrt{A^2 + B^2 + C^2}} \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \vec{x}_0 - (\underbrace{\langle \vec{x}_0, \vec{n}\rangle - d(\vec{0}, G)}_{=d'(\vec{x}_0, G)})\vec{n},$$

in Übereinstimmung mit der aus der Anschauung gewonnenen Lösung.

Wir zeigen nachfolgend, dass das in BSP. (13.9.2) verwendete Verfahren auch für allgemeine Extremwertaufgaben mit Nebenbedingungen Gültigkeit besitzt.

Satz 13.26 *Gegeben seien ein Gebiet $\Omega \subset \mathbf{R}^n$ und Funktionen $f, g \in C^1(\Omega)$. Die Funktion f besitze in einem Punkt $\vec{x}_0 \in \Omega$ ein relatives Extremum unter der Nebenbedingung $g(\vec{x}_0) = 0$. Gilt* $\operatorname{grad} g(\vec{x}_0) \neq \vec{0}$*, so gibt es eine Zahl $\lambda \in \mathbf{R}$ mit*

$$\boxed{\operatorname{grad} f(\vec{x}_0) = \lambda \operatorname{grad} g(\vec{x}_0).}$$

Die Zahl λ heiße Lagrange–**Multiplikator**. *Für* $\operatorname{grad} g(\vec{x}_0) = \vec{0}$ *hat man keine Aussage.*

Begründung: Auf der Äquipotentialfläche $G := \{\vec{x} \in \Omega : g(\vec{x}) = 0\}$ wähle man einen beliebigen Weg $\vec{x} = \vec{x}(t) \in G$ so, dass $\vec{x}(0) = \vec{x}_0$ und $\vec{x} \in C^1$ gelten. Hat f in dem vorgegebenen Punkt $\vec{x}_0$ ein relatives Extremum, so muss auch die Funktion $h(t) := f(\vec{x}(t))$ bei $t = 0$ extremal sein. Wegen $h \in C^1$ erhalten wir unter Verwendung der 2.Kettenregel

$$0 = \frac{d}{dt}\, h(t)|_{t=0} = \langle \operatorname{grad} f(\vec{x}_0), \dot{\vec{x}}(0)\rangle.$$

Also ist der Vektor $\operatorname{grad} f(\vec{x}_0)$ senkrecht zu allen Wegen in G durch den Punkt $\vec{x}_0$. Das heißt, es gilt $\operatorname{grad} f(\vec{x}_0) \parallel \operatorname{grad} g(\vec{x}_0)$, oder äquivalent $\operatorname{grad} f(\vec{x}_0) = \lambda \operatorname{grad} g(\vec{x}_0)$. □

Bemerkung 13.15 (a) Der Satz 13.26 vermittelt lediglich ein **notwendiges Kriterium** für das Auffinden eines relativen Extremums $\vec{x}_0$ unter der Nebenbedingung $g(\vec{x}_0) = 0$. Es genügt, nur solche Punkte $\vec{x}_0$ zu untersuchen, die Lösungen des folgenden Gleichungssystems sind:

$$\boxed{\begin{array}{lcl} g(\vec{x}_0) & = & 0, \\ \operatorname{grad} f(\vec{x}_0) & = & \lambda \operatorname{grad} g(\vec{x}_0). \end{array}} \tag{9.2}$$

Das sind $n+1$ (nichtlineare) Gleichungen für die $n+1$ Unbekannten $x_{01}, x_{02}, \ldots, x_{0n}, \lambda$.

(b) Führt man die Funktion

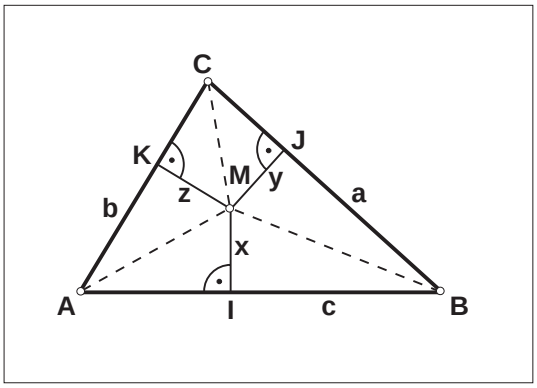
$$F(\vec{x}, \lambda) := f(\vec{x}) - \lambda g(\vec{x})$$

ein, so ist die folgende Gleichung eine äquivalente Formulierung von (9.2):

$$\boxed{\operatorname{grad} F(\vec{x}_0, \lambda) = \vec{0}.} \tag{9.3}$$

Hierin ist die Differentiation auch auf die Variable λ zu erstrecken.

(c) *Hinreichende* Bedingungen wie im Fall der Extremwertaufgaben *ohne* Nebenbedingungen sollen hier nicht formuliert werden. □



Das Dreieck aus BSP. (13.9.3)

BSP. (13.9.3) Gegeben sei ein Dreieck ABC mit den Seiten $c = \overline{AB}$, $a = \overline{BC}$, $b = \overline{CA}$ und der Fläche $F > 0$. Für einen Punkt M im Inneren des Dreiecks seien I, J, K die Fußpunkte der Lote von M auf die Seiten $\overline{AB}, \overline{BC}$ bzw. $\overline{CA}$. Man berechne das Maximum der Funktion $d(x, y, z) := xyz$ mit $x := \overline{MI}$, $y := \overline{MJ}$, $z := \overline{MK}$, siehe obige Skizze.

Lösung: Offensichtlich ist x die Höhe des Dreiecks ABM, und c ist seine Grundlinie, so dass sein Flächeninhalt $\frac{1}{2}cx$ beträgt. Mit analoger Überlegung resultiert als Gesamtfläche $2F = cx + ay + bz = const.$ Wir setzen nun $g(\vec{x}) := cx + ay + bz - 2F$, so dass die Extremwertaufgabe wie folgt formuliert werden kann:

$$d(\vec{x}) := xyz \stackrel{!}{=} \text{Max} \quad \text{unter der Nebenbedingung } g(\vec{x}) = 0.$$

Eine Lösung $\vec{x}_0 \in \mathbf{R}^3$ muss notwendig ein kritischer Punkt der Funktion $F(\vec{x}, \lambda) := d(\vec{x}) - \lambda\, g(\vec{x})$ sein, das heißt, es muss gelten:

$$\operatorname{grad} F(\vec{x}, \lambda) = \begin{bmatrix} yz - \lambda c \\ xz - \lambda a \\ xy - \lambda b \\ cx + ay + bz - 2F \end{bmatrix} = \vec{0} \quad \Rightarrow \quad \left.\begin{array}{rcl} xyz &=& \lambda cx, \\ xyz &=& \lambda ay, \\ xyz &=& \lambda bz, \end{array}\right\} \quad \Rightarrow \quad cx = ay = bz.$$

Mit diesem Resultat folgt aus der Nebenbedingung $g(\vec{x}) = 0$ die Lösung $3cx = 3ay = 3bz = 2F$, also

$$x = \frac{2F}{3c}, \quad y = \frac{2F}{3a}, \quad z = \frac{2F}{3b}, \quad \vec{x}_0 = \frac{2F}{3} \begin{bmatrix} \frac{1}{c} \\ \frac{1}{a} \\ \frac{1}{b} \end{bmatrix}.$$

Offenbar ist $\vec{x}_0$ das gesuchte Maximum der Funktion d, denn wegen $x \geq 0, y \geq 0, z \geq 0$ können Minima von d nur für $x = 0$ oder $y = 0$ oder $z = 0$ vorliegen.

Die geometrische Interpretation der Lösung ist die folgende: Bezeichnen h_a, h_b, h_c die Höhen des Dreiecks, so ist $2F = ah_a = bh_b = ch_c$. Demgemäß gilt

$$x = \frac{h_c}{3}, \quad y = \frac{h_a}{3}, \quad z = \frac{h_b}{3},$$

das heißt, der gesuchte Punkt M ist der **Schnittpunkt der Höhenlinien** des Dreiecks.

Häufig gelingt es, die Nebenbedingung $g(\vec{x}) = 0$ nach einer der Variablen $\vec{x} = (x_1, x_2, \ldots, x_n)^T$ *explizit* aufzulösen; zum Beispiel gelte $x_n = h(\vec{x}')$, $\vec{x}' := (x_1, x_2, \ldots, x_{n-1})^T$. Dann kann die Extremwertaufgabe

$$f(\vec{x}) \stackrel{!}{=} \text{ Extr.} \quad \text{unter der Nebenbedingung } g(\vec{x}) = 0$$

auch als Extremwertaufgabe *ohne* Nebenbedingung für die Funktion

$$F(\vec{x}') := f(x_1, x_2, \ldots, x_{n-1}, x_n = h(\vec{x}'))$$

formuliert werden.

BSP. (13.9.4) Es seien $d(x, y, z) := xyz$ und $g(x, y, z) := cx + ay + bz - 2F$ die Funktionen aus BSP. (13.9.3). Wir lösen die Gleichung $g = 0$ nach der Variablen z auf und setzen das Resultat in die Funktion d ein. Es gilt $z = \frac{1}{b}(2F - cx - ay)$, und wir haben das Maximum der Funktion $F(x, y) := \frac{xy}{b}(2F - cx - ay)$ zu bestimmen. Deren kritische Punkte sind die Lösungen der beiden Gleichungen

$$F_x(x, y) = \frac{y}{b}(2F - 2cx - ay) = 0, \quad F_y(x, y) = \frac{x}{b}(2F - cx - 2ay) = 0.$$

Die Fälle $x = 0$ und/oder $y = 0$ liefern $d = 0$, also nicht das gesuchte Maximum. Somit müssen x, y Lösungen des linearen Gleichungssystems

$$2cx + ay = 2F, \quad cx + 2ay = 2F$$

sein. Dessen eindeutig bestimmte Lösung führt wiederum auf das bereits bekannte Resultat

$$x = \frac{2F}{3c}, \quad y = \frac{2F}{3a}, \quad z = \frac{2F}{3b}.$$

Ist die Äquipotentialfläche $G := \{\vec{x} \in D(f) : g(\vec{x}) = 0\}$ **kompakt**, also beschränkt und abgeschlossen, so müssen bei stetigem $f \in \mathrm{Abb}(\mathbf{R}^n, \mathbf{R})$ Maximum und Minimum notwendig auf G angenommen werden. Also ist in diesem Fall die Existenz von Extremwerten (und somit die Lösbarkeit der Extremwertaufgabe) gewährleistet.

BSP. (13.9.5) Es sind die relativen Extrema der Funktion $f(x, y) := x^2 - y^2$ unter der Nebenbedingung $g(x, y) := x^2 + y^2 - 1 = 0$ zu bestimmen. Die Äquipotentialfläche $g = 0$ ist hier die Kreislinie $S_1(\vec{0})$ vom Radius 1 um den Mittelpunkt $\vec{0}$, also eine kompakte Punktmenge. Da die Gleichung $g = 0$ nach y^2 auflösbar ist, kann die Variable y aus der Extremwertaufgabe eliminiert werden:

$$F(x) := f(x, y^2 = 1 - x^2) = 2x^2 - 1.$$

Die Extremwertaufgabe für F ist nun mit den Mitteln der gewöhnlichen Differentialrechnung lösbar: Die Bedingung $F'(x) = 4x \stackrel{!}{=} 0$ führt über $x = 0$ und $y^2 = 1$ auf die beiden Minima $f(0, \pm 1) = -1$. Aus Stetigkeitsgründen muss f auf der kompakten Menge $S_1(\vec{0})$ aber auch Maxima haben, die noch nicht durch die Bedingung $F'(x) = 0$ erfasst wurden. Wegen $x \in [-1, 1]$ können diese nur an den Intervallenden $x = \pm 1$ liegen. Dort gilt $y^2 = 0$ und somit $f(\pm 1, 0) = 1$.

Extremwertaufgaben mit mehreren Nebenbedingungen: Auf einem Gebiet $\Omega \subset \mathbf{R}^n$ seien reelle Funktionen $f, g_j \in C^1(\Omega)$, $1 \le j \le m < n$, so vorgegeben, dass die Äquipotentialflächen $g_j = 0$ eine nichtleere Schnittmenge besitzen:

$$G^o := \bigcap_{j=1}^{m} \{\vec{x} \in \Omega \,:\, g_j(\vec{x}) = 0\} \neq \emptyset.$$

Dann kann man das folgende Resultat zeigen:

Satz 13.27 *Ist $\vec{x}_0 \in \Omega$ unter der Nebenbedingung $\vec{x}_0 \in G^o$ ein Extremwert der Funktion f, und ist das Vektor–System* $\operatorname{grad} g_j(\vec{x}_0)$, $j = 1, 2, \ldots, m$, *linear unabhängig, so gibt es eindeutig bestimmte Zahlen $\lambda_1, \lambda_2, \ldots, \lambda_m$ derart, dass gilt:*

$$\boxed{\begin{array}{lcl} \operatorname{grad} f(\vec{x}_0) & = & \displaystyle\sum_{j=1}^{m} \lambda_j \operatorname{grad} g_j(\vec{x}_0), \\ g_j(\vec{x}_0) & = & 0, \quad j = 1, 2, \ldots, m. \end{array}} \tag{9.4}$$

Das System (9.4) liefert $n + m$ (nichtlineare) Bestimmungsgleichungen für die $n + m$ Unbekannten $\vec{x}_0 = (x_{01}, x_{02}, \ldots, x_{0n})^T$, $\lambda_1, \lambda_2, \ldots, \lambda_m$. Die geforderte lineare Unabhängigkeit ist notwendig für die Eindeutigkeit der Lösungen λ_j. Bei linearer Abhängigkeit der Gradienten wird die Situation weitaus komplizierter; wir verweisen auf die Literatur, z.B. H. Dallmann/K.–H. Elster, Einführung in die Höhere Mathematik, Band 2. Vieweg, Braunschweig (1981).

BSP. (13.9.6) Es seien die Funktionen $f(x, y, z) := 2x + 3y + 2z$, $g_1(x, y, z) := x^2 + y^2 - 2$ und $g_2(x, y, z) := x + z - 1$ auf $\Omega := \mathbf{R}^3$ gegeben. Zu bestimmen sind die Extemwerte von f unter den Nebenbedingungen $g_1 = 0 = g_2$. Offenbar sind die Äquipotentialflächen $g_1 = 0$ und $g_2 = 0$ ein Kreiszylinder bzw. eine Ebene. Die Schnittmenge $G^o = \{\vec{x} \in \mathbf{R}^3 \,:\, g_1(\vec{x}) = 0 = g_2(\vec{x})\}$ ist kompakt. Relative Extrema $\vec{x}_0$ von f unter der Nebenbedingung $\vec{x}_0 \in G^o$ müssen also die Gleichungen (9.4) erfüllen:

$$\begin{array}{rcl} \vec{0} & = & \operatorname{grad} f(\vec{x}_0) - \lambda_1 \operatorname{grad} g_1(\vec{x}_0) - \lambda_2 \operatorname{grad} g_2(\vec{x}_0) = \begin{bmatrix} 2 - \lambda_1 \cdot 2x_0 - \lambda_2 \cdot 1 \\ 3 - \lambda_1 \cdot 2y_0 - \lambda_2 \cdot 0 \\ 2 - \lambda_1 \cdot 0 - \lambda_2 \cdot 1 \end{bmatrix}, \\ 0 & = & g_1(\vec{x}_0) = x_0^2 + y_0^2 - 2, \\ 0 & = & g_2(\vec{x}_0) = x_0 + z_0 - 1. \end{array}$$

Aus diesen Gleichungen erhält man die Lösungen $x_0 = 0$, $y_0 = \pm\sqrt{2}$, $z_0 = 1$ sowie die Lagrange–Multiplikatoren $\lambda_1 = \pm\frac{3}{4}\sqrt{2}$, $\lambda_2 = 2$. Das heißt, mögliche Extremwerte von f unter den beiden Nebenbedingungen $g_1 = 0 = g_2$ liegen in den Punkten $\vec{x}_0 := (0, \pm\sqrt{2}, 1)^T$. Es gelten $f(0, \sqrt{2}, 1) = 2 + 3\sqrt{2}$ sowie $f(0, -\sqrt{2}, 1) = 2 - 3\sqrt{2}$, und dies sind in der Tat Maximum und Minimum von f auf der kompakten Menge G^o.

Kapitel 14

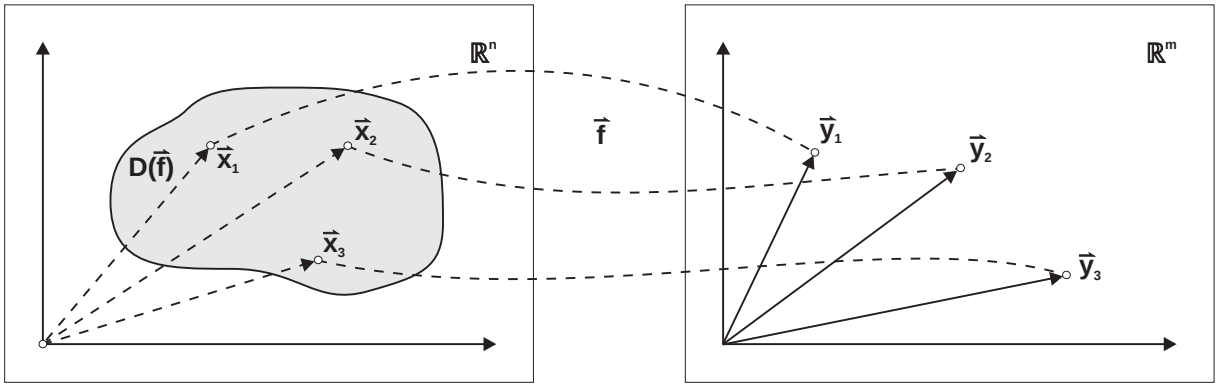
Differentialrechnung vektorwertiger Funktionen

14.1 Definitionen und Beispiele

Definition 14.1 *Abbildungen* $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$, $m > 1$, *heißen* (m*–dimensionale*) **Vektorfunktionen** *von* n (*unabhängigen*) *Veränderlichen, oder kurz* **vektorwertige Funktionen**.

Eine vektorwertige Funktion $\vec{f}$ ordnet mithin jedem Punkt $\vec{x} \in D(\vec{f})$ einen Vektor $\vec{y} \in \mathbf{R}^m$ zu:

$$\vec{y} = \vec{f}(\vec{x}), \qquad \vec{x} \in D(\vec{f}) \subset \mathbf{R}^n.$$



Die Urbildmenge einer vektorwertigen Funktion **Die Bildmenge einer vektorwertigen Funktion**

Da jede Koordinate des Bildvektors $\vec{y} = (y_1, y_2, \ldots, y_m)^T$ eine (skalare) Funktion des Urbildvektors $\vec{x} \in D(\vec{f})$ ist, können vektorwertige Funktionen generell auch durch m skalare **Komponentenfunktionen** dargestellt werden:

$$\vec{f}(\vec{x}) = (f_1(\vec{x}), f_2(\vec{x}), \ldots, f_m(\vec{x}))^T, \qquad f_j \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}).$$

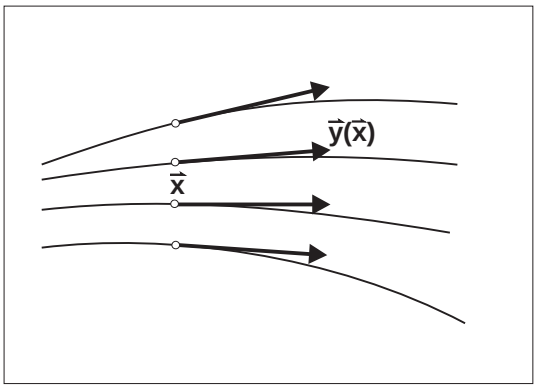
BSP. (14.1.1) Eine elektrische **Ladung** Q im Punkt $\vec{0} \in \mathbf{R}^3$ erzeugt gemäß den Maxwellschen Gesetzen ein elektrisches **Feld** der **Feldstärke**

$$\vec{E} := \frac{Q}{4\pi\epsilon}\left(\frac{x_1}{r^3}, \frac{x_2}{r^3}, \frac{x_3}{r^3}\right)^T, \quad \vec{x} = (x_1, x_2, x_3)^T \in \mathbf{R}^3 \setminus \{\vec{0}\}, \quad r := \|\vec{x}\|. \tag{1.1}$$

Hier ist eine vektorwertige Funktion $\vec{E} =: \vec{f} \in \mathrm{Abb}\,(\mathbf{R}^3, \mathbf{R}^3)$ mit $D(\vec{f}) := \mathbf{R}^3 \setminus \{\vec{0}\}$ und den Komponentenfunktionen

$$f_j(\vec{x}) := \frac{Q}{4\pi\epsilon}\,\frac{x_j}{r^3}, \quad j = 1, 2, 3,$$

vorgelegt. Die in der Physik auftretenden **Vektorfelder** sind überwiegend vektorwertige Funktionen der Art $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^3, \mathbf{R}^3)$. Dem **Ortsvektor** $\vec{x} \in \mathbf{R}^3$ wird jeweils genau ein **Feldvektor** $\vec{y} = \vec{f}(\vec{x})$ zugeordnet. Diejenigen Raumkurven, die die Feldvektoren als **Tangenten** besitzen, bilden die **Feldlinien**. Die Gesamtheit der Feldlinien ergibt das **Vektorfeld** der Funktion $\vec{f}$.



Feldlinien eines Vektorfeldes

BSP. (14.1.2) Das Vektorfeld (1.1) kann auch in der Form

$$\vec{E} = -\mathrm{grad}\,\varphi(\vec{x}), \qquad \varphi(\vec{x}) := \frac{Q}{4\pi\epsilon r},$$

geschrieben werden; wir verweisen auf BSP. (13.5.5). Allgemein heißen vektorwertige Funktionen $\vec{f}$ vom Typ

$$\vec{f}(\vec{x}) := -\mathrm{grad}\,\varphi(\vec{x}) \quad \text{mit gegebenem } \varphi \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$$

Gradientenfelder oder **Potentialfelder**. Die Funktion φ heißt dann ein **Potential** von $\vec{f}$. Im Fall $n = 3$ sind Potentialfelder spezielle Vektorfelder. Da der Vektor $\mathrm{grad}\,\varphi(\vec{x})$ senkrecht auf der Äquipotentialfläche der Funktion φ durch den Punkt $\vec{x}$ steht, stehen auch die Feldlinien eines Potentialfeldes stets senkrecht auf den Äquipotentialflächen. Natürlich wird hier die Regularität $\varphi \in C^1$ vorausgesetzt.

BSP. (14.1.3) Die **linearen** Vektorfunktionen $\vec{f} \in \mathrm{L}\,(\mathbf{R}^n, \mathbf{R}^m)$ sind gemäß Satz 5.3(b) genau die $m \times n$–Matrizen $A \in \mathbf{R}^{(m,n)}$. Es ist in diesem Fall üblich, anstelle des Funktionssymbols $\vec{f}$ das Matrixsymbol A zu verwenden. Die Spezifikation eines Definitionsbereiches erübrigt sich, da $A \in \mathbf{R}^{(m,n)}$ stets auf dem gesamten Vektorraum $\mathbf{R}^n$ operiert.

BSP. (14.1.4) Für $m > n$ heiße die durch

$$\vec{f}(\vec{x}) := (x_1, x_2, \ldots, x_n, 0, \ldots, 0)^T \in \mathbf{R}^m, \qquad \vec{x} \in \mathbf{R}^n,$$

definierte lineare Abbildung $\vec{f} \in \mathrm{L}\,(\mathbf{R}^n, \mathbf{R}^m)$ die **Projektionsabbildung** Pr von $\mathbf{R}^n$ in $\mathbf{R}^m$. Die ihr durch die Vorschrift $Pr(\vec{x}) = A\vec{x}$ zugeordnete Matrix $A \in \mathbf{R}^{(m,n)}$ ist die Matrix

$$A := \left.\begin{bmatrix} Id_n \\ \vec{0}^T \\ \vdots \\ \vec{0}^T \end{bmatrix}\right\} \; m - n \text{ mal.}$$

BSP. (14.1.5) **Affine** Funktionen $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ sind genau die Funktionen

$$\vec{f}(\vec{x}) := A\vec{x} + \vec{b}, \quad \vec{x} \in \mathbf{R}^n, \qquad A \in \mathbf{R}^{(m,n)}, \quad \vec{b} \in \mathbf{R}^m.$$

BSP. (14.1.6) **Koordinatentransformationen** $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^n)$ sind vektorwertige Funktionen. Wir hatten bereits in Abschnitt 13.1 **Kugelkoordinaten** in $\mathbf{R}^3$ ausführlich diskutiert:

$$\vec{x} = \vec{f}(r,\varphi,\vartheta) := \begin{bmatrix} r\cos\varphi\sin\vartheta \\ r\sin\varphi\sin\vartheta \\ r\cos\vartheta \end{bmatrix}, \quad D(\vec{f}) = \{(r,\varphi,\vartheta) \,:\, r>0,\ 0\le\varphi<2\pi,\ 0<\vartheta<\pi\} \subset \mathbf{R}^3.$$

14.2 Stetigkeit und Ableitung

Wie wir in Abschnitt 13.2 gezeigt haben, ist jeder der Vektorräume $\mathbf{R}^p$, $p \in \mathbf{N}$, unter der Metrik

(D3)
$$\boxed{d_p(\vec{x},\vec{y}) := \Big(\sum_{j=1}^{p} |x_j - y_j|^2\Big)^{1/2}, \quad \vec{x},\ \vec{y} \in \mathbf{R}^p,}$$

ein **vollständiger metrischer Raum**. Somit liegt gemäß Definition 13.7 den vektorwertigen Funktionen $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ der folgende Stetigkeitsbegriff zugrunde:

Definition 14.2 *Eine Funktion* $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ *heiße* **stetig** *im Punkte* $\vec{x}_0 \in D(\vec{f})$*, wenn für jede Folge* $(\vec{x}_k)_{k\in\mathbf{N}} \subset D(\vec{f})$ *mit* $\lim\limits_{k\to\infty} d_n(\vec{x}_k, \vec{x}_0) = 0$ *gilt:*

$$\boxed{\lim_{k\to\infty} d_m(\vec{f}(\vec{x}_k), \vec{f}(\vec{x}_0)) = 0.}$$

Genau wie im Sonderfall $n = 1$ ist die Stetigkeit von $\vec{f}$ im Punkt $\vec{x}_0 \in D(\vec{f})$ äquivalent mit der Stetigkeit jeder der Komponentenfunktionen von $\vec{f}$ in $\vec{x}_0$. Wir haben also das folgende wichtige Analogon zum Satz 6.12:

Satz 14.1 *Genau dann ist die vektorwertige Funktion*

$$\vec{f}(\vec{x}) := (f_1(\vec{x}), f_2(\vec{x}), \ldots, f_n(\vec{x}))^T, \quad f_k : D(\vec{f}) \to \mathbf{R}\ \ \forall\, k = 1,2,\ldots,n, \quad D(\vec{f}) \subset \mathbf{R}^n,$$

im Punkt $\vec{x}_0 \in D(\vec{f})$ *stetig, wenn jede ihrer Komponentenfunktionen* f_k*,* $k = 1,2,\ldots,n$*, in* $\vec{x}_0$ *stetig ist.*

Dieser Satz beinhaltet ein einfaches Kriterium zum Nachprüfen der Stetigkeit vektorwertiger Funktionen. Wir müssen gegenüber skalaren Funktionen $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ nichts Neues hinzulernen.

BSP. (14.2.1) Es sei $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R}^2)$ die vektorwertige Funktion

$$\vec{f}(x,y) := \Big(\frac{x^2}{e^{x+y}}, \frac{y}{x^2+1}\Big) =: (f_1(x,y), f_2(x,y))^T.$$

Auf dem Definitionsbereich $D(\vec{f}) := \mathbf{R}^2$ ist jede der beiden Komponentenfunktionen f_1, f_2 stetig, und somit ist auch $\vec{f}$ auf ganz $\mathbf{R}^2$ stetig.

BSP. (14.2.2) Jede **lineare** Abbildung $A \in \mathrm{L}\,(\mathbf{R}^n, \mathbf{R}^m) = \mathbf{R}^{(m,n)}$ ist stetig. Denn wegen

$$d_m(A\vec{x}, A\vec{x}_0) = \|A\vec{x} - A\vec{x}_0\| = \|A(\vec{x} - \vec{x}_0)\| =: \|A\vec{y}\|$$

braucht man Stetigkeit bei linearen Abbildungen nur im Punkt $\vec{x}_0 := \vec{0}$ zu untersuchen. Aus der Stetigkeit bei $\vec{0}$ folgt dann bereits die Stetigkeit für alle $\vec{x} \in \mathbf{R}^n$. Zum Nachweis der Stetigkeit im Punkt $\vec{0}$ bedienen wir uns der Tatsache, dass die FROBENIUS–Norm $\|A\|_F$ der Matrix $A = (a_{jk})$ mit der euklidischen Vektornorm $\|\vec{x}\|$ kompatibel ist:

$$d_m(A\vec{x}, \vec{0}) = \|A\vec{x}\| \leq \Big(\sum_{j=1}^{m} \sum_{k=1}^{n} |a_{jk}|^2 \Big)^{1/2} \|\vec{x}\| =: \|A\|_F \|\vec{x}\|;$$

man vergleiche Definition 6.22. Nun ist die Stetigkeit bei $\vec{0}$ offenkundig.

Für das Rechnen mit stetigen vektorwertigen Funktionen ist der folgende Satz noch sehr hilfreich:

Satz 14.2 (a) *Sind $\vec{f}, \vec{g} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ stetig in einem Punkt $\vec{x}_0 \in D(\vec{f}) \cap D(\vec{g})$, so sind auch die folgenden Funktionen in $\vec{x}_0$ stetig:*

$$\lambda \vec{f} + \mu \vec{g} \;\; \forall\, \lambda, \mu \in \mathbf{R}, \qquad \langle \vec{f}, \vec{g} \rangle, \qquad \|\vec{f}\|, \qquad \vec{f} \times \vec{g}, \text{ sofern } m = 3 \text{ gilt.}$$

(b) *Ist $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ im Punkte $\vec{x}_0 \in D(\vec{f})$ stetig sowie $\vec{g} \in \mathrm{Abb}\,(\mathbf{R}^m, \mathbf{R}^k)$ stetig im Punkt $\vec{f}(\vec{x}_0)$, so ist das Kompositum $\vec{g} \circ \vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^k)$ stetig in $\vec{x}_0$.*

BSP. (14.2.3) Auf dem Definitionsbereich $D(f) := \{(x, y) \in \mathbf{R}^2 : x^2 + y^2 < 1\}$ ist die skalare Funktion $f(x, y) := 1 - x^2 - y^2$ sicher stetig und sogar positiv. Ebenso ist die vektorwertige Funktion $\vec{g}(u) := (u, \sqrt{u}, \ln u)^T$ auf der positiven Halbachse $D(\vec{g}) := \{u \in \mathbf{R} : u > 0\}$ überall stetig. Wir folgern aus Satz 14.2, dass die zusammengesetzte Funktion

$$(\vec{g} \circ f)(x, y) = \vec{g}(f(x, y)) := \begin{bmatrix} 1 - x^2 - y^2 \\ \sqrt{1 - x^2 - y^2} \\ \ln(1 - x^2 - y^2) \end{bmatrix}$$

auf der gesamten Menge $D(\vec{f})$ stetig ist.

Der Vektorraum $\mathbf{R}^m$, $m \geq 2$, ist nicht geordnet, so dass es für Funktionen $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ keinen *Zwischenwertsatz* und keine *Maxima* bzw. *Minima* geben kann. Hingegen kann der Begriff der *gleichmäßigen Stetigkeit* in offenkundiger Weise auch für vektorwertige Funktionen erklärt werden. In Analogie zum Satz 13.10 haben wir:

Satz 14.3 *Eine stetige Funktion $\vec{f} : K \to \mathbf{R}^m$ ist auf der* **kompakten** *Teilmenge $K \subset \mathbf{R}^n$ sogar gleichmäßig stetig.*

Besitzen die Komponentenfunktionen $f_j \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ einer gegebenen vektorwertigen Funktion $\vec{f}$ in einem inneren Punkt $\vec{x}_0 \in D(\vec{f})$ partielle Ableitungen $\frac{\partial f_j}{\partial x_k}(\vec{x}_0)$, $j = 1, 2, \ldots, m$, so ist es sinnvoll, den folgenden Vektor zu betrachten:

$$\frac{\partial \vec{f}}{\partial x_k}(\vec{x}_0) := \Big(\frac{\partial f_1}{\partial x_k}(\vec{x}_0), \frac{\partial f_2}{\partial x_k}(\vec{x}_0), \ldots, \frac{\partial f_m}{\partial x_k}(\vec{x}_0) \Big)^T. \tag{2.1}$$

Definition 14.3 *Existieren die partiellen Ableitungen $\frac{\partial f_j}{\partial x_k}(\vec{x}_0)$, $1 \leq j \leq m$, in einem inneren Punkt $\vec{x}_0 \in D(\vec{f})$, so heiße der Vektor* (2.1) *die* **partielle Ableitung** *der vektorwertigen Funktion $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ im Punkte $\vec{x}_0$* **nach der k–ten Komponenten**. *Existieren die partiellen Ableitungen von $\vec{f}$ nach allen Komponenten $k = 1, 2, \ldots, n$, so heiße $\vec{f}$ (in $\vec{x}_0$)* **partiell differenzierbar**.

BSP. (14.2.4) Es sei $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R}^2)$ die Funktion aus BSP. (14.2.1). Dann existieren die partiellen Ableitungen

$$\frac{\partial \vec{f}}{\partial x}\,(x,y) = \begin{bmatrix} -(x^2-2x)e^{-(x+y)} \\ -2xy/(x^2+1)^2 \end{bmatrix}, \qquad \frac{\partial \vec{f}}{\partial y}\,(x,y) = \begin{bmatrix} -x^2\, e^{-(x+y)} \\ 1/(x^2+1) \end{bmatrix}$$

in jedem Punkt $\vec{x} = (x,y) \in \mathbf{R}^2$.

Ist die Funktion $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ in einem inneren Punkt $\vec{x}_0 \in D(\vec{f})$ partiell differenzierbar, so können die Spaltenvektoren

$$\frac{\partial \vec{f}}{\partial x_k}\,(\vec{x}_0) := \Big(\frac{\partial f_1}{\partial x_k}\,(\vec{x}_0), \frac{\partial f_2}{\partial x_k}\,(\vec{x}_0), \ldots, \frac{\partial f_m}{\partial x_k}\,(\vec{x}_0)\Big)^T, \quad k = 1,2,\ldots,n,$$

in einer Matrix $J_{\vec{f}}(\vec{x}_0) \in \mathbf{R}^{(m,n)}$ angeordnet werden. Diese Matrix spielt in der Analysis eine wichtige Rolle.

Definition 14.4 *Die Funktion $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ sei in einem inneren Punkt $\vec{x}_0 \in D(\vec{f})$ partiell differenzierbar. Dann heiße die $m \times n$–Matrix*

$$\boxed{J_{\vec{f}}(\vec{x}_0) := \frac{\partial(f_1,\ldots,f_m)}{\partial(x_1,\ldots,x_n)}\,(\vec{x}_0) = \Big(\frac{\partial \vec{f}}{\partial x_1}\,(\vec{x}_0), \ldots, \frac{\partial \vec{f}}{\partial x_n}\,(\vec{x}_0)\Big) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}\,(\vec{x}_0) & \ldots & \frac{\partial f_1}{\partial x_n}\,(\vec{x}_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}\,(\vec{x}_0) & \ldots & \frac{\partial f_m}{\partial x_n}\,(\vec{x}_0) \end{bmatrix}}$$

die JACOBI– *oder* **Funktionalmatrix** *von $\vec{f}$ an der Stelle $\vec{x}_0$.*

BSP. (14.2.5) Es sei $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R}^3)$ auf $D(\vec{f}) := \mathbf{R}^2$ gemäß $\vec{f}(x,y) := (2x+y, 3x^2+y^2, xy)^T$ erklärt. Wir berechnen ihre JACOBI–Matrix:

$$J_{\vec{f}}(x,y) = \begin{bmatrix} f_{1,x}(x,y) & f_{1,y}(x,y) \\ f_{2,x}(x,y) & f_{2,y}(x,y) \\ f_{3,x}(x,y) & f_{3,y}(x,y) \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 6x & 2y \\ y & x \end{bmatrix}; \qquad \text{speziell:} \quad J_{\vec{f}}(1,2) = \begin{bmatrix} 2 & 1 \\ 6 & 4 \\ 2 & 1 \end{bmatrix}.$$

BSP. (14.2.6) Die durch die Matrix $A = (\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n) \in \mathbf{R}^{(m,n)}$ definierte vektorwertige Funktion $\vec{f}(\vec{x}) := A\vec{x}$, $\vec{x} \in \mathbf{R}^n$, hat offenbar die partiellen Ableitungen $\frac{\partial \vec{f}}{\partial x_k} = \vec{a}_k$, $k = 1,2,\ldots,n$. Somit resultiert die JACOBI–Matrix

$$J_A(\vec{x}) = A = const \;\; \forall\, \vec{x} \in \mathbf{R}^n.$$

Die Auszeichnung der partiellen Ableitungen in Richtung der Vektoren der Standardbasis des $\mathbf{R}^n$ ist natürlich nicht zwingend. Auch bei vektorwertigen Funktionen kann wie im skalaren Fall eine Ableitung in beliebiger Richtung $\vec{h} \in \mathbf{R}^n$ definiert werden. In Anlehnung an die Definition 13.15 formulieren wir hier:

Definition 14.5 *Gegeben seien eine Funktion $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ und ein innerer Punkt $\vec{x}_0 \in D(\vec{f})$. Ferner sei $\vec{h} \in \mathbf{R}^n$, $\|\vec{h}\| = 1$, ein* **Einheitsvektor**. *Existiert der Grenzwert*

$$\boxed{\frac{\partial \vec{f}}{\partial \vec{h}}\,(\vec{x}_0) := \lim_{t \to 0} \frac{1}{t}\,(\vec{f}(\vec{x}_0 + t\vec{h}) - \vec{f}(\vec{x}_0)) = \frac{d}{dt}\,\vec{f}(\vec{x}_0 + t\vec{h})|_{t=0},}$$

so heiße $\frac{\partial \vec{f}}{\partial \vec{h}}\,(\vec{x}_0)$ die **Richtungsableitung** *von $\vec{f}$ im Punkte $\vec{x}_0$* **in Richtung** *$\vec{h}$.*

BSP. (14.2.7) Die Konsistenz dieser Definition mit der Definition der partiellen Ableitung von $\vec{f}$ nach der k–ten Komponenten ist sichergestellt. Setzen wir nämlich $\vec{h} := \vec{e}_k$ in die obige Definition ein, so folgt korrekt $\frac{\partial \vec{f}}{\partial \vec{e}_k}(\vec{x}_0) = \frac{\partial \vec{f}}{\partial x_k}(\vec{x}_0)$.

BSP. (14.2.8) Die Funktion $\vec{f} \in \mathrm{Abb}(\mathbf{R}^2, \mathbf{R}^2)$ sei auf $D(\vec{f}) := \mathbf{R}^2$ gemäß $\vec{f}(x,y) := (x\cos y, x\sin y)^T$ definiert. Wir berechnen die Richtungsableitung von $\vec{f}$ im Punkte $(x,y) \in \mathbf{R}^2$ in allgemeiner Richtung $\vec{h} = (h_1, h_2)^T$, $\|\vec{h}\| = 1$:

$$\frac{\partial \vec{f}}{\partial \vec{h}}(x,y) = \frac{d}{dt}\left[\begin{array}{c}(x+th_1)\cos(y+th_2)\\(x+th_1)\sin(y+th_2)\end{array}\right]\Big|_{t=0} = \left[\begin{array}{c}h_1\cos y - h_2 x\sin y\\h_1\sin y + h_2 x\cos y\end{array}\right] = \underbrace{\left[\begin{array}{cc}\cos y & -x\sin y\\ \sin y & x\cos y\end{array}\right]}_{=J_{\vec{f}}(x,y)}\vec{h}.$$

Die Gleichung $\frac{\partial \vec{f}}{\partial \vec{h}}(\vec{x}) = J_{\vec{f}}(\vec{x})\,\vec{h}$ hat sich hier nicht zufällig ergeben, wie wir weiter unten in Satz 14.4 begründen werden. Wir erinnern an die Richtungsableitung $\frac{\partial f}{\partial \vec{h}}(\vec{x}) = \langle \mathrm{grad}\, f(\vec{x}), \vec{h}\rangle$ für skalare Funktionen $f \in \mathrm{Abb}(\mathbf{R}^n, \mathbf{R})$. Man kann hier bereits vermuten, dass die JACOBI–Matrix bei vektorwertigen Funktionen dieselbe Rolle spielt, wie der **Gradient** bei skalaren Funktionen.

BSP. (14.2.9) Auf einem Gebiet $\Omega \subset \mathbf{R}^n$ sei eine skalare Funktion $f \in C^2(\Omega)$ gegeben. Dann gilt für jedes feste $k = 1, 2, \ldots, n$:

$$\frac{\partial}{\partial x_k}\,\mathrm{grad}\, f(\vec{x}_0) = (f_{x_1x_k}(\vec{x}_0), f_{x_2x_k}(\vec{x}_0), \ldots, f_{x_nx_k}(\vec{x}_0))^T.$$

Hieraus erschließen wir, dass die JACOBI–Matrix von $\mathrm{grad}\, f$ und die HESSE–Matrix von f im Punkte $\vec{x}_0$ übereinstimmen:

$$J_{\mathrm{grad}\, f}(\vec{x}_0) = \Big(\frac{\partial^2 f}{\partial x_j \partial x_k}(\vec{x}_0)\Big) = H(\vec{x}_0).$$

Bemerkung 14.1 (a) Wie bei skalaren Funktionen kann man auch hier aus der partiellen Differenzierbarkeit der Funktion $\vec{f} \in \mathrm{Abb}(\mathbf{R}^n, \mathbf{R}^m)$ im Punkte $\vec{x}_0$ **nicht** auf die Stetigkeit von $\vec{f}$ in $\vec{x}_0$ schließen.

(b) Das BSP. (14.2.8) lässt erahnen, dass ein geeigneter Ableitungsbegriff ganz analog wie in Definition 13.17 nun mit Hilfe der JACOBI–Matrix formuliert werden kann. □

Um diese Ahnung zu konkretisieren, gehen wir wieder von dem Ansatz aus, eine Approximation der Funktionswerte $\vec{f}(\vec{x})$ einer gegebenen Funktion $\vec{f} \in \mathrm{Abb}(\mathbf{R}^n, \mathbf{R}^m)$ durch eine affine Funktion in der Umgebung eines Punktes $\vec{x}_0 \in D(\vec{f})$ zu bestimmen. Die Schar der affinen Funktionen $\vec{T} \in \mathrm{Abb}(\mathbf{R}^n, \mathbf{R}^m)$ durch den festen Punkt $(\vec{x}_0, \vec{f}(\vec{x}_0)) \in G(\vec{f})$ ist durch

$$\vec{T}(\vec{x}) = \vec{f}(\vec{x}_0) + A(\vec{x} - \vec{x}_0), \qquad \vec{x} \in \mathbf{R}^n, \quad A \in \mathbf{R}^{(m,n)},$$

gegeben. Hiermit macht die folgende Definition einen Sinn.

Definition 14.6 *Gegeben seien eine Funktion $\vec{f} \in \mathrm{Abb}(\mathbf{R}^n, \mathbf{R}^m)$ und ein innerer Punkt $\vec{x}_0 \in D(\vec{f})$. Genau dann heiße $\vec{f}$ in $\vec{x}_0$* **differenzierbar**, *wenn es eine Matrix $A \in \mathbf{R}^{(m,n)}$ gibt mit*

$$\boxed{\vec{f}(\vec{x}) = \vec{f}(\vec{x}_0) + A(\vec{x} - \vec{x}_0) + \vec{\mathcal{O}}(\|\vec{x} - \vec{x}_0\|) \quad \text{für} \quad \vec{x} \to \vec{x}_0.} \tag{2.2}$$

Schreiben wir die gesuchte Matrix $A \in \mathbf{R}^{(m,n)}$ mit Hilfe ihrer **Zeilenvektoren** $\vec{a}_j^{\,T} \in \mathbf{R}^n$, $1 \leq j \leq m$, in der Form

$$A = \begin{bmatrix} \vec{a}_1^{\,T} \\ \vec{a}_2^{\,T} \\ \vdots \\ \vec{a}_m^{\,T} \end{bmatrix},$$

auf, so gestattet die Relation (2.2) die *komponentenweise* Darstellung

$$f_j(\vec{x}) = f_j(\vec{x}_0) + \langle \vec{a}_j, \vec{x} - \vec{x}_0 \rangle + \mathcal{O}(\|\vec{x} - \vec{x}_0\|) \quad \text{für} \quad \vec{x} \to \vec{x}_0, \quad j = 1, 2, \ldots, m. \tag{2.3}$$

Gemäß Definition 13.17 ist der Vektor $\vec{a}_j$ somit genau der Gradient der Funktion f_j im Punkt $\vec{x}_0$. Das heißt, wir haben die Matrix $A \in \mathbf{R}^{(m,n)}$ mit der JACOBI–Matrix von $\vec{f}$ identifiziert:

$$A = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\vec{x}_0) & \ldots & \frac{\partial f_1}{\partial x_n}(\vec{x}_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\vec{x}_0) & \ldots & \frac{\partial f_m}{\partial x_n}(\vec{x}_0) \end{bmatrix} = \frac{\partial(f_1, \ldots, f_m)}{\partial(x_1, \ldots, x_n)}(\vec{x}_0) = J_{\vec{f}}(\vec{x}_0).$$

Aus diesem Zusammenhang erhalten wir:

Satz 14.4 *Gegeben seien eine Funktion* $\vec{f} \in \mathrm{Abb}(\mathbf{R}^n, \mathbf{R}^m)$ *und ein innerer Punkt* $\vec{x}_0 \in D(\vec{f})$.
(a) *Genau dann ist* $\vec{f}$ *in* $\vec{x}_0$ *differenzierbar, wenn jede Komponentenfunktion* $f_j : D(\vec{f}) \to \mathbf{R}$, $j = 1, 2, \ldots, m$, *in* $\vec{x}_0$ *differenzierbar ist. Die* **Ableitung** *von* $\vec{f}$ *in* $\vec{x}_0$ *ist die* JACOBI*–Matrix*

$$\boxed{J_{\vec{f}}(\vec{x}_0) = \frac{\partial(f_1, \ldots, f_m)}{\partial(x_1, \ldots, x_n)}(\vec{x}_0) \stackrel{\text{def}}{=} \frac{d\vec{f}}{d\vec{x}}(\vec{x}_0) \stackrel{\text{def}}{=} \vec{f}\,'(\vec{x}_0).}$$

(b) *Ist* $\vec{f}$ *in* $\vec{x}_0$ *differenzierbar, so existieren die* **Richtungsableitungen** *von* $\vec{f}$ *im Punkte* $\vec{x}_0$ *in* **jeder** *Richtung* $\vec{h} \in \mathbf{R}^n$, $\|\vec{h}\| = 1$, *und es gilt*

$$\boxed{\frac{\partial \vec{f}}{\partial \vec{h}}(\vec{x}_0) = J_{\vec{f}}(\vec{x}_0)\,\vec{h}.} \tag{2.4}$$

Speziell für $\vec{h} := \vec{e}_j$ *existieren also die partiellen Ableitungen* $\frac{\partial \vec{f}}{\partial x_j}(\vec{x}_0)$, *und es gilt*

$$\boxed{\frac{\partial \vec{f}}{\partial x_j}(\vec{x}_0) = J_{\vec{f}}(\vec{x}_0)\,\vec{e}_j = \Big(\frac{\partial f_1}{\partial x_j}(\vec{x}_0), \frac{\partial f_2}{\partial x_j}(\vec{x}_0), \ldots, \frac{\partial f_m}{\partial x_j}(\vec{x}_0)\Big)^T, \quad 1 \leq j \leq n.} \tag{2.5}$$

(c) *Ist die Funktion* $\vec{f}$ *im Punkte* $\vec{x}_0$ *differenzierbar, so ist sie dort auch* **stetig**.

(d) *Besitzt* $\vec{f}$ *im Punkte* $\vec{x}_0$ **stetige** *partielle Ableitungen* $\frac{\partial \vec{f}}{\partial x_j}(\vec{x}_0)$, $j = 1, 2, \ldots, n$, *so ist* $\vec{f}$ *in* $\vec{x}_0$ *auch differenzierbar.*

Begründungen: Die Aussage ergibt (a) ergibt sich aus dem Vorspann des Satzes. Die Aussagen (b)–(d) erhält man aus Satz 13.14 durch Anwendung auf jede Komponentenfunktion f_j einzeln. □

BSP. (14.2.10) Es seien $r := \|\vec{x}\|$, $\vec{x} \in \mathbf{R}^n$, sowie $\vec{f}(\vec{x}) := \mathrm{grad}\,(\ln r)$, $r > 0$. Dann gilt wie in BSP. (13.5.2) gezeigt $\vec{f}(\vec{x}) = \frac{\vec{x}}{r^2}$, $r > 0$, und somit

$$J_{\vec{f}}(\vec{x}) = \frac{1}{r^4}\left(r^2\delta_{jk} - 2x_j x_k\right)_{\substack{j=1,\ldots,n \\ k=1,\ldots,n}} = \frac{1}{r^2}\,Id_n - \frac{2}{r^4}\,(x_j x_k)_{\substack{j=1,\ldots,n \\ k=1,\ldots,n}} = \frac{1}{r^2}\Big(Id_n - \frac{2}{r^2}\,(\vec{x} \otimes \vec{x})\Big).$$

Für einen beliebigen Einheitsvektor $\vec{h} \in \mathbf{R}^n$, $\|\vec{h}\| = 1$, ergibt sich daraus die Richtungsableitung

$$\frac{\partial \vec{f}}{\partial \vec{h}}(\vec{x}) = \frac{1}{r^2}\,\vec{h} - \frac{2}{r^4}\,\langle \vec{h}, \vec{x}\rangle\,\vec{x} = \frac{1}{r^2}\left(Id_n - \frac{2}{r^2}\,(\vec{x} \otimes \vec{x})\right)\vec{h}.$$

14.3 Rechenregeln für differenzierbare Funktionen $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$

Sind die Funktionen $\vec{f}, \vec{g} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ in einem inneren Punkt $\vec{x}_0 \in D(\vec{f}) \cap D(\vec{g})$ differenzierbar, so macht es keine Mühe einzusehen, dass die folgenden Verknüpfungen ebenfalls in $\vec{x}_0$ differenzierbar sind:

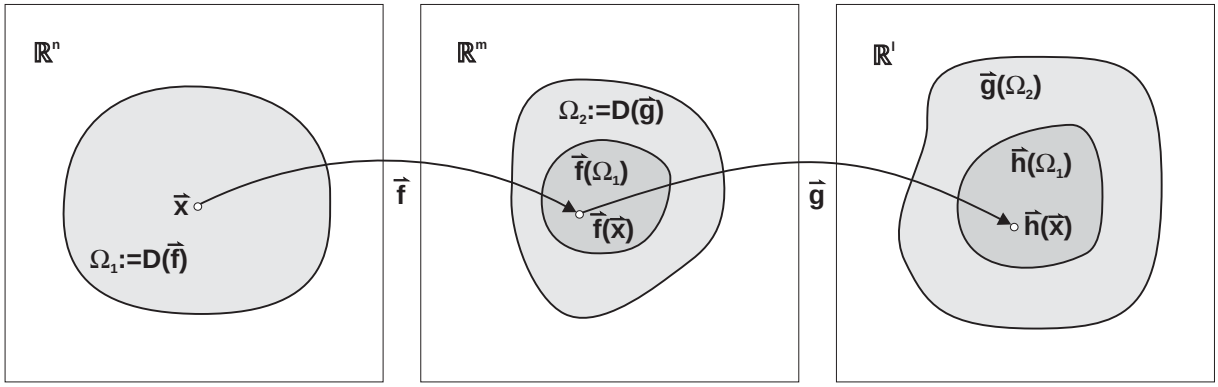
$$\lambda\,\vec{f} + \mu\,\vec{g}\ \ \forall\ \lambda, \mu \in \mathbf{R}, \qquad \langle \vec{f}, \vec{g}\rangle, \qquad \|\vec{f}\|, \qquad \vec{f} \times \vec{g} \quad \text{für} \quad m = 3.$$

Natürlich gelten die bekannten Formeln $(\lambda\,\vec{f} + \mu\,\vec{g})'(\vec{x}_0) = \lambda\,\vec{f}\,'(\vec{x}_0) + \mu\,\vec{g}\,'(\vec{x}_0)$ usw. Interessant ist in diesem Zusammenhang die Differentiation des **Kompositums** $\vec{g} \circ \vec{f}$ für Funktionen $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ und $\vec{g} \in \mathrm{Abb}\,(\mathbf{R}^m, \mathbf{R}^l)$. Im skalaren Fall $n = m = l = 1$ gilt ja die einfache Kettenregel

$$\frac{d}{dx}\,(g \circ f)(x) = \frac{d}{dx}\,g(f(x)) = g'(f(x)) \cdot f'(x),$$

sofern f in $x \in D(f)$ und g in $f(x) \in D(g)$ differenzierbar sind; *zum Beispiel* $\frac{d}{dx}\,(\mathrm{arc\,tan}_H\, e^{ax}) = \frac{1}{1+e^{2ax}} \cdot ae^{ax}$.

Erste Verallgemeinerungen dieser elementaren Kettenregel haben wir bereits in Satz 13.16 mit den Formeln (5.10(c) und (d)) getroffen. Sind nun $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ und $\vec{g} \in \mathrm{Abb}\,(\mathbf{R}^m, \mathbf{R}^l)$ unter der Nebenbedingung $\vec{f}(D(\vec{f})) \subset D(\vec{g})$ vorgegeben, so induziert die Hintereinanderausführung $\vec{h}(\vec{x}) := \vec{g}(\vec{f}(\vec{x}))$ eine vektorwertige Funktion $\vec{h} = \vec{g} \circ \vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^l)$, für die wir die folgende **Kettenregel** formulieren.



Das Kompositum $\vec{h} := \vec{g} \circ \vec{f}$ der zwei Funktionen $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ und $\vec{g} \in \mathrm{Abb}\,(\mathbf{R}^m, \mathbf{R}^l)$

Satz 14.5 (Kettenregel)
Die Funktionen $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ und $\vec{g} \in \mathrm{Abb}\,(\mathbf{R}^m, \mathbf{R}^l)$ seien in den inneren Punkten $\vec{x}_0 \in D(\vec{f})$ bzw. in $\vec{y}_0 := \vec{f}(\vec{x}_0) \in D(\vec{g})$ differenzierbar. Dann ist das Kompositum $\vec{h} := \vec{g} \circ \vec{f}$ in $\vec{x}_0$ differenzierbar, und es gilt im Sinne der Matrizenmultiplikation

$$\boxed{J_{\vec{h}}(\vec{x}_0) = \frac{d\vec{h}}{d\vec{x}}\,(\vec{x}_0) = \frac{d\vec{g}}{d\vec{y}}\,(\vec{y}_0) \cdot \frac{d\vec{f}}{d\vec{x}}\,(\vec{x}_0) = J_{\vec{g}}(\vec{y}_0) \cdot J_{\vec{f}}(\vec{x}_0).}$$

Das heißt, in expliziten Formeln gilt:

$$\vec{h}\,'(\vec{x}_0) = \left[\begin{array}{ccc} \frac{\partial h_1}{\partial x_1} & \cdots & \frac{\partial h_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_l}{\partial x_1} & \cdots & \frac{\partial h_l}{\partial x_n} \end{array}\right]_{|_{\vec{x}=\vec{x}_0}} = \left[\begin{array}{ccc} \frac{\partial g_1}{\partial y_1} & \cdots & \frac{\partial g_1}{\partial y_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_l}{\partial y_1} & \cdots & \frac{\partial g_l}{\partial y_m} \end{array}\right]_{|_{\vec{y}=\vec{f}(\vec{x}_0)}} \left[\begin{array}{ccc} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{array}\right]_{|_{\vec{x}=\vec{x}_0}} .$$

Gut einprägsam ist die Kettenregel in der Form

$$\underbrace{\mathbf{R}^n}_{\ni \vec{x}} \sim\sim\stackrel{\vec{f}}{\sim}\sim\rightsquigarrow \underbrace{\mathbf{R}^m}_{\ni \vec{y}} \sim\sim\stackrel{\vec{g}}{\sim}\sim\rightsquigarrow \underbrace{\mathbf{R}^l}_{\ni \vec{z}}$$

$$\frac{\partial(z_1,\ldots,z_l)}{\partial(x_1,\ldots,x_n)} = \frac{\partial(z_1,\ldots,z_l)}{\partial(y_1,\ldots,y_m)} \cdot \frac{\partial(y_1,\ldots,y_m)}{\partial(x_1,\ldots,x_n)}$$

BSP. (14.3.1) Wir berechnen die Ableitung einer Funktion $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R}^l)$ bei Transformation auf **Polarkoordinaten** $\vec{p}(r,\varphi) := (r\cos\varphi, r\sin\varphi)^T$ mit $r \geq 0$ und $0 \leq \varphi < 2\pi$. Es liegt folgendes Abbildungsschema vor

$$\underbrace{\mathbf{R}^2}_{\ni (r,\varphi)^T} \sim\sim\stackrel{\vec{p}}{\sim}\sim\rightsquigarrow \underbrace{\mathbf{R}^2}_{\ni (x,y)^T} \sim\sim\stackrel{\vec{f}}{\sim}\sim\rightsquigarrow \underbrace{\mathbf{R}^l}_{\ni (f_1,\ldots,f_l)^T} .$$

Die Funktion $\vec{f}$ sei zum Beispiel für $l = 3$ gemäß $\vec{f}(x,y) := (xy, x+y^2, x-y)^T$ spezifiziert. Wir berechnen

$$\vec{f}\,'(r,\varphi) = \frac{\partial(f_1,f_2,f_3)}{\partial(r,\varphi)} = \frac{\partial(f_1,f_2,f_3)}{\partial(x,y)} \cdot \frac{\partial(x,y)}{\partial(r,\varphi)} = \left[\begin{array}{cc} y & x \\ 1 & 2y \\ 1 & -1 \end{array}\right] \left[\begin{array}{cc} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{array}\right]$$

$$= \left[\begin{array}{cc} r\sin\varphi & r\cos\varphi \\ 1 & 2r\sin\varphi \\ 1 & -1 \end{array}\right] \left[\begin{array}{cc} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{array}\right] = \left[\begin{array}{cc} r\sin 2\varphi & r^2\cos 2\varphi \\ \cos\varphi + 2r\sin^2\varphi & -r\sin\varphi + r^2\sin 2\varphi \\ \cos\varphi - \sin\varphi & -r(\sin\varphi + \cos\varphi) \end{array}\right].$$

Wir gelangen zum selben Resultat, wenn wir zuerst das Kompositum $\vec{h} := \vec{f} \circ \vec{p}$ berechnen und danach die JACOBI–Matrix der Funktion $\vec{h}$ bestimmen:

$$\vec{h}(r,\varphi) = \left[\begin{array}{c} r^2\sin\varphi\cos\varphi \\ r\cos\varphi + r^2\sin^2\varphi \\ r\cos\varphi - r\sin\varphi \end{array}\right], \quad J_{\vec{h}}(r,\varphi) = \left[\begin{array}{cc} r\sin 2\varphi & r^2\cos 2\varphi \\ \cos\varphi + 2r\sin^2\varphi & -r\sin\varphi + r^2\sin 2\varphi \\ \cos\varphi - \sin\varphi & -r(\sin\varphi + \cos\varphi) \end{array}\right].$$

BSP. (14.3.2) Als ein Sonderfall der Kettenregel erhält man für eine differenzierbare Funktion $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ und für einen differenzierbaren Weg $\vec{x} \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R}^n)$ die **Wegableitung** von $\vec{f}$ **längs des Weges** $\vec{x}$:

$$\frac{d}{dt}\,\vec{f}(\vec{x}(t)) = \frac{\partial(f_1,\ldots,f_m)}{\partial(x_1,\ldots,x_n)}\,\dot{\vec{x}}(t).$$

Es sei *zum Beispiel* $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R}^3)$ die Funktion aus BSP. (14.3.1) und $\vec{x}(t)$ der Ellipsenbogen $\vec{x}(t) := (a\cos t, b\sin t)^T$, $0 \leq t < 2\pi$. Dann ist die Wegableitung von $\vec{f}$ längs des Weges $\vec{x}$ wie folgt bestimmt:

$$\frac{d}{dt}\,\vec{f}(\vec{x}(t)) = \left[\begin{array}{cc} y & x \\ 1 & 2y \\ 1 & -1 \end{array}\right] \left[\begin{array}{c} -a\sin t \\ b\cos t \end{array}\right] = \left[\begin{array}{cc} b\sin t & a\cos t \\ 1 & 2b\sin t \\ 1 & -1 \end{array}\right] \left[\begin{array}{c} -a\sin t \\ b\cos t \end{array}\right] = \left[\begin{array}{c} ab\cos 2t \\ -a\sin t + b^2\sin 2t \\ -a\sin t - b\cos t \end{array}\right].$$

BSP. (14.3.3) Bei **Koordinatenwechsel** $\vec{x} \mapsto \vec{y}$ transformieren sich die partiellen Ableitungen einer skalaren Funktion $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ in der folgenden Weise:

$$\boxed{\frac{\partial f}{\partial y_j}\,(\vec{x}(\vec{y})) = f_{x_1}\,\frac{\partial x_1}{\partial y_j} + f_{x_2}\,\frac{\partial x_2}{\partial y_j} + \cdots + f_{x_n}\,\frac{\partial x_n}{\partial y_j}, \quad j = 1, 2, \ldots, n.}$$

Wir betrachten *zum Beispiel* den Koordinatenwechsel $(x, y)^T \mapsto (r, \varphi)^T$ von kartesischen Koordinaten auf Polarkoordinaten $x = r\cos\varphi$, $y = r\sin\varphi$ bei gegebener differenzierbarer Funktion $f \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R})$:

$$\left.\begin{array}{rclcl} f_r & = & f_x\,x_r + f_y\,y_r & = & f_x\,\cos\varphi + f_y\,\sin\varphi \\ f_\varphi & = & f_x\,x_\varphi + f_y\,y_\varphi & = & -f_x\,r\sin\varphi + f_y\,r\cos\varphi \end{array}\right\} \;\Rightarrow\; (f_r, f_\varphi) = (f_x, f_y)\begin{bmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{bmatrix}.$$

Wir wenden die Kettenregel im speziellen Fall einer **injektiven** Funktion $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^n)$ an. Da $\vec{f}$ auf der Bildmenge von $\vec{f}(D(\vec{f}))$ invertierbar ist, existiert die Inverse $\vec{f}^{-1} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^n)$, und es gilt $\vec{x} = \vec{f}^{-1}(\vec{f}(\vec{x}))\ \forall\ \vec{x} \in D(\vec{f})$. Das heißt, setzen wir $g(\vec{y}) := \vec{f}^{-1}(\vec{y})$ und nehmen wir $\vec{f}, \vec{f}^{-1} \in C^1$ an, so erhalten wir aus der Kettenregel des Satzes 14.5 die Relation

$$Id = J_{\vec{f}^{-1}}(\vec{y}) \cdot J_{\vec{f}}(\vec{x})\ \ \forall\ \vec{x} \in D(\vec{f}), \quad \vec{y} := \vec{f}(\vec{x}),$$

und somit

$$\boxed{J_{\vec{f}^{-1}}(\vec{y}) = \left(J_{\vec{f}}(\vec{f}^{-1}(\vec{y}))\right)^{-1}.} \tag{3.1}$$

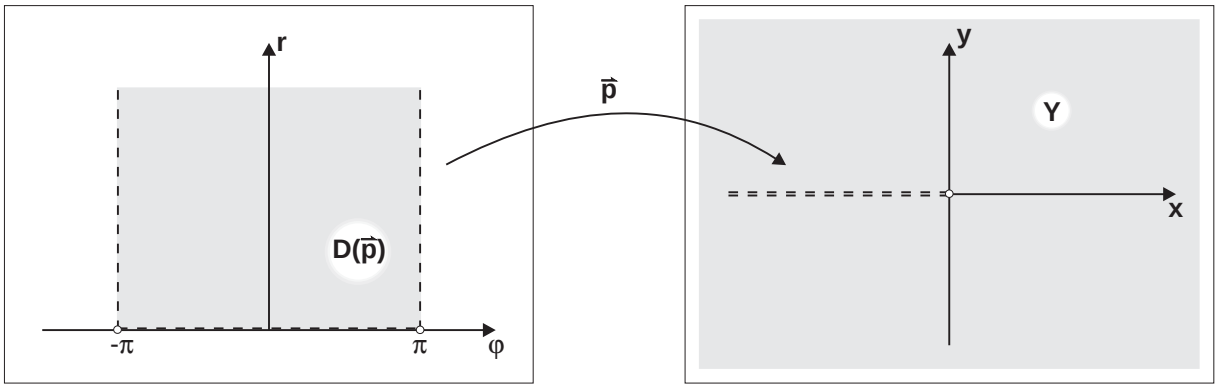
Folgerung 14.1 *Die stetig differenzierbare Funktion* $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^n)$ *sei invertierbar. Dann muss* **notwendig** *die Bedingung* $\det J_{\vec{f}}(\vec{x}) \neq 0\ \forall\ \vec{x} \in D(\vec{f})$ *gelten.*

Im folgenden Abschnitt werden wir zeigen, dass diese Bedingung auch **hinreichend** ist. Zusammen mit der Regularitätsvoraussetzung $\vec{f} \in C^1$ erhalten wir dann die Existenz einer Inversen $\vec{f}^{-1} \in C^1$, deren Ableitung man gemäß (3.1) durch Invertierung der JACOBI–Matrix von $\vec{f}$ gewinnt.

BSP. (14.3.4) Wir betrachten wieder ebene **Polarkoordinaten** $\vec{p}(r, \varphi) := (r\cos\varphi, r\sin\varphi)^T$, und zwar auf dem Definitionsbereich

$$D(\vec{p}) := \{(r, \varphi) \in \mathbf{R}^2\ :\ 0 < r,\ \ -\pi < \varphi < \pi\}.$$

Der Bildbereich ist nun die Teilmenge $Y := \vec{p}(D(\vec{p})) = \mathbf{R}^2 \setminus \{(x, 0)\ :\ x \leq 0\}$, siehe folgende Skizze.



Der Definitionsbereich von ebenen Polarkoordinaten

Der Bildbereich von ebenen Polarkoordinaten

Man überzeugt sich unmittelbar davon, dass die Abbildung $\vec{p} : D(\vec{p}) \to Y$ **bijektiv** ist und die folgende Umkehrfunktion besitzt:

$$\vec{p}^{\,-1}(x,y) = \begin{bmatrix} \sqrt{x^2+y^2} \\ (\operatorname{sign} y)\operatorname{arc\,cos}_H \frac{x}{\sqrt{x^2+y^2}} \end{bmatrix} = \begin{bmatrix} r \\ \varphi \end{bmatrix}.$$

Die JACOBI–Matrix und ihre Inverse berechnen sich wie folgt:

$$J_{\vec{p}}(r,\varphi) = \frac{\partial(x,y)}{\partial(r,\varphi)} = \begin{bmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{bmatrix}, \quad \left(J_{\vec{p}}(r,\varphi)\right)^{-1} = \begin{bmatrix} \cos\varphi & \sin\varphi \\ -\frac{1}{r}\sin\varphi & \frac{1}{r}\cos\varphi \end{bmatrix}.$$

Somit gilt

$$J_{\vec{p}^{\,-1}}(x,y) = \Big(J_{\vec{p}}(r(x,y),\varphi(x,y))\Big)^{-1} = \frac{\partial(r,\varphi)}{\partial(x,y)} = \begin{bmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{bmatrix}.$$

Wie die obige Rechnung zeigt, kann die JACOBI–Matrix $J_{\vec{p}^{\,-1}}(r,\varphi)$ **ohne explizite Kenntnis** der Umkehrfunktion $\vec{p}^{\,-1}$ durch Invertierung der JACOBI–Matrix $J_{\vec{p}}(r,\varphi)$ gewonnen werden. Wegen (3.1) gilt dieser Zusammenhang allgemein, und man verwendet ihn bei **Koordinatenwechseln**:

$$\underbrace{\mathbf{R}^n}_{\ni \vec{x}} \sim\sim\overset{\vec{p}^{\,-1}}{\sim}\sim\sim\to \underbrace{\mathbf{R}^n}_{\ni \vec{y}} \sim\sim\overset{\vec{p}}{\sim}\sim\sim\to \underbrace{\mathbf{R}^n}_{\ni \vec{x}}.$$

Hat man im Vektorraum $\mathbf{R}^n$ ein neues Koordinatensystem $\vec{y} = \vec{p}^{\,-1}(\vec{x})$, so können die Ableitungen einer Funktion $\vec{f} \in \operatorname{Abb}(\mathbf{R}^n, \mathbf{R}^m)$ nach den alten Koordinaten $\vec{x}$ durch die Ableitungen nach den neuen Koordinaten $\vec{y}$ in der folgenden Weise ausgedrückt werden:

$$\frac{d\vec{f}}{d\vec{x}}(\vec{y}) = \frac{\partial(f_1,\ldots,f_m)}{\partial(x_1,\ldots,x_n)} = \frac{d(\vec{f}\circ\vec{p}^{\,-1})}{d\vec{x}}(\vec{x}) \overset{\text{Satz\_14.5}}{=} \frac{\partial(f_1,\ldots,f_m)}{\partial(y_1,\ldots,y_n)} \cdot \frac{d\vec{p}^{\,-1}}{d\vec{x}}(\vec{p}(\vec{y})).$$

Hier setzen wir stetige Differenzierbarkeit der Funktionen $\vec{f}, \vec{p}, \vec{p}^{\,-1}$ voraus. Wegen (3.1) wird die Umkehrfunktion $\vec{p}^{\,-1}$ aber gar nicht benötigt; man erhält unter Verwendung der inversen JACOBI–Matrix von $\vec{p}$:

$$\boxed{\frac{d\vec{f}}{d\vec{x}}(\vec{y}) = \frac{\partial(f_1,\ldots,f_m)}{\partial(x_1,\ldots,x_n)} = \frac{\partial(f_1,\ldots,f_m)}{\partial(y_1,\ldots,y_n)} \cdot \Big(\frac{\partial(x_1,\ldots,x_n)}{\partial(y_1,\ldots,y_n)}\Big)^{-1}.} \tag{3.2}$$

BSP. (14.3.5) Koordinatenwechsel in $\mathbf{R}^3$ auf **Zylinderkoordinaten**. Wir haben hier

$$\vec{x} = (x,y,z)^T = \vec{p}(r,\varphi,z) := (r\cos\varphi, r\sin\varphi, z)^T, \quad D(\vec{p}) := \{(r,\varphi,z) \,:\, 0 < r,\ 0 \le \varphi < 2\pi,\ z \in \mathbf{R}\}.$$

Die JACOBI–Matrix von $\vec{p}$ ist wegen

$$\det\Big(\frac{\partial(x,y,z)}{\partial(r,\varphi,z)}\Big) = \begin{vmatrix} \cos\varphi & -r\sin\varphi & 0 \\ \sin\varphi & r\cos\varphi & 0 \\ 0 & 0 & 1 \end{vmatrix} = r \neq 0 \quad \text{auf ganz } D(\vec{p})$$

invertierbar, und ihre Inverse ist die Matrix

$$\Big(\frac{\partial(x,y,z)}{\partial(r,\varphi,z)}\Big)^{-1} = \begin{bmatrix} \cos\varphi & \sin\varphi & 0 \\ -\frac{1}{r}\sin\varphi & \frac{1}{r}\cos\varphi & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad r > 0.$$

Für eine differenzierbare Funktion $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^3, \mathbf{R})$ erhalten wir somit in Zylinderkoordinaten (r, φ, z):

$$\begin{aligned}(f_x, f_y, f_z) &= (f_r, f_\varphi, f_z)\Big(\frac{\partial(x,y,z)}{\partial(r,\varphi,z)}\Big)^{-1} = (f_r, f_\varphi, f_z)\begin{bmatrix} \cos\varphi & \sin\varphi & 0 \\ -\frac{1}{r}\sin\varphi & \frac{1}{r}\cos\varphi & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \Big(f_r\cos\varphi - \frac{1}{r}f_\varphi\sin\varphi, f_r\sin\varphi + \frac{1}{r}f_\varphi\cos\varphi, f_z\Big).\end{aligned}$$

Der **Gradient** einer differenzierbaren Funktion $f \in \mathrm{Abb}\,(\mathbf{R}^3, \mathbf{R})$ hat also in Zylinderkoordinaten die folgende Darstellung:

$$\boxed{\mathrm{grad}\, f(r,\varphi,z) = \begin{bmatrix} f_x(r,\varphi,z) \\ f_y(r,\varphi,z) \\ f_z(r,\varphi,z) \end{bmatrix} = \begin{bmatrix} f_r\cos\varphi - \frac{1}{r}f_\varphi\sin\varphi \\ f_r\sin\varphi + \frac{1}{r}f_\varphi\cos\varphi \\ f_z \end{bmatrix}.}$$

BSP. (14.3.6) Koordinatenwechsel in $\mathbf{R}^3$ auf **Kugelkoordinaten**. Wir haben nun

$$\vec{x} = (x,y,z)^T = \vec{p}(r,\varphi,\vartheta) := (r\cos\varphi\sin\vartheta, r\sin\varphi\sin\vartheta, r\cos\vartheta)^T$$

mit dem Definitionsbereich

$$D(\vec{p}) := \{(r,\varphi,\vartheta) \,:\, 0 < r,\ 0 \le \varphi < 2\pi,\ 0 < \vartheta < \pi\}.$$

Die JACOBI–Matrix von $\vec{p}$ ist wegen

$$\det\Big(\frac{\partial(x,y,z)}{\partial(r,\varphi,\vartheta)}\Big) = \begin{vmatrix} \cos\varphi\sin\vartheta & -r\sin\varphi\sin\vartheta & r\cos\varphi\cos\vartheta \\ \sin\varphi\sin\vartheta & r\cos\varphi\sin\vartheta & r\sin\varphi\cos\vartheta \\ \cos\vartheta & 0 & -r\sin\vartheta \end{vmatrix} = -r^2\sin\vartheta \neq 0 \quad \text{auf ganz } D(\vec{p})$$

invertierbar, und ihre Inverse ist die Matrix

$$\Big(\frac{\partial(x,y,z)}{\partial(r,\varphi,\vartheta)}\Big)^{-1} = \begin{bmatrix} \cos\varphi\sin\vartheta & \sin\varphi\sin\vartheta & \cos\vartheta \\ -\frac{\sin\varphi}{r\sin\vartheta} & \frac{\cos\varphi}{r\sin\vartheta} & 0 \\ \frac{1}{r}\cos\varphi\cos\vartheta & \frac{1}{r}\sin\varphi\cos\vartheta & -\frac{1}{r}\sin\vartheta \end{bmatrix}, \quad r > 0,\ \vartheta \in (0,\pi).$$

Für eine differenzierbare Funktion $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^3, \mathbf{R})$ erhalten wir also in Kugelkoordinaten (r, φ, ϑ):

$$(f_x, f_y, f_z) = (f_r, f_\varphi, f_\vartheta)\Big(\frac{\partial(x,y,z)}{\partial(r,\varphi,\vartheta)}\Big)^{-1}$$

mit den expliziten Formeln

$$\begin{aligned} f_x(r,\varphi,\vartheta) &= f_r\cos\varphi\sin\vartheta - f_\varphi\frac{\sin\varphi}{r\sin\vartheta} + f_\vartheta\frac{1}{r}\cos\varphi\cos\vartheta, \\ f_y(r,\varphi,\vartheta) &= f_r\sin\varphi\sin\vartheta + f_\varphi\frac{\cos\varphi}{r\sin\vartheta} + f_\vartheta\frac{1}{r}\sin\varphi\cos\vartheta, \\ f_z(r,\varphi,\vartheta) &= f_r\cos\vartheta - f_\vartheta\frac{1}{r}\sin\vartheta. \end{aligned}$$

Der **Gradient** einer differenzierbaren Funktion $f \in \mathrm{Abb}\,(\mathbf{R}^3, \mathbf{R})$ hat somit in Kugelkoordinaten die folgende Darstellung:

$$\boxed{\operatorname{grad} f(r,\varphi,\vartheta) = \begin{bmatrix} f_r \cos\varphi \sin\vartheta - f_\varphi \frac{\sin\varphi}{r\sin\vartheta} + f_\vartheta \frac{1}{r}\cos\varphi\cos\vartheta \\ f_r \sin\varphi\sin\vartheta + f_\varphi \frac{\cos\varphi}{r\sin\vartheta} + f_\vartheta \frac{1}{r}\sin\varphi\cos\vartheta \\ f_r\cos\vartheta - f_\vartheta \frac{1}{r}\sin\vartheta \end{bmatrix}.}$$

Gilt $f \in C^2(\mathbf{R}^3)$, so können in gleicher Weise auch die zweiten Ableitungen auf Kugelkoordinaten (r,φ,ϑ) umgerechnet werden. Es gilt

$$H(r,\varphi,\vartheta) = \begin{bmatrix} f_{xx} & f_{xy} & f_{xz} \\ f_{xy} & f_{yy} & f_{yz} \\ f_{xz} & f_{yz} & f_{zz} \end{bmatrix} = \frac{\partial(f_x,f_y,f_z)}{\partial(x,y,z)} = \frac{\partial(f_x,f_y,f_z)}{\partial(r,\varphi,\vartheta)} \cdot \Big(\frac{\partial(x,y,z)}{\partial(r,\varphi,\vartheta)}\Big)^{-1}.$$

Die Rechnungen werden sehr umfangreich; sie sollen hier nicht vorgeführt werden. Man kann auf diese Weise zum Beispiel Differentialoperatoren auf neue Koordinaten umrechnen. Im Fall von Kugelkoordinaten berechnet man

$$\Delta f(r,\varphi,\vartheta) = \frac{2}{r}\frac{\partial f}{\partial r} + \frac{\partial^2 f}{\partial r^2} + \frac{\cot\vartheta}{r^2}\frac{\partial f}{\partial\vartheta} + \frac{1}{r^2}\frac{\partial^2 f}{\partial\vartheta^2} + \frac{1}{r^2\sin^2\vartheta}\frac{\partial^2 f}{\partial\varphi^2};$$

das heißt, der LAPLACE–Operator $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$ hat in räumlichen Kugelkoordinaten die Darstellung

$$\boxed{\Delta = \frac{2}{r}\frac{\partial}{\partial r} + \frac{\partial^2}{\partial r^2} + \frac{\cot\vartheta}{r^2}\frac{\partial}{\partial\vartheta} + \frac{1}{r^2}\frac{\partial^2}{\partial\vartheta^2} + \frac{1}{r^2\sin^2\vartheta}\frac{\partial^2}{\partial\varphi^2}.}$$

14.4 Nichtlineare Gleichungssysteme und der Satz über implizite Funktionen

Das Beispiel des **Koordinatenwechsels** in $\mathbf{R}^n$ weist schon auf die Wichtigkeit der Frage nach der Existenz der **inversen Abbildung** einer gegebenen Funktion $\vec{h} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^n)$ hin. Wir verallgemeinern die Problemstellung, indem wir generell nach Lösungen $\vec{y} \in D(\vec{h})$ von **nichtlinearen Gleichungssystemen**

$$\boxed{\vec{h}(\vec{y}) = \vec{x}, \qquad \vec{x} \in \mathbf{R}^n \quad \text{fest},} \tag{4.1}$$

mit einer gegebenen Vektorfunktion $\vec{h} \in \mathrm{Abb}\,(\mathbf{R}^m, \mathbf{R}^n)$ fragen. Äquivalent mit der Problemstellung (4.1) ist die Lösung der Gleichung

$$\vec{F}(\vec{x},\vec{y}) := \vec{h}(\vec{y}) - \vec{x} = \vec{0}$$

nach der gesuchten Variablen $\vec{y}$. Das heißt, eine mögliche Lösungsmenge der Gleichung (4.1) bildet die spezielle Äquipotentialfläche $\ddot{A}P := \{(\vec{x},\vec{y}) \in \mathbf{R}^n \times \mathbf{R}^m \,:\, \vec{F}(\vec{x},\vec{y}) = \vec{0}\}$. Es bleibt zu prüfen, ob die Fläche $\ddot{A}P$ der **Graph** einer Funktion $\vec{y} = \vec{f}(\vec{x})$, $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$, ist. Die Funktion $\vec{f}$ heiße in diesem Fall **implizit durch die Gleichung** $\vec{F}(\vec{x},\vec{y}) = \vec{0}$ **definiert**.

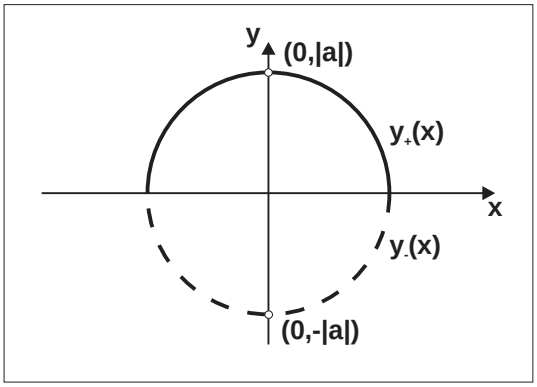
Im simpelsten Fall hat man die Frage zu beantworten, ob die skalare implizite Gleichung $F(x,y) = 0$ zu vorgegebenem $F \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R})$ (eindeutig) nach y auflösbar ist.

BSP. (14.4.1) Für festes $a \in \mathbf{R}$ betrachten wir die Funktion $F(x,y) := x^2 + y^2 + a^2$, $x, y \in \mathbf{R}$. Die Gleichung $F(x,y) = 0$ führt auf $y^2 = -x^2 - a^2$, und man erkennt sofort, dass es für $a \neq 0$ keine reellen Lösungen gibt. Das heißt, die durch $F(x,y) = 0$ definierte ÄP ist als Teilmenge von $\mathbf{R} \times \mathbf{R}$ leer.

BSP. (14.4.2) Wir ändern die Funktion F aus dem vorangegangenen Beispiel ein wenig ab: Es sei nun $F(x,y) := x^2 + y^2 - a^2$, $a \in \mathbf{R}$ fest. Die durch die Gleichung $F(x,y) = 0$ definierte Äquipotentiallinie ist **lokal eindeutig** durch die beiden Funktionen

$$f_+(x) := +\sqrt{a^2 - x^2}, \quad f_-(x) := -\sqrt{a^2 - x^2}, \qquad -|a| \leq x \leq |a|,$$

darstellbar. Hier sind die Funktionen $y_\pm = f_\pm(x)$ in einer Umgebung der Punkte $(0, +|a|)$ bzw. $(0, -|a|)$ jeweils die eindeutigen Lösungen der Gleichung $F(x,y) = 0$. Wir sagen, die Gleichung $F(x,y) = 0$ ist **lokal nach y auflösbar**.



Die eindeutigen Lösungszweige der Gleichung $x^2 + y^2 - a^2 = 0$

BSP. (14.4.3) Es seien Matrizen $A = (a_{jk}) \in \mathbf{R}^{(m,n)}$ und $B = (b_{jk}) \in \mathbf{R}^{(m,m)}$ sowie ein fester Vektor $\vec{c} \in \mathbf{R}^m$ gegeben. Es sei ferner $\vec{F} \in \mathrm{Abb}\,(\mathbf{R}^n \times \mathbf{R}^m, \mathbf{R}^m)$ die affine Funktion

$$\vec{F}(\vec{x}, \vec{y}) := A\vec{x} + B\vec{y} - \vec{c}, \qquad \vec{x} \in \mathbf{R}^n, \quad \vec{y} \in \mathbf{R}^m. \tag{4.2}$$

Wird $\det B \neq 0$ angenommen, so existiert die inverse Matrix B^{-1}, und die Gleichung $\vec{F}(\vec{x}, \vec{y}) = \vec{0}$ ist jetzt **global eindeutig nach $\vec{y}$ auflösbar**:

$$\vec{y} = \vec{f}(\vec{x}) := -B^{-1}A\vec{x} + B^{-1}\vec{c}, \qquad \vec{x} \in \mathbf{R}^n.$$

Bemerkung 14.2 (a) Sicher ist die Bedingung

$$\exists\, (\vec{x}_0, \vec{y}_0) \in \mathbf{R}^n \times \mathbf{R}^m \;:\; \vec{F}(\vec{x}_0, \vec{y}_0) = \vec{0} \tag{4.3}$$

notwendig für die lokale Auflösbarkeit der Gleichung $\vec{F}(\vec{x}, \vec{y}) = \vec{0}$ nach der Veränderlichen $\vec{y}$. Andernfalls wäre die durch $\vec{F}(\vec{x}, \vec{y}) = \vec{0}$ definierte ÄP leer, wie wir in BSP. (14.4.1) gesehen haben. In BSP. (14.4.2) ist die Bedingung (4.3) zum Beispiel in den Punkten $(x_0, y_0) := (0, +|a|)$ und $(x_0, y_0) := (0, -|a|)$ erfüllt. In BSP. (14.4.3) gilt dies im Punkte $(\vec{x}_0, \vec{y}_0) := (\vec{0}, B^{-1}\vec{c})$.

(b) Wir hatten in Abschnitt 14.3 erkannt, dass die Existenz einer Inversen $\vec{h}^{-1} \in C^1$ der Funktion $\vec{h} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^n)$ neben der Regularität $\vec{h} \in C^1$ **notwendig** die Invertierbarkeit der JACOBI–Matrix $J_{\vec{h}}(\vec{y})$ erfordert. Wir greifen nochmals BSP. (14.4.3) auf und schreiben die Gleichung $\vec{F}(\vec{x}, \vec{y}) = \vec{0}$ in der Form (4.1), nämlich:

$$\vec{h}(\vec{y}) := B\vec{y} = \vec{c} - A\vec{x}.$$

Nun erkennt man, dass die Bedingung $\det J_{\vec{h}}(\vec{y}) \equiv \det B \neq 0$ notwendig für die Auflösbarkeit nach $\vec{y}$ ist. Sie ist aber auch hinreichend, wie wir in BSP. (14.4.3) erörtert haben. Natürlich kann die

Bedingung $\det J_{\vec{h}}(\vec{y}) \neq 0$ nur für **quadratische Jacobi**–Matrizen gefordert werden, und man könnte meinen, die Aufgabe (4.1) wäre nur im Fall $n = m$ sinnvoll gestellt. Tatsächlich darf man aber, wie das BSP. (14.4.3) lehrt, ganz allgemein von Funktionsvorgaben $\vec{F} \in \mathrm{Abb}\,(\mathbf{R}^n \times \mathbf{R}^m, \mathbf{R}^m)$ ausgehen. □

Eine Präzisierung der oben diskutierten Auflösbarkeitsbedingungen nehmen wir im folgenden Satz vor:

Satz 14.6 (Hauptsatz über implizite Funktionen)
Gegeben seien eine offene Teilmenge $\Omega \subset \mathbf{R}^n \times \mathbf{R}^m$ *und eine Vektorfunktion* $\vec{F} \in \mathrm{Abb}\,(\mathbf{R}^n \times \mathbf{R}^m, \mathbf{R}^m)$ *mit* $\vec{F} \in C^1(\Omega)$. *Es existiere ferner ein Punkt* $(\vec{x}_0, \vec{y}_0) \in \Omega$ *mit*

$$\boxed{\vec{F}(\vec{x}_0, \vec{y}_0) = \vec{0} \quad \textit{und} \quad \det\Big[\frac{\partial(F_1, \ldots, F_m)}{\partial(y_1, \ldots, y_m)}\Big]_{|(\vec{x}_0, \vec{y}_0)} \neq 0.} \tag{4.4}$$

Dann ist die Gleichung $\vec{F}(\vec{x}, \vec{y}) = \vec{0}$ **lokal eindeutig nach** $\vec{y}$ **auflösbar**. *Das heißt, es existiert eine offene* δ*–Kugel* $B_\delta(\vec{x}_0) \subset \mathbf{R}^n$ *und genau eine Funktion* $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ *mit den Eigenschaften*

$$\boxed{\vec{y}_0 = \vec{f}(\vec{x}_0) \quad \textit{und} \quad \vec{F}(\vec{x}, \vec{f}(\vec{x})) = \vec{0} \ \ \forall\, \vec{x} \in B_\delta(\vec{x}_0),} \tag{4.5}$$

$$\boxed{\vec{f} \in C^1(B_\delta(\vec{x}_0)),} \tag{4.6}$$

und die Funktion $\vec{y} = \vec{f}(\vec{x})$ *hat in jedem Punkt* $\vec{x} \in B_\delta(\vec{x}_0)$ *die Ableitung*

$$\boxed{\vec{f}\,'(\vec{x}) := J_{\vec{y}}(\vec{x}) = -\Big[\frac{\partial(F_1, \ldots, F_m)}{\partial(y_1, \ldots, y_m)}\Big]^{-1}{}_{|(\vec{x}, \vec{f}(\vec{x}))} \frac{\partial(F_1, \ldots, F_m)}{\partial(x_1, \ldots, x_n)}{}_{|(\vec{x}, \vec{f}(\vec{x}))}.}$$

Skizze für eine Begründung: Die Beweisidee beruht auf der Verwendung des Banachschen Fixpunktsatzes 13.2. Dazu werden wir einen vollständigen metrischen Raum M und einen kontrahierenden Operator $T : M \to M$ in geeigneter Weise einführen, und zwar wird M die abgeschlossene Kugel $M := \overline{B_{\beta/2}}(\vec{y}_0) \subset \mathbf{R}^m$ mit noch zu wählendem Radius $\beta/2$ sein. Versehen mit der euklidischen Metrik (D3) – vgl. BSP. (13.2.1) – ist dann M ein vollständiger metrischer Raum. Wir konstruieren nun den Operator T, der von dem Parameter $\vec{x}$ abhängig ist: $T = T_{\vec{x}}$. Dazu definieren wir die Matrix $B_0 \in \mathbf{R}^{(m,m)}$ gemäß

$$B_0 := \frac{\partial(F_1, \ldots, F_m)}{\partial(y_1, \ldots, y_m)}{}_{|(\vec{x}_0, \vec{y}_0)},$$

die wegen (4.4) invertierbar ist. Die Gleichung $\vec{F}(\vec{x}, \vec{y}) = \vec{0}$ kann deshalb in der Form $\vec{y} = \vec{y} - B_0^{-1}\vec{F}(\vec{x}, \vec{y})$ geschrieben werden. Das heißt, setzen wir

$$T_{\vec{x}}\vec{y} := \vec{y} - B_0^{-1}\vec{F}(\vec{x}, \vec{y}), \qquad (\vec{x}, \vec{y}) \in \Omega, \tag{4.7}$$

so haben wir zu $\vec{x}$ einen Fixpunkt $\vec{y} = \vec{y}(\vec{x})$ von $T_{\vec{x}}$ zu bestimmen. Zunächst folgern wir aus $Id = B_0^{-1}B_0$ die Relation

$$T_{\vec{x}}\vec{y}_1 - T_{\vec{x}}\vec{y}_2 = B_0^{-1}[B_0(\vec{y}_1 - \vec{y}_2) - (\vec{F}(\vec{x}, \vec{y}_1) - \vec{F}(\vec{x}, \vec{y}_2))], \qquad (\vec{x}, \vec{y}_1),\ (\vec{x}, \vec{y}_2) \in \Omega. \tag{4.8}$$

Nun sei $\epsilon > 0$ so gewählt, dass $\epsilon\,\|B_0^{-1}\|_F < \frac{1}{2}$ gilt. Hier bezeichnet $\|\cdot\|_F$ wieder die Frobenius–Norm einer Matrix, vgl. Abschnitt 6.5. Die Regularitätsvoraussetzung $\vec{F} \in C^1(\Omega)$ garantiert nun gemäß Definition 14.6 die Gültigkeit der Relation

$$\vec{F}(\vec{x}, \vec{y}) - \vec{F}(\vec{x}, \vec{y}_0) = B_0(\vec{y} - \vec{y}_0) + \vec{\mathcal{O}}(\|\vec{y} - \vec{y}_0\|) \quad \text{für} \quad \vec{y} \to \vec{y}_0.$$

Es existiert demnach eine offene β–Kugel $B_\beta(\vec{y}_0) \subset \mathbf{R}^m$ mit

$$\|\vec{F}(\vec{x}_0, \vec{y}) - \vec{F}(\vec{x}_0, \vec{y}_0) - B_0(\vec{y} - \vec{y}_0)\| < \epsilon \, \|\vec{y} - \vec{y}_0\| \quad \forall \, \vec{y} \in B_\beta(\vec{y}_0).$$

Hieraus lässt sich mit einem Stetigkeitsschluss, eventuell unter Verkleinerung von β, die folgende Ungleichung herleiten:

$$\|\vec{F}(\vec{x}_0, \vec{y}_1) - \vec{F}(\vec{x}_0, \vec{y}_2) - B_0(\vec{y}_1 - \vec{y}_2)\| < \epsilon \, \|\vec{y}_1 - \vec{y}_2\| \quad \forall \, \vec{y}_1, \, \vec{y}_2 \in B_\beta(\vec{y}_0).$$

Da $\vec{F}$ auch in der Variablen $\vec{x}$ stetig ist, kann sogar die Existenz einer offenen δ–Kugel $B_\delta(\vec{x}_0) \subset \mathbf{R}^n$ erschlossen werden mit

$$\|\vec{F}(\vec{x}, \vec{y}_1) - \vec{F}(\vec{x}, \vec{y}_2) - B_0(\vec{y}_1 - \vec{y}_2)\| < \epsilon \, \|\vec{y}_1 - \vec{y}_2\| \quad \forall \, \vec{y}_1, \, \vec{y}_2 \in B_\beta(\vec{y}_0) \quad \forall \vec{x} \in B_\delta(\vec{x}).$$

Somit resultiert unter Verwendung der Relation (4.8):

$$\|T_{\vec{x}}\vec{y}_1 - T_{\vec{x}}\vec{y}_2\| \leq \|B_0^{-1}\|_F \epsilon \, \|\vec{y}_1 - \vec{y}_2\| < \frac{1}{2} \, \|\vec{y}_1 - \vec{y}_2\| \quad \forall \, \vec{y}_1, \, \vec{y}_2 \in B_\beta(\vec{y}_0) \quad \forall \vec{x} \in B_\delta(\vec{x}). \tag{4.9}$$

Andererseits haben wir wegen der Voraussetzung (4.4)

$$T_{\vec{x}_0}\vec{y}_0 = \vec{y}_0 - B_0^{-1}\vec{F}(\vec{x}_0, \vec{y}_0) = \vec{y}_0 \in B_\beta(\vec{y}_0).$$

Somit dürfen wir annehmen, β sei so klein gewählt, dass gilt:

$$\|T_{\vec{x}}\vec{y} - \vec{y}_0\| < \frac{1}{2} \, \beta \quad \forall \, \vec{y} \in B_\beta(\vec{y}_0) \quad \forall \vec{x} \in B_\delta(\vec{x}_0). \tag{4.10}$$

Aus (4.9) und (4.10) folgt nun, dass der Operator $T_{\vec{x}}$ für jedes feste $\vec{x} \in B_\delta(\vec{x}_0)$ eine Kontraktion auf der Menge $M := \overline{B_{\beta/2}}(\vec{y}_0)$ ist. Also erschließen wir aus dem Banachschen Fixpunktsatz:

$$\forall \, \vec{x} \in B_\delta(\vec{x}_0) \; \exists! \, \vec{y} \in M \; : \; \vec{y} = T_{\vec{x}}\vec{y} := \vec{y} - B_0^{-1}\vec{F}(\vec{x}, \vec{y}).$$

Die eindeutige Zuordnung $\vec{x} \mapsto \vec{y} =: \vec{f}(\vec{x})$ induziert eine Abbildung $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^m)$ mit $\vec{F}(\vec{x}, \vec{f}(\vec{x})) = \vec{0} \;\; \forall \;\; \vec{x} \in B_\delta(\vec{x}_0)$. Insbesondere gilt $\vec{y}_0 = \vec{f}(\vec{x}_0)$ wegen $\vec{F}(\vec{x}_0, \vec{y}_0) = \vec{0}$. Stetigkeit und stetige Differenzierbarkeit von $\vec{f}$ resultieren aus den entsprechenden Eigenschaften von $\vec{F}$. Eine Skizze des Stetigkeitsbeweises geht so: Es seien $\vec{x}_1, \vec{x}_2 \in B_\delta(\vec{x}_0)$ gewählt und $\vec{y}_j := \vec{f}(\vec{x}_j)$ gesetzt. Dann gilt

$$\vec{f}(\vec{x}_1) - \vec{f}(\vec{x}_2) = T_{\vec{x}_1}\vec{y}_1 - T_{\vec{x}_2}\vec{y}_2 = T_{\vec{x}_1}\vec{y}_1 - T_{\vec{x}_1}\vec{y}_2 + T_{\vec{x}_1}\vec{y}_2 - T_{\vec{x}_2}\vec{y}_2.$$

Unter Verwendung der Ungleichung (4.9) ergibt sich daraus:

$$\|\vec{f}(\vec{x}_1) - \vec{f}(\vec{x}_2)\| \leq \frac{1}{2} \, \|\vec{f}(\vec{x}_1) - \vec{f}(\vec{x}_2)\| + \|B_0^{-1}\|_F \, \|\vec{F}(\vec{x}_1, \vec{y}_2) - \vec{F}(\vec{x}_2, \vec{y}_2)\|,$$

also schließlich

$$\|\vec{f}(\vec{x}_1) - \vec{f}(\vec{x}_2)\| \leq const \, \|\vec{F}(\vec{x}_1, \vec{y}_2) - \vec{F}(\vec{x}_2, \vec{y}_2)\|.$$

Da $\vec{F}(\vec{x}, \vec{y})$ in der Variablen $\vec{x}$ stetig ist, strebt die rechte Seite im Limes $\vec{x}_1 \to \vec{x}_2$ gegen Null. Die Ableitungsformel für $\vec{f}\,'$ folgt noch aus der Kettenregel, wenn diese auf die Gleichung $\vec{F}(\vec{x}, \vec{f}(\vec{x})) = \vec{0}$ angewendet wird:

$$\vec{0} = \frac{\partial(F_1, \ldots, F_m)}{\partial(x_1, \ldots, x_n)} + \frac{\partial(F_1, \ldots, F_m)}{\partial(y_1, \ldots, y_m)} \cdot \frac{\partial(y_1, \ldots, y_m)}{\partial(x_1, \ldots, x_n)}.$$

Gemäß Voraussetzung (4.4) gilt $\det \left[\frac{\partial(F_1, \ldots, F_m)}{\partial(y_1, \ldots, y_m)} \right] \neq 0$ in einer Umgebung des Punktes $(\vec{x}_0, \vec{y}_0)$, und deshalb besitzt die Jacobi–Matrix $\frac{\partial(F_1, \ldots, F_m)}{\partial(y_1, \ldots, y_m)}$ dort eine Inverse. Aus der obigen Gleichung folgt dann die behauptete Formel für $\vec{f}\,'(\vec{x})$. $\square$

Sonderfall der skalaren Gleichung $F(x,y)=0$

Im einfachsten Fall einer skalaren Funktion $F(x,y)$, also für $n=m=1$, sind die folgenden Bedingungen gemäß Satz 14.6 **hinreichend** dafür, dass durch die Gleichung $F(x,y)=0$ in der Nähe eines Punktes (x_0,y_0) **implizit** eine Funktion $y=f(x)$ mit $y_0=f(x_0)$ definiert wird (oder äquivalent, dass die Gleichung $F(x,y)=0$ lokal in der Nähe des Punktes (x_0,y_0) nach $y=f(x)$ aufgelöst werden kann):

- $F(x,y)$ ist stetig,
- $F(x,y)$ besitzt stetige partielle Ableitungen $F_x(x,y)$ und $F_y(x,y)$,
- $F(x_0,y_0)=0$ und $F_y(x_0,y_0)\neq 0$.

In diesem Fall existiert stets auch die stetige Ableitung der impliziten Funktion $y=f(x)$, und es gilt

$$\boxed{f'(x)=-\frac{F_x(x,y)}{F_y(x,y)},\qquad y:=f(x).} \tag{4.11}$$

Verzichtet man in den Voraussetzungen auf die Existenz der partiellen Ableitung $F_x(x,y)$, so ist immer noch die Existenz der stetigen impliziten Funktion $y=f(x)$ gewährleistet; diese braucht jedoch nicht mehr differenzierbar zu sein.

Wegen der Symmetrie in den Variablen x,y erhält man in gleicher Weise die lokale Auflösbarkeit der Gleichung $F(x,y)=0$ nach $x=h(y)$, wenn in den obigen Voraussetzungen die Bedingung $F_x(x_0,y_0)\neq 0$ an die Stelle von $F_y(x_0,y_0)\neq 0$ tritt.

BSP. (14.4.4) Wir betrachten hier für den Fall $a:=1$ nochmals die Funktion $F(x,y):=x^2+y^2-1$ aus BSP. (14.4.2). Auf der Menge $\Omega:=\mathbf{R}^2$ ist sicher $F\in C^2(\Omega)$ erfüllt. Nun diskutieren wir die lokale Auflösbarkeit der Gleichung $F(x,y)=0$ in der Nähe der vier Punkte $P(x_0,y_0):=P_1(1,0)$, $P_2(0,1)$, $P_3(-1,0)$ und $P_4(0,-1)$. In jedem dieser Punkte gilt $F(x_0,y_0)=0$ sowie

$$F_y(x_0,y_0)=2y_0=\begin{cases}0 & \text{in } P_1 \text{ und } P_3,\\ \pm 2 & \text{in } P_2 \text{ bzw. } P_4.\end{cases}$$

Das heißt, die Bedingungen (4.4) werden nur in den Punkten P_2 und P_4 erfüllt. Gemäß Satz 14.6 erhalten wir hier die bereits in BSP. (14.4.2) bestimmten lokalen Lösungen

$$y=f_\pm(x):=\pm\sqrt{1-x^2},\qquad -1<x<1,$$

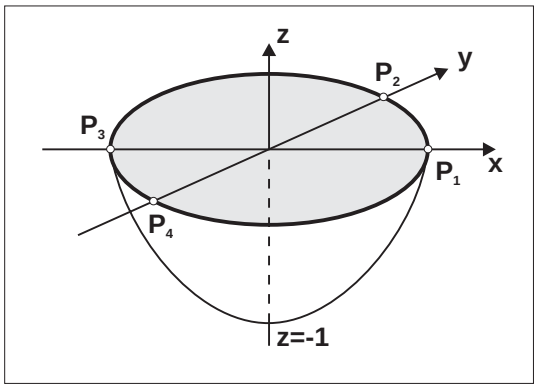
die zwar stetig fortsetzbar nach $x=\pm 1$ sind, dort aber nicht mehr differenzierbar sind. In den Punkten P_1 und P_3 wird jedoch wegen

$$F_x(x_0,y_0)=2x_0=\begin{cases}\pm 2 & \text{in } P_1 \text{ bzw. } P_3,\\ 0 & \text{in } P_2 \text{ und } P_4,\end{cases}$$

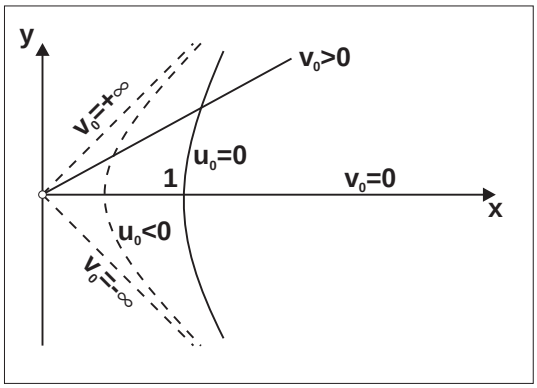
die lokale Auflösbarkeit der Gleichung $F(x,y)=0$ nach $x=h(y)$ gewährleistet:

$$x=h_\pm(y):=\pm\sqrt{1-y^2},\qquad -1<y<1.$$

Wir weisen hier nochmals auf die geometrische Bedeutung der erzielten Lösungen $y=f_\pm(x)$ und $x=h_\pm(y)$ hin. Sie sind lokale Darstellungen der Äquipotentiallinie der Fläche $z=F(x,y)$ zum Niveau $z=0$.



Die Niveaulinie $z = 0$ der Fläche $z = F(x, y) := x^2 + y^2 - 1$



Skizze zum BSP. (14.4.6)

Implizites Differenzieren

Der Satz 14.6 über implizite Funktionen enthält außer der abstrakten Existenzaussage keine Angaben darüber, wie die durch die Gleichung $\vec{F}(\vec{x}, \vec{y}) = \vec{0}$ implizit definierte Funktion $\vec{y} = \vec{f}(\vec{x})$ tatsächlich gefunden werden kann. Allgemein ist es auch gar nicht möglich, die Funktion $\vec{f}$ formelmäßig anzugeben. Hingegen kann man aber die Ableitungen von $\vec{f}$ im Punkt $\vec{x}_0$ berechnen, sofern der Funktionswert $\vec{y}_0 = \vec{f}(\vec{x}_0)$ als bekannt vorausgesetzt wird. Wir betrachten nun nur noch den skalaren Sonderfall $n = m = 1$. Es sei also auf einem Gebiet $\Omega \subset \mathbf{R}^2$ eine Funktion $F \in C^k(\Omega)$ und ein Punkt $(x_0, y_0) \in \Omega$ mit $F(x_0, y_0) = 0$ und $F_y(x_0, y_0) \neq 0$ vorgegeben. Dann existiert ja eine implizit definierte Funktion $y = f(x)$, und es gilt in einer geeigneten Kugelumgebung $U := B_\delta(x_0) \subset \mathbf{R}$ die Gleichung

$$0 = F(x, f(x)), \qquad x \in U.$$

Durch Differentiation dieser Gleichung nach x resultiert unter Verwendung der Kettenregel:

$$\begin{aligned} 0 &= F_x + F_y \cdot f'(x) \\ 0 &= F_{xx} + F_{xy} \cdot f'(x) + F_{yx} \cdot f'(x) + F_{yy} \cdot f'^2(x) + F_y \cdot f''(x), \end{aligned}$$

usw. Hier und im folgenden sind die Argumente $(x, f(x))$ bei den partiellen Ableitungen der Funktion F fortgelassen worden. Aus Stetigkeitsgründen muss $F_y(x, f(x)) \neq 0$ auch noch in einer geeigneten Umgebung von x_0 gelten, so dass die obigen Gleichungen jeweils nach $f'(x)$ bzw. nach $f''(x)$ aufgelöst werden können. Aus der ersten Gleichung erhalten wir wieder die Beziehung (4.11), also

$$\boxed{\begin{aligned} f'(x) &= -\frac{F_x}{F_y}\Big|_{(x,f(x))}, \\ f''(x) &= -\frac{1}{F_y^3}\left(F_x^2 F_{yy} - 2F_x F_y F_{xy} + F_y^2 F_{xx}\right)\Big|_{(x,f(x))}. \end{aligned}}$$

BSP. (14.4.5) Auf dem positiven Kegel $\Omega := \{(x, y) \in \mathbf{R}^2 : x > 0, \quad y > 0\}$ betrachten wir die Funktion $F(x, y) := x^y - y^x$. Im Punkt $(1, 1) \in \Omega$ sind die Ableitungen der durch die Gleichung $F(x, y) = 0$ implizit definierten Funktion $y = f(x)$ zu berechnen. Es ist klar, dass eine formelmäßige Darstellung von f nicht existiert.

Lösung: Es sind zunächst die Voraussetzungen des Satzes 14.6 über implizite Funktionen zu prüfen.

Sicher gilt $F \in C^\infty(\Omega)$, so dass alle Stetigkeitserfordernisse erfüllt sind. Wegen $F(x,y) = e^{y \ln x} - e^{x \ln y}$ erhält man auch $F(1,1) = 0$ und daraus

$$\begin{aligned} F_y(x,y) &= \ln x\, e^{y \ln x} - \frac{x}{y}\, e^{x \ln y}, \qquad F_y(1,1) = -1 \neq 0, \\ F_x(x,y) &= \frac{y}{x}\, e^{y \ln x} - \ln y\, e^{x \ln y}, \qquad F_x(1,1) = 1. \end{aligned}$$

Da die Voraussetzungen des Satzes über implizite Funktionen erfüllt sind, erhalten wir die Ableitung $f'(1)$ der impliziten Funktion f zu

$$f'(1) = -\frac{F_x(1,1)}{F_y(1,1)} = 1.$$

Die Berechnung der höheren Ableitungen ist bereits mit erheblichem Aufwand verbunden; wir verzichten hier auf explizite Formeln. Es resultiert zum Beispiel $f''(1) = 0$.

Wir kehren nun zurück zum Problem der Existenz einer inversen Funktion, das heißt, zum Problem der Lösbarkeit von Gleichung (4.1). Gegeben sei auf einer offenen Teilmenge $\Omega \subset \mathbf{R}^n$ eine Funktion $\vec{h} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^n)$ mit der Regularität $\vec{h} \in C^1(\Omega)$. Zu festem $\vec{y}_0 \in \Omega$ setzen wir $\vec{x}_0 := \vec{h}(\vec{y}_0)$ und betrachten die Funktion $\vec{F}(\vec{x}, \vec{y}) := \vec{h}(\vec{y}) - \vec{x}$. Sofern die Bedingung

$$\det \Big[\frac{\partial(F_1, \ldots, F_n)}{\partial(y_1, \ldots, y_n)}\Big]_{|(\vec{x}_0, \vec{y}_0)} = \left| \begin{array}{ccc} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial y_1} & \cdots & \frac{\partial h_n}{\partial y_n} \end{array} \right|_{|\vec{y}_0} = \det J_{\vec{h}}(\vec{y}_0) \neq 0$$

erfüllt ist, trifft Satz 14.6 zu: Es existiert lokal eine Funktion $\vec{y} = \vec{f}(\vec{x})$ mit $\vec{F}(\vec{x}, \vec{f}(\vec{x})) = \vec{0} = \vec{h}(\vec{f}(\vec{x})) - \vec{x}$ in einer Umgebung des Punktes $\vec{x}_0$. Das heißt, $\vec{f}$ ist eine Rechtsinverse der Funktion $\vec{h}$. Wir haben also mit modifizierten Bezeichnungen das folgende Ergebnis:

Satz 14.7 (Hauptsatz über inverse Funktionen)
Gegeben seien eine offene Teilmenge $\Omega \subset \mathbf{R}^n$ und eine Funktion $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^n)$ mit $\vec{f} \in C^1(\Omega)$. Gilt in einem Punkt $\vec{y}_0 \in \Omega$ die Bedingung

$$\boxed{\det J_{\vec{f}}(\vec{y}_0) \neq 0,} \tag{4.12}$$

so existiert eine offene δ–Kugel $U := B_\delta(\vec{x}_0)$ um den Punkt $\vec{x}_0 := \vec{f}(\vec{y}_0)$ und genau eine lokale Umkehrfunktion $\vec{f}^{-1} \in C^1(U)$ mit

$$\boxed{\vec{f}(\vec{f}^{-1}(\vec{x})) = \vec{x} \;\; \forall\, \vec{x} \in U.} \tag{4.13}$$

Es gelten ferner noch (vgl. (3.1) in Abschnitt 14.3):

$$\boxed{J_{\vec{f}^{-1}}(\vec{x}) = (J_{\vec{f}}(\vec{y}))^{-1}, \qquad \vec{x} = \vec{f}(\vec{y}) \in U.} \tag{4.14}$$

BSP. (14.4.6) Hier seien $\Omega := \mathbf{R}^2$ und $\vec{f} \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R}^2)$ gemäß

$$(x,y)^T := \vec{f}(u,v) := \begin{bmatrix} e^u \cosh v \\ e^u \sinh v \end{bmatrix}, \qquad (u,v) \in \Omega,$$

erklärt. Wegen

$$\det J_{\vec{f}}(u,v) = \begin{vmatrix} e^u \cosh v & e^u \sinh v \\ e^u \sinh v & e^u \cosh v \end{vmatrix} = e^{2u} \neq 0 \quad \forall\, (u,v) \in \Omega$$

existiert eine Inverse $\vec{f}^{-1}$ auf der Bildmenge $\vec{f}(\Omega) \subset \mathbf{R}^2$. Da $x = e^u \cosh v > 0$ und $\frac{y}{x} = \tanh v \in (-1,1)$ gelten, haben wir

$$\vec{f}(\Omega) = \{(x,y) \in \mathbf{R}^2 \,:\, x > 0,\ -x < y < x\}, \quad v = \operatorname{Ar}\tanh \frac{y}{x} = \frac{1}{2} \ln \frac{x+y}{x-y}, \quad u = \frac{1}{2} \ln(x^2 - y^2).$$

Somit kann $\vec{f}^{-1}$ in der folgenden Weise formelmäßig dargestellt werden:

$$(u,v)^T = \vec{f}^{-1}(x,y) = \begin{bmatrix} \frac{1}{2} \ln(x^2 - y^2) \\ \frac{1}{2} \ln \frac{x+y}{x-y} \end{bmatrix}, \quad (x,y) \in \vec{f}(\Omega).$$

Man berechnet außerdem die JACOBI–Matrix

$$J_{\vec{f}^{-1}}(u,v) = (J_{\vec{f}}(u,v))^{-1} = \begin{bmatrix} e^{-u} \cosh v & -e^{-u} \sinh v \\ -e^{-u} \sinh v & e^{-u} \cosh v \end{bmatrix}.$$

Wir können dieses Beispiel als Beispiel eines **Koordinatenwechsels** $\vec{x} = \vec{p}(u,v)$ in $\mathbf{R}^2$ auffassen. Hat man im allgemeinen Fall des $\mathbf{R}^n$ neue Koordinaten $u, v, \ldots$ durch einen Koordinatenwechsel $\vec{x} = \vec{p}(u,v,\ldots)$ eingeführt, so erhält man **Koordinatenlinien**, wenn man jeweils $n-1$ der Koordinaten konstant hält. Im vorliegenden Beispiel sind die v–Koordinatenlinien die Kurven $u = u_0 = const.$ Diese sind rechtwinklige Hyperbeln mit dem Scheitelabstand e^{u_0}:

$$x^2 - y^2 = e^{2u_0}.$$

Die u–Koordinatenlinien sind die Kurven $v = v_0 = const$, also die Geraden $y = x \tanh v_0$. Der Vektor $\frac{\partial \vec{p}}{\partial u}$ liegt tangential zu den u–Koordinatenlinien. Die Forderung $\det J_{\vec{p}}(u,v,\ldots) = \det\left(\frac{\partial \vec{p}}{\partial u}, \frac{\partial \vec{p}}{\partial v}, \ldots\right) \neq 0$ stellt somit sicher, dass Koordinatenlinien sich stets unter einem Winkel $\neq 0$ schneiden.

14.5 Singuläre Kurvenpunkte ebener Kurven

Definition 14.7 *Es sei ein reelles Polynom vom Grade $m \in \mathbf{N}_0$ in den zwei Variablen $(x,y) \in \mathbf{R}^2$ gegeben,*

$$F(x,y) := \sum_{j=0}^{m} a_j x^j y^{m-j}, \qquad a_j \in \mathbf{R}.$$

Dann heißen die Äquipotentiallinien $F(x,y) = 0$ **ebene algebraische Kurven**.

BSP. (14.5.1) Die Äquipotentiallinien $F(x,y) = 0$ der allgemeinen quadratischen Funktion

$$F(x,y) := Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + G, \qquad A,B,C,D,E,G \in \mathbf{R},$$

sind gerade die **Kegelschnitte**, siehe Abschnitt 11.6. Diese sind also algebraische Kurven.

Wir haben im vorangegangenen Abschnitt 14.4 gezeigt, dass die Steigung der Tangente an die Äquipotentiallinie $F(x,y) = 0$ durch den Punkt (x_0,y_0) unter der Voraussetzung $\operatorname{grad} F(x_0,y_0) \neq \vec{0}$ auch ohne Kenntnis einer **expliziten Darstellung** $y = f(x)$ berechnet werden kann. Sofern wir allgemein eine stetig differenzierbare Funktion $F(x,y)$ betrachten, gilt ja unter der Voraussetzung

$\operatorname{grad} F(x_0, y_0) \neq \vec{0}$ nach der Regel des impliziten Differenzierens für die Ableitung $y'(x_0)$ der implizit definierten Funktion $y = f(x)$ – sofern diese existiert:

$$F_x(x_0, y_0) + F_y(x_0, y_0) \cdot y'(x_0) = 0 \quad \Rightarrow \quad y'(x_0) = \begin{cases} -\dfrac{F_x(x_0, y_0)}{F_y(x_0, y_0)} & : \ F_y(x_0, y_0) \neq 0, \\ \pm\infty & : \ \text{sonst.} \end{cases} \tag{5.1}$$

Gilt hingegen $\operatorname{grad} F(x_0, y_0) = \vec{0}$, so bleibt die durch (5.1) definierte Tangentensteigung $y'(x_0)$ unbestimmt. Für algebraische Kurven erklärt man in diesem Fall:

Definition 14.8 *Es sei (x_0, y_0) ein Punkt auf der durch die Gleichung $F(x, y) = 0$ definierten algebraischen Kurve, deren Existenz vorausgesetzt sei. Gilt $\operatorname{grad} F(x_0, y_0) = \vec{0}$, so heiße (x_0, y_0) ein* **singulärer Punkt**.

BSP. (14.5.2) Die **Ellipsen** mit Halbachsen $a, b > 0$ sind die durch die Gleichung

$$F(x, y) := \left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 - 1 \stackrel{!}{=} 0$$

definierten ebenen algebraischen Kurven. Es gilt hier $\operatorname{grad} F(x_0, y_0) = (\frac{2x_0}{a^2}, \frac{2y_0}{b^2})^T = \vec{0}$ genau für $(x_0, y_0) = (0, 0)$. Wegen $F(0, 0) \neq 0$ liegt der Punkt $(0, 0)$ nicht auf der durch $F(x, y) = 0$ definierten Kurve. Das heißt, Ellipsen besitzen **keine** singulären Punkte.

Das Verhalten von algebraischen Kurven in der Umgebung eines singulären Punkten (x_0, y_0) kann im Regelfall durch eine asymptotische Analyse charakterisiert werden. Es ist üblich, einen **Typenkatalog** von singulären Punkten zu erstellen. Wir erörtern hier einige der Standardtypen.

Definition 14.9 *Schneiden sich verschiedene Zweige der durch $F(x, y) = 0$ definierten ebenen algebraischen Kurve in einem singulären Punkt (x_0, y_0) unter nichtverschwindendem Winkel, wobei jeder Zweig in (x_0, y_0) eine eigene Tangente besitze, so heiße (x_0, y_0) ein* **Doppelpunkt** *oder* **Knotenpunkt**.

BSP. (14.5.3) Das DESCARTES**sche Blatt** (*Folium Cartesii*) ist die durch die Gleichung

$$F(x, y) := x^3 + y^3 - 3axy = 0, \qquad a > 0,$$

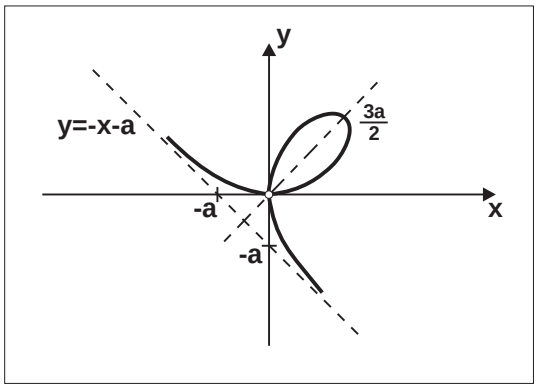
definierte ebene algebraische Kurve. Wir haben hier $\operatorname{grad} F(x_0, y_0) = (3x_0^2 - 3ay_0, 3y_0^2 - 3ax_0)^T = \vec{0}$ genau für $(x_0, y_0) = (0, 0)$ und $(x_0, y_0) = (a, a)$. Wegen $F(0, 0) = 0$ und $F(a, a) \neq 0$ ist nur der Punkt $(0, 0)$ ein **singulärer Kurvenpunkt**. Nahe $(0, 0)$ gilt die asymptotische Entwicklung

$$F(x, y) = -3axy + \mathcal{O}(|x^3 + y^3|) \quad \text{für} \quad |x| + |y| \to 0,$$

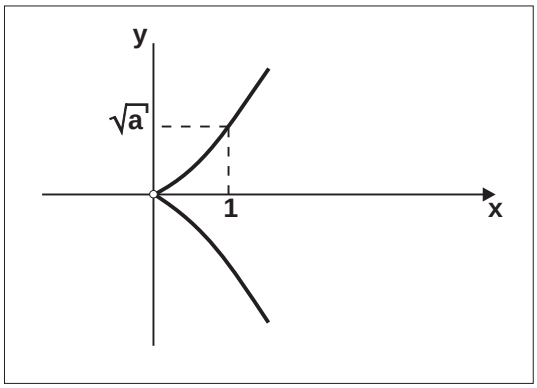
so dass das Verhalten der ÄP $F(x, y) = 0$ nahe $(0, 0)$ asymptotisch durch die zwei Geraden $x = 0$ und $y = 0$ wiedergegeben wird. Da sich diese in $(0, 0)$ unter rechtem Winkel schneiden, liegt bei $(0, 0)$ ein **Doppelpunkt**. Mit

$$x(t) := \frac{3at}{1 + t^3}, \quad y(t) = \frac{3at^2}{1 + t^3}, \qquad t \neq -1,$$

liegt eine **Parameterdarstellung** des DESCARTESschen Blattes vor. Aus den Beziehungen $\lim\limits_{t \to -1} \frac{x(t)}{y(t)} = -1$ und $\lim\limits_{t \to -1} (x(t) + y(t)) = -a$ erkennt man, dass die Gerade $y = -x - a$ eine Asymptote des DESCARTESschen Blattes ist.



Das DESCARTESsche Blatt besitzt in $(0,0)$ einen Doppelpunkt



Die NEILsche Parabel besitzt in $(0,0)$ einen Rückkehrpunkt 1.Art

Definition 14.10 *Enden zwei verschiedene Zweige in einem singulären Punkt (x_0, y_0) der durch $F(x,y) = 0$ definierten ebenen algebraischen Kurve, so heiße (x_0, y_0)*

- *eine* **Spitze**, *falls in (x_0, y_0) eine gemeinsame Tangente beider Zweige vorliegt,*
- *ein* **Knickpunkt**, *falls jeder Zweig in (x_0, y_0) eine eigene Tangente besitzt,*
- *ein* **Rückkehrpunkt 1.Art**, *falls (x_0, y_0) eine Spitze ist und die beiden Zweige auf verschiedenen Seiten der gemeinsamen Tangente liegen,*
- *ein* **Rückkehrpunkt 2.Art** *oder eine* **Schnabelspitze**, *falls (x_0, y_0) eine Spitze ist und die beiden Zweige auf einer Seite der gemeinsamen Tangente liegen.*

BSP. (14.5.4) Die NEILsche **Parabel** (oder *semikubische Parabel*) ist die durch die Gleichung

$$F(x,y) := y^2 - 3ax^3 = 0, \qquad a > 0,$$

definierte ebene algebraische Kurve. Wir haben nun $\operatorname{grad} F(x_0, y_0) = (-3ax_0^2, 2y_0)^T = \vec{0}$ genau für $(x_0, y_0) = (0,0)$. Wegen $F(0,0) = 0$ ist also der Punkt $(0,0)$ ein **singulärer Kurvenpunkt**. Die ÄP $F(x,y) = 0$ durch den singulären Punkt $(0,0)$ zerfällt in zwei Kurvenzweige mit den expliziten Darstellungen

$$y_\pm(x) := \pm\sqrt{ax^3}, \qquad y'_\pm(x) = \pm\frac{3}{2}\sqrt{ax}, \qquad x \geq 0.$$

Wegen $y'_\pm(0+) = 0$ liegt in $(0,0)$ eine gemeinsame Tangente beider Zweige vor; das heißt, der Punkt $(0,0)$ ist ein **Rückkehrpunkt 1.Art**.

BSP. (14.5.5) Wir betrachten jetzt die durch die Gleichung

$$F(x,y) := (y - x^2)^2 - x^5 = 0$$

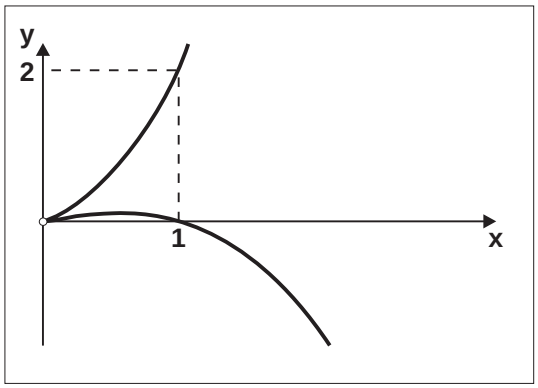
definierte ebene algebraische Kurve. Es gilt hier $\operatorname{grad} F(x_0, y_0) = (-4x_0(y_0 - x_0^2) - 5x_0^4, 2(y_0 - x_0^2)) = \vec{0}$ genau für $(x_0, y_0) = (0,0)$, und wegen $F(0,0) = 0$ ist der Punkt $(0,0)$ wiederum ein **singulärer Kurvenpunkt**. Nahe $(0,0)$ gilt die asymptotische Entwicklung

$$F(x,y) = (y - x^2)^2 + \mathcal{O}(|x^5|) \quad \text{für} \quad |x| + |y| \to 0,$$

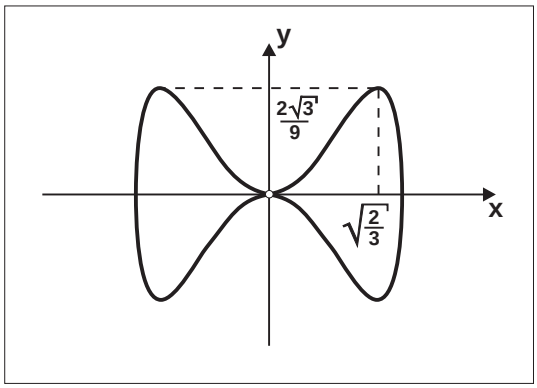
so dass das Verhalten der ÄP $F(x,y) = 0$ nahe $(0,0)$ asymptotisch durch die Parabel $y = x^2$ wiedergegeben wird. Man identifiziert mit dieser Asymptotik offenbar nur einen Kurvenzweig. Erst durch explizites Auflösen der Gleichung $F(x,y) = 0$ ergeben sich beide Zweige mit der Darstellung

$$y_\pm(x) = x^2 \pm \sqrt{x^5}, \qquad y'_\pm(x) = 2x \pm \frac{5}{2}\sqrt{x^3}, \qquad x \geq 0.$$

Wegen $y'_\pm(0+) = 0$ muss also der Punkt $(0,0)$ ein **Rückkehrpunkt 2.Art** (eine Schnabelspitze) sein.



Die algebraische Kurve $(y-x^2)^2-x^5=0$
hat eine Schnabelspitze in $(0,0)$



Die algebraische Kurve $y^2-x^4+x^6=0$
hat einen Berührungspunkt in $(0,0)$

Definition 14.11 *Berühren sich zwei verschiedene Zweige einer durch* $F(x,y)=0$ *definierten ebenen algebraischen Kurve in einem singulären Punkt* (x_0,y_0)*, so heiße dieser Punkt ein* (*Selbst–*) **Berührungspunkt**.

BSP. (14.5.6) Wir betrachten hier die durch die Gleichung

$$F(x,y) := y^2 - x^4 + x^6 = 0$$

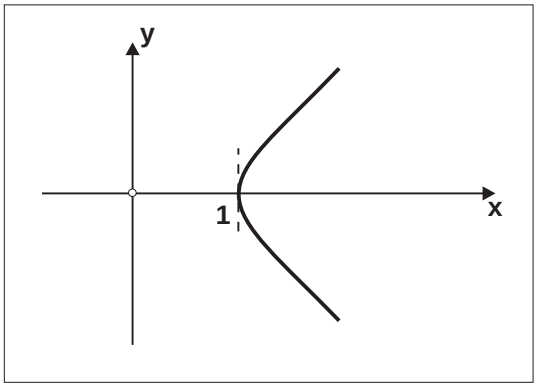
definierte ebene algebraische Kurve. Es gilt nun $\operatorname{grad} F(x_0,y_0) = (-4x_0^3+6x_0^5, 2y_0) = \vec{0}$ genau für $(x_0,y_0)=(0,0)$ und $(x_0,y_0)=(\pm\sqrt{\frac{2}{3}},0)$. Wegen $F(0,0)=0$ und $F(\pm\sqrt{\frac{2}{3}},0)\neq 0$ ist der Punkt $(0,0)$ der einzige **singuläre Kurvenpunkt**. Nahe $(0,0)$ gilt die asymptotische Entwicklung

$$F(x,y) = y^2 - x^4 + \mathcal{O}(|x^6|) \quad \text{für} \quad |x|+|y| \to 0,$$

so dass das Verhalten der ÄP $F(x,y)=0$ nahe $(0,0)$ asymptotisch durch die zwei Parabelzweige $\tilde{y}_\pm(x) = \pm x^2$ wiedergegeben wird. Da sich diese Zweige in $(0,0)$ mit gemeinsamer horizontaler Tangente berühren, muss der Punkt $(0,0)$ ein **Selbstberührungspunkt** der ebenen algebraischen Kurve sein. Man erhält übrigens auch ohne Asymptotik die zwei Kurvenäste in expliziter Darstellung:

$$y_\pm(x) = \pm x^2\sqrt{1-x^2}, \;\; |x|\le 1, \qquad y'_\pm(x) = \pm\frac{x(2-3x^2)}{\sqrt{1-x^2}} = \begin{cases} 0 & : \; x=0, \\ \mp\infty & : \; x\to 1, \\ \pm\infty & : \; x\to -1. \end{cases}$$

Definition 14.12 *Ein singulärer Punkt* (x_0,y_0)*, in dessen Umgebung keine weiteren Punkte der durch die Gleichung* $F(x,y)=0$ *definierten ebenen algebraischen Kurve liegen, heiße ein* **isolierter Punkt** *oder* **Einsiedlerpunkt**.



Die algebraische Kurve $y^2-x^2(x-1)=0$
hat einen Einsiedlerpunkt in $(0,0)$

BSP. (14.5.7) Wir betrachten schließlich die durch die Gleichung

$$F(x,y) := y^2 - x^2(x-1) = 0$$

definierte ebene algebraische Kurve. Es gilt $\operatorname{grad} F(x_0,y_0) = (-2x_0 + 3x_0^2, 2y_0) = \vec{0}$ genau für $(x_0,y_0) = (0,0)$ und $(x_0,y_0) = (\frac{2}{3},0)$. Wegen $F(0,0) = 0$ und $F(\frac{2}{3},0) \neq 0$ ist der Punkt $(0,0)$ der einzige **singuläre Kurvenpunkt**. Durch explizites Auflösen der Gleichung $F(x,y) = 0$ ergeben sich zwei Kurvenzweige mit der Darstellung

$$y_\pm(x) = \pm x\sqrt{x-1}, \quad x = 0 \quad \text{oder} \quad x \geq 1, \qquad y'_\pm(x) = \pm\frac{3x-2}{2\sqrt{x-1}}, \quad x \geq 1.$$

Der singuläre Punkt $(0,0)$ ist ein **Einsiedlerpunkt**.

Kapitel 15

Lineare Optimierung

15.1 Problemstellung, Normalform, Beispiele

In Abschnitt 13.9 haben wir bereits **Extremwertaufgaben mit Nebenbedingungen** studiert, dabei allerdings nur solche Nebenbedingungen zugelassen, die als **Gleichungsrestriktionen** formuliert werden konnten. In diesem Kapitel werden wir nochmals Extremwertaufgaben studieren, allerdings unter weitaus allgemeineren Nebenbedingungen, die auch die Form von **Ungleichungsrestriktionen** haben dürfen. Wir sprechen dann von **Optimierungsaufgaben**. Wir werden uns allerdings ausschließlich mit **linearen** Aufgaben befassen.

Die *allgemeine Aufgabenstellung* der Optimierung kann man wie folgt formulieren:

Gegeben seien Funktionen

$$f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}), \quad \vec{g} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^s), \quad \vec{h} \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R}^t).$$

Gesucht ist ein Vektor $\vec{x} \in \mathbf{R}^n$ mit

$$f(\vec{x}) = \min_{\vec{u} \in \mathbf{R}^n} f(\vec{u}), \quad \vec{g}(\vec{x}) = \vec{0}, \quad \vec{h}(\vec{x}) \geq \vec{0}. \tag{OP}$$

Definition 15.1 *Die Funktion $f(\vec{x})$ heiße* **Zielfunktion** *oder* **Kostenfunktion**; *die Bedingungen $\vec{g}(\vec{x}) = \vec{0}$ heißen* **Gleichungs–Nebenbedingungen**, *und die Bedingungen $\vec{h}(\vec{x}) \geq \vec{0}$ heißen* **Ungleichungs–Nebenbedingungen**. *Nebenbedingungen werden auch* **Restriktionen** *genannt. Dabei erfüllen zwei Vektoren $\vec{u} = (u_1, u_2, \ldots, u_m)^T$, $\vec{v} = (v_1, v_2, \ldots, v_m)^T \in \mathbf{R}^m$ die Relation $\vec{u} \leq \vec{v}$ (bzw. $\vec{u} \geq \vec{v}$), wenn* **komponentenweise** *$u_j \leq v_j$ (bzw. $u_j \geq v_j$) $\forall\, j = 1, 2, \ldots, m$ gilt.*

Bemerkung 15.1 (a) Wird die Zielfunktion $f(\vec{x})$ mit -1 multipliziert, so erhält man anstelle der Minimierungsaufgabe eine **Maximierungsaufgabe**. Letztere braucht somit nicht gesondert betrachtet zu werden.

(b) Auch die Ungleichungsrestriktionen $\vec{h}(\vec{x}) \geq \vec{0}$ werden durch Multiplikation mit -1 in die Restriktionen $-\vec{h}(\vec{x}) =: \vec{h}^*(\vec{x}) \leq \vec{0}$ konvertiert. Somit sind durch den Standardtyp (OP) weitere Optimierungsaufgaben abgedeckt. □

Sind die Funktionen $f, \vec{g}$ und $\vec{h}$ **linear**, so liegt mit (OP) ein **lineares Optimierungsproblem** (LOP) vor, auch **lineares Programm** genannt. Dieses gestattet die folgende Formulierung:

Gegeben seien ein **Kostenvektor** $\vec{c} \in \mathbf{R}^n$, zwei **Grenzvektoren** $\vec{\ell}, \vec{u} \in \overline{\mathbf{R}}^{n+m}$ sowie die **Matrix der Nebenbedingungen** $A \in \mathbf{R}^{(m,n)}$.

Gesucht ist ein Vektor $\vec{x} \in \mathbf{R}^n$ mit

$$\vec{\ell} \le \begin{bmatrix} \vec{x} \\ A\vec{x} \end{bmatrix} \le \vec{u} \quad \text{und} \quad z := \langle \vec{c}, \vec{x} \rangle \stackrel{!}{=} \text{Min.} \tag{LOP}$$

In dieser Terminologie sind **Gleichungs**–Restriktionen durch die Vorgabe $\ell_j = u_j$ charakterisiert. Wir fordern, dass **Ungleichungs**–Restriktionen jeweils nur auf **einer Ungleichungsseite** von (LOP) auftreten und lassen deshalb die Werte $\ell_j = -\infty$ sowie $u_j = +\infty$ zu. Das heißt, alle Restriktionen durch $\pm\infty$ sind **Leer–Restriktionen**.

Definition 15.2 (a) *Treten in* (LOP) *Ungleichungs–Restriktionen in der Form* $x_j \ge 0$ *für gewisse Variable* x_j *auf, so heißen diese Variablen* **vorzeichenbeschränkt**. *Variable mit der trivialen Ungleichungs–Restriktion* $-\infty < x_j < +\infty$ *heißen* **unrestringiert** *oder* **freie Variable**.

(b) *Die* **Normalform** *des* (LOP) *besteht in der Aufgabe*

$$\boxed{\textit{Minimiere } \langle \vec{c}, \vec{x} \rangle \textit{ auf der Menge } M := \{\vec{x} \in \mathbf{R}^n \,:\, \vec{x} \ge \vec{0},\ A\vec{x} = \vec{b}\}.} \tag{P}$$

Die Menge M *heiße* **Menge der zulässigen Lösungen**. *Das Problem* (P) *heiße* **zulässig**, *wenn* $M \neq \emptyset$ *gilt. Eine zulässige Lösung* $\vec{x}^* \in M$ *heiße eine* (*globale*) **Lösung** *von* (P), *wenn gilt:*

$$\boxed{\langle \vec{c}, \vec{x}^* \rangle \le \langle \vec{c}, \vec{x} \rangle \ \forall\, \vec{x} \in M.}$$

Es ist klar, dass durch eine Transformation $\vec{x}\,' := B\vec{x} - \vec{v}$, $B \in \text{Inv}\,(\mathbf{R}^n)$, $\vec{v} \in \mathbf{R}^n$ fest, lediglich die **Form** der Aufgabe (LOP) verändert wird, nicht aber ihr **Lösungscharakter**. Ist $\vec{x} \in \mathbf{R}^n$ eine Lösung von (LOP) und setzt man $A' := AB^{-1}$ sowie $\vec{c}\,' := (B^{-1})^T\vec{c}$, so löst $\vec{x}\,' \in \mathbf{R}^n$ die Aufgabe

$$\vec{\ell}_B \le \begin{bmatrix} B^{-1}\vec{x}\,' \\ A'\vec{x}\,' \end{bmatrix} \le \vec{u}_B \quad \text{und} \quad z' := \langle \vec{c}\,', \vec{x}\,' \rangle \stackrel{!}{=} \text{Min.} \tag{1.1}$$

Hier haben wir $\vec{\ell}_B := \vec{\ell} - \vec{v}_B, \vec{u}_B := \vec{u} - \vec{v}_B$ mit $\vec{v}_B := \begin{bmatrix} B^{-1}\vec{v} \\ A'\vec{v} \end{bmatrix}$ gesetzt. Umgekehrt führt eine Lösung $\vec{x}\,' \in \mathbf{R}^n$ von (1.1) auf eine Lösung $\vec{x} = B^{-1}\vec{x}\,' + B^{-1}\vec{v}$ von (LOP). Aus dieser Erkenntnis folgern wir:

Satz 15.1 *Die Aufgabe* (LOP) *lässt sich ohne Veränderung ihres Lösungscharakters – eventuell unter Vergrößerung der Dimension* n *– so standardisieren, dass alle Variablen* x_j **vorzeichenbeschränkt** *sind:*

$$\vec{\ell}\,' \le A'\vec{x}\,' \le \vec{u}\,', \quad \vec{x}\,' \ge \vec{0} \ \text{ und } \ z' := \langle \vec{c}\,', \vec{x}\,' \rangle \stackrel{!}{=} \text{Min.} \tag{LOP}'$$

Begründungen: (a) Wir gehen davon aus, dass in (LOP) **keine** Gleichungs–Restriktionen in der Form $\ell_j = x_j = u_j$ vorliegen. In diesem Fall ist die Variable x_j nämlich bereits eindeutig bestimmt, und die Dimension von (LOP) kann um 1 vermindert werden.

(b) Alle *wesentlichen* Ungleichungs–Restriktionen in der Form $x_j \le u_j < +\infty$ in (LOP) können durch Multiplikation mit -1 in die Normalform

$$-\infty < \ell_j \le -x_j, \quad \ell_j := -u_j,$$

gebracht werden. Setzt man in der Matrix $B := \mathrm{diag}\,(\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ die Koeffizienten $\epsilon_j := -1$ an diejenigen Stellen, an denen die Multiplikation mit -1 erfolgt ist, sonst aber $\epsilon_j := +1$, so ist $B = B^T = B^{-1} \in \mathrm{Inv}\,(\mathbf{R}^n)$, und man erhält aus (LOP) die neue Restriktionsbedingung

$$\vec{p} \le B\vec{x}, \quad \vec{\ell}\,' \le A\vec{x} \le \vec{u}\,', \quad \vec{p} \in \overline{\mathbf{R}}^n \;\text{ mit }\; p_j := \begin{cases} -u_j & , \ \epsilon_j = -1, \\ \ell_j & , \ \epsilon_j = +1. \end{cases}$$

Hierin sind $\vec{\ell}\,', \vec{u}\,' \in \overline{\mathbf{R}}^m$ die Projektionen von $\vec{\ell}$ bzw. $\vec{u}$ auf die letzten m Komponenten. Wir definieren jetzt den Vektor $\vec{v}$ gemäß $v_j := p_j$, falls $p_j > -\infty$, und $v_j := 0$, falls $p_j = -\infty$. Für $\vec{x}\,' := B\vec{x} - \vec{v}$ gilt nun $x'_j \ge 0$ für alle restringierten Variablen sowie $-\infty < x'_j = x_j < +\infty$ für die unrestringierten Variablen. Wir haben ferner mit $A' := AB^{-1} = AB$ und $\vec{c}\,' := (B^{-1})^T\vec{c} = B\vec{c}$:

$$\vec{\ell}\,' - A'\vec{v} \le A'\vec{x}\,' \le \vec{u}\,' - A'\vec{v}, \quad z' = \langle \vec{c}\,', \vec{x}\,' \rangle \stackrel{!}{=} \mathrm{Min}.$$

(c) Ist x_j eine *unrestringierte* Variable, so ersetzen wir x_j durch ein Paar (x_j^+, x_j^-) vorzeichenbeschränkter Variabler

$$x_j = x_j^+ - x_j^-, \qquad x_j^+ := \max\{0, x_j\} \ge 0, \quad x_j^- := \max\{0, -x_j\} \ge 0. \tag{1.2}$$

Die Dimension n steigt auf $n+1$ an, und es sind folgende Ersetzungen vorzunehmen:

$$\begin{aligned}
\vec{x} &= (\ldots, x_{j-1}, x_j, x_{j+1}, \ldots)^T \in \mathbf{R}^n &&\Rightarrow\ \vec{x}\,' = (\ldots, x_{j-1}, x_j^+, x_j^-, x_{j+1}, \ldots)^T \in \mathbf{R}^{n+1}, \\
\vec{c} &= (\ldots, c_{j-1}, c_j, c_{j+1}, \ldots)^T \in \mathbf{R}^n &&\Rightarrow\ \vec{c}\,' = (\ldots, c_{j-1}, c_j, -c_j, c_{j+1}, \ldots)^T \in \mathbf{R}^{n+1}, \\
A &= (\ldots, \vec{a}_{j-1}, \vec{a}_j, \vec{a}_{j+1}, \ldots) \in \mathbf{R}^{(m,n)} &&\Rightarrow\ A' = (\ldots, \vec{a}_{j-1}, \vec{a}_j, -\vec{a}_j, \vec{a}_{j+1}, \ldots) \in \mathbf{R}^{(m,n+1)}.
\end{aligned}$$

Analog verfährt man mit den anderen unrestringierten Variablen. Man beachte, dass durch diese Schritte der **Rang** der Matrix A **nicht** verändert wird: $\mathrm{Rang}\,A = \mathrm{Rang}\,A'$. Die Variablen x_j^+ und x_j^- können aus der Aufgabe (LOP)' nicht eindeutig bestimmt werden. Ist ein Lösungspaar (x_j^+, x_j^-) bekannt, so macht man sich schnell klar, dass auch $(x_j^+ + c, x_j^- + c)$ für jedes $c \ge 0$ ein Lösungspaar ist. Man erhält aber stets dasselbe $x_j = x_j^+ - x_j^-$, und mit (1.2) *a posteriori* ein eindeutig bestimmtes Paar (x_j^+, x_j^-). □

Wir beseitigen schließlich noch in (LOP)' die Ungleichungsrestriktion $\vec{\ell}\,' \le A'\vec{x}\,' \le \vec{u}\,'$ durch Einführung von sogenannten **Schlupfvariablen**.

Satz 15.2 *Die Aufgabe* (LOP)' *lässt sich ohne Veränderung ihres Lösungscharakters – eventuell unter Vergrößerung der Dimension n' – auf die Normalform* (P) *transformieren.*

Begründung: (a) Gilt in (LOP)' für einen Index $j \in \{1, 2, \ldots, m\}$ die Gleichungs–Restriktion $\ell'_j = (A\vec{x}\,')_j = u'_j =: b_j$, so ist nichts zu unternehmen.
(b) Gilt in (LOP)' für einen Index $j \in \{1, 2, \ldots, m\}$ die Ungleichungs–Restriktion

$$(A'\vec{x}\,')_j = \sum_{k=1}^{n'} a'_{jk}x'_k \ge \ell'_j =: b_j,$$

so kann diese Bedingung äquivalent umformuliert werden in

$$\sum_{k=1}^{n'} a'_{jk}x'_k - y_i = b_j, \quad y_i \ge 0,$$

womit eine neue vorzeichenbeschränkte Variable y_i eingeführt wird.

(c) Gilt in (LOP)' für einen Index $j \in \{1, 2, \ldots, m\}$ die Ungleichungs–Restriktion

$$(A'\vec{x}\,')_j = \sum_{k=1}^{n'} a'_{jk} x'_k \leq u'_j =: -b_j,$$

so setzen wir $a''_{jk} := -a'_{jk}$, $k = 1, 2, \ldots, n'$, und schreiben unter Einführung einer Schlupfvariablen y_i äquivalent:

$$\sum_{k=1}^{n'} a''_{jk} x'_k - y_i = b_j, \quad y_i \geq 0.$$

Auf diese Weise müssen maximal m Schlupfvariable y_i hinzugefügt werden. Die Aufgabe (LOP)' hat am Ende die Normalform

$$(A'', -Id) \begin{bmatrix} \vec{x}\,' \\ \vec{y} \end{bmatrix} = \vec{b}, \quad \begin{bmatrix} \vec{x}\,' \\ \vec{y} \end{bmatrix} \geq \begin{bmatrix} \vec{0} \\ \vec{0} \end{bmatrix}, \quad z' = \left\langle \begin{bmatrix} \vec{c}\,' \\ \vec{0} \end{bmatrix}, \begin{bmatrix} \vec{x}\,' \\ \vec{y} \end{bmatrix} \right\rangle \stackrel{!}{=} \text{Min}.$$

Wir fassen nachfolgend die erforderlichen Schritte zur Herstellung der Normalform (P) eines linearen Programmes (LOP) zusammen. In den Nebenbedingungen

$$\vec{\ell}^{\,(1)} \leq \vec{x} \leq \vec{u}^{\,(1)}, \qquad \vec{\ell}^{\,(2)} \leq A\vec{x} \leq \vec{u}^{\,(2)}$$

gehen wir von folgenden Komponentendarstellungen aus:

$$\begin{array}{lll} \vec{\ell}^{\,(1)} = (\ell_1^{(1)}, \ell_2^{(1)}, \ldots, \ell_n^{(1)})^T, & \vec{x} = (x_1, x_2, \ldots, x_n)^T, & \vec{u}^{\,(1)} = (u_1^{(1)}, u_2^{(1)}, \ldots, u_n^{(1)})^T, \\ \vec{\ell}^{\,(2)} = (\ell_1^{(2)}, \ell_2^{(2)}, \ldots, \ell_m^{(2)})^T, & A = (\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n), & \vec{u}^{\,(2)} = (u_1^{(2)}, u_2^{(2)}, \ldots, u_m^{(2)})^T, \\ & \vec{c} = (c_1, c_2, \ldots, c_n)^T. & \end{array}$$

Die Matrix A werden wir auch durch ihre Zeilenvektoren darstellen:

$$A = \begin{bmatrix} \vec{z}_1 \\ \vec{z}_2 \\ \vdots \\ \vec{z}_m \end{bmatrix}.$$

Nun gelte für einen Index j:

$\boxed{\ell_j^{(1)} = x_j = u_j^{(1)} :}$ Dann ist die Komponente x_j bereits eindeutig bestimmt. Man streiche x_j sowie in A den Spaltenvektor $\vec{a}_j$ und in $\vec{c}$ die Komponente c_j. Man ersetze schließlich n durch $n - 1$.

$\boxed{x_j \leq u_j^{(1)} < +\infty :}$ Der Spaltenvektor $\vec{a}_j$ ist durch $\vec{a}_j\,' := -\vec{a}_j$ zu ersetzen, und es sind die folgenden Substitutionen vorzunehmen:

$$\vec{\ell}\,' := \vec{\ell}^{\,(2)} + u_j^{(1)} \vec{a}_j\,', \quad \vec{u}\,' := \vec{u}^{\,(2)} + u_j^{(1)} \vec{a}_j\,', \quad \vec{c}\,' := (c_1, \ldots, -c_j, \ldots, c_n)^T.$$

Es sei A' die neue Matrix der Nebenbedingungen. Ist $\vec{x}\,' = (x_1, \ldots, x_j, \ldots, x_n)^T$ eine Lösung von (P), so ist $\vec{x} = (x_1, \ldots, u_j^{(1)} - x_j, \ldots, x_n)^T$ eine Lösung von (LOP). Die Behandlung der neuen Nebenbedingung $\vec{\ell}\,' \leq A'\vec{x}\,' \leq \vec{u}\,'$ erfolgt weiter unten.

$\boxed{-\infty < \ell_j^{(1)} \le x_j :}$ Es sind die folgenden Ersetzungen vorzunehmen:

$$\vec{\ell}\,' := \vec{\ell}^{\,(2)} - \ell_j^{(1)}\vec{a}_j, \quad \vec{u}\,' := \vec{u}^{\,(2)} - \ell_j^{(1)}\vec{a}_j.$$

Ist $\vec{x}\,' = (x_1, \ldots, x_j, \ldots, x_n)^T$ eine Lösung von (P), so ist $\vec{x} = (x_1, \ldots, x_j - \ell_j^{(1)}, \ldots, x_n)^T$ eine Lösung von (LOP). Die Behandlung der neuen Nebenbedingung $\vec{\ell}\,' \le A\vec{x}\,' \le \vec{u}\,'$ erfolgt weiter unten.

$\boxed{-\infty < x_j < +\infty :}$ Die Dimension n ist auf $n+1$ heraufzusetzen, und es sind die folgenden Ersetzungen vorzunehmen:

$$\begin{aligned}
\vec{x} &= (\ldots, x_{j-1}, x_j, x_{j+1}, \ldots)^T \in \mathbf{R}^n &&\Rightarrow\ \vec{x}\,' = (\ldots, x_{j-1}, x_j^+, x_j^-, x_{j+1}, \ldots)^T \in \mathbf{R}^{n+1},\\
\vec{c} &= (\ldots, c_{j-1}, c_j, c_{j+1}, \ldots)^T \in \mathbf{R}^n &&\Rightarrow\ \vec{c}\,' = (\ldots, c_{j-1}, c_j, -c_j, c_{j+1}, \ldots)^T \in \mathbf{R}^{n+1},\\
A &= (\ldots, \vec{a}_{j-1}, \vec{a}_j, \vec{a}_{j+1}, \ldots) \in \mathbf{R}^{(m,n)} &&\Rightarrow\ A' = (\ldots, \vec{a}_{j-1}, \vec{a}_j, -\vec{a}_j, \vec{a}_{j+1}, \ldots) \in \mathbf{R}^{(m,n+1)}.
\end{aligned}$$

Ist $\vec{x}\,' = (\ldots, x_{j-1}, x_j^+, x_j^-, x_{j+1}, \ldots)^T$ eine Lösung von (P), so ist $\vec{x} = (\ldots, x_{j-1}, x_j^+ - x_j^-, x_{j+1}, \ldots,)^T$ eine Lösung von (LOP).

$\boxed{(A\vec{x})_j \le u_j^{(2)} < +\infty :}$ Der Zeilenvektor $\vec{z}_j$ der Matrix A ist durch $\vec{z}_j\,' := -\vec{z}_j$ zu ersetzen. Die neue Matrix A' ist mit dem Einheitsvektor $-\vec{e}_j$ der Standardbasis des $\mathbf{R}^m$ zur Matrix $A'' := (\vec{a}_1\,', \ldots, \vec{a}_n\,', -\vec{e}_j)$ zu erweitern. Ebenso ist $\vec{x}$ mit der Schlupfvariablen y_i zum Vektor $\vec{x}\,' := (x_1, \ldots, x_n, y_i)^T$ zu erweitern, und es ist ein neuer Kostenvektor $\vec{c}\,' := (c_1, \ldots, c_n, 0)^T$ einzuführen. In der Gleichungs–Restriktion $A''\vec{x}\,' = \vec{b}$ muss $b_j := -u_j^{(2)}$ gesetzt werden. Hier können wir $i = 1$ wählen. Bei p Schlupfvariablen haben wir $i = 1, 2, \ldots, p$ zu setzen und die Dimension n auf $n + p$ anzuheben.

$\boxed{-\infty < \ell_j^{(2)} \le (A\vec{x})_j :}$ Die Matrix A ist mit dem Einheitsvektor $-\vec{e}_j$ der Standardbasis des $\mathbf{R}^m$ zur Matrix $A' := (\vec{a}_1, \ldots, \vec{a}_n, -\vec{e}_j)$ zu erweitern. Ebenso ist $\vec{x}$ mit der Schlupfvariablen y_i zum Vektor $\vec{x}\,' := (x_1, \ldots, x_n, y_i)^T$ zu erweitern, und es ist ein neuer Kostenvektor $\vec{c}\,' := (c_1, \ldots, c_n, 0)^T$ einzuführen. In der Gleichungs–Restriktion $A'\vec{x}\,' = \vec{b}$ muss $b_j := \ell_j^{(2)}$ gesetzt werden. Hier können wir $i = 1$ wählen. Bei p Schlupfvariablen haben wir $i = 1, 2, \ldots, p$ zu setzen und die Dimension n auf $n + p$ anzuheben.

BSP. (15.1.1) **Produktionsplanungsproblem.** In einem Unternehmen werden Produkte $P_1, P_2, \ldots, P_n$ hergestellt. Dafür stehen Ressourcen $R_1, R_2, \ldots, R_m$ zur Verfügung (*zum Beispiel* Rohstoffe, Maschinen, Lagerraum etc.), und zwar jeweils b_j Einheiten von $R_j, j = 1, 2, \ldots, m$. Zur Herstellung einer Mengeneinheit des Produktes P_j werden a_{jk} Einheiten der k–ten Ressource R_k benötigt, und der Verkauf einer Mengeneinheit des Produktes P_j bringt einen Gewinn von c_j DM. Ein **Produktionsplan** besteht in der Angabe eines Vektors $\vec{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbf{R}^n$, in welchem die Komponente $x_j \ge 0$ die Anzahl der herzustellenden Einheiten des Produkts P_j angibt. Ein Produktionsplan ist daher **zulässig**, wenn gilt:

$$\sum_{k=1}^{n} a_{jk}x_k \le b_j \ \ \forall\, j = 1, 2, \ldots, m \ \text{ und } \ x_k \ge 0 \ \ \forall\, k = 1, 2, \ldots, n.$$

Der Gesamtgewinn aus dem Produktionsplan ergibt sich zu

$$z := \sum_{k=1}^{n} c_k x_k = \langle \vec{c}, \vec{x} \rangle, \quad \vec{c} := (c_1, c_2, \ldots, c_n)^T \in \mathbf{R}^n.$$

Setzt man noch $A := (a_{jk}) \in \mathbf{R}^{(m,n)}$, $\vec{b} := (b_1, b_2, \ldots, b_m)^T \in \mathbf{R}^m$, so ist das Ziel die Gewinnmaximierung. Das heißt, das *Produktionsplanungsproblem* ist von der Form (LOP), nämlich

$$\boxed{\text{Maximiere } \langle \vec{c}, \vec{x} \rangle \text{ auf der Menge } M := \{\vec{x} \in \mathbf{R}^n \,:\, \vec{x} \geq \vec{0},\ A\vec{x} \leq \vec{b}\}.} \qquad (\mathrm{P}_0)$$

Wir gelangen schnell zur Normalform durch Einführung von m Schlupfvariablen $\vec{y} := (y_1, y_2, \ldots, y_m)^T \in \mathbf{R}^m$:

$$\boxed{\begin{array}{l} \text{Minimiere } \left\langle \begin{bmatrix} -\vec{c} \\ \vec{0} \end{bmatrix}, \begin{bmatrix} \vec{x} \\ \vec{y} \end{bmatrix} \right\rangle \text{ auf der Menge} \\ \qquad M := \left\{ \begin{bmatrix} \vec{x} \\ \vec{y} \end{bmatrix} \in \mathbf{R}^{n+m} \,:\, \begin{bmatrix} \vec{x} \\ \vec{y} \end{bmatrix} \geq \begin{bmatrix} \vec{0} \\ \vec{0} \end{bmatrix}, \quad (A,\, Id) \begin{bmatrix} \vec{x} \\ \vec{y} \end{bmatrix} = \vec{b} \right\}. \end{array}} \qquad (\mathrm{P})$$

Wir haben $(A,\, Id)\binom{\vec{x}}{\vec{y}} = A\vec{x} + \vec{y} = \vec{b}$, so dass $\vec{y} = -A\vec{x} + \vec{b}$ resultiert. Wegen dieses Zusammenhangs kann das (LOP) in das folgende Schema gebracht werden:

$$\boxed{\begin{array}{l} y_j = -\sum_{k=1}^{n} a_{jk} x_k + b_j \geq 0, \quad j = 1, 2, \ldots, m, \\ x_k \geq 0, \quad k = 1, 2, \ldots, n, \\ z = -\sum_{k=1}^{n} c_k x_k \stackrel{!}{=} \text{Min}. \end{array}}$$

Zahlenbeispiel und seine graphische Lösung. Es sollen $n, m, A, \vec{b}, \vec{c}$ in (P_0) in der folgenden Weise spezifiziert werden:

$$n = 2,\ m = 3, \quad A = \begin{bmatrix} 1 & 3 \\ 3 & 1 \\ -1 & 1 \end{bmatrix}, \quad \vec{b} = \begin{bmatrix} 13 \\ 15 \\ 3 \end{bmatrix}, \quad \vec{c} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

In der Normalform liegt somit das folgende lineare Programm vor:

$$\boxed{\begin{array}{rcrcrcrcl} y_1 = & - & x_1 & - & 3x_2 & + & 13 & \geq & 0, \\ y_2 = & - & 3x_1 & - & x_2 & + & 15 & \geq & 0, \\ y_3 = & & x_1 & - & x_2 & + & 3 & \geq & 0, \\ & & x_1 & & & & & \geq & 0, \\ & & & & x_2 & & & \geq & 0, \\ z = & - & x_1 & - & x_2 & & & \stackrel{!}{=} & \text{Min}. \end{array}} \qquad (1.3)$$

Da nur *zwei* Unbekannte x_1, x_2 auftreten, kann das lineare Programm (1.3) **graphisch** in der (x_1, x_2)–Ebene gelöst werden. Die Menge aller Punkte $P(x_1, x_2)$, welche einer linearen Ungleichung

$$y_j = \alpha_j x_1 + \beta_j x_2 + \gamma_j \geq 0$$

genügen, bilden eine *abgeschlossene Halbebene* $HE_j := \{(x_1, x_2)^T \,:\, \alpha_j x_1 + \beta_j x_2 + \gamma_j \geq 0\}$ mit *Randlinie* (= Gerade)

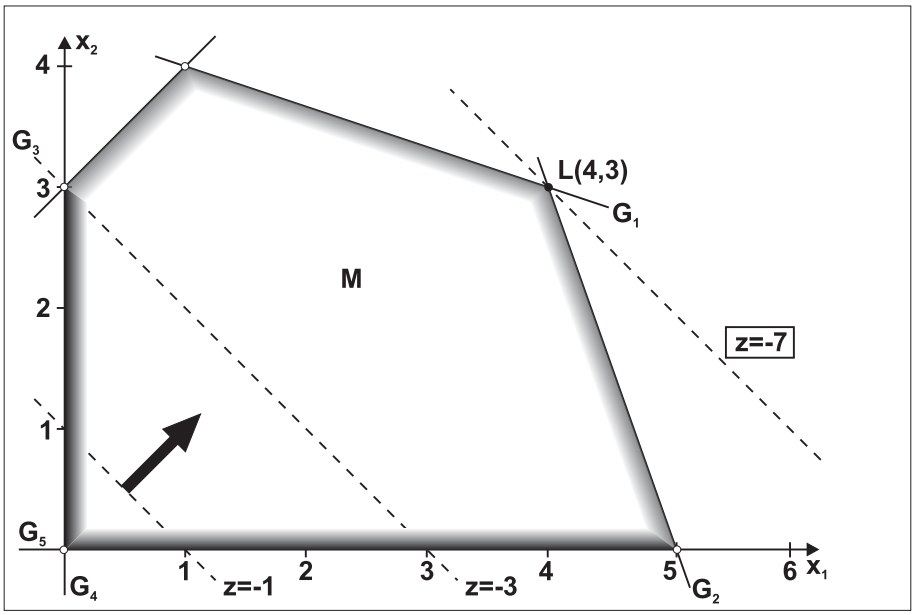
$$G_j := \{(x_1, x_2)^T \,:\, \alpha_j x_1 + \beta_j x_2 + \gamma_j = 0\}.$$

Ist $\gamma_j \geq 0$, so gehört offensichtlich der Koordinatenursprung $(0,0)^T$ zu HE_j; für $\gamma_j < 0$ liegt er im Komplement $\mathbf{R}^2 \setminus HE_j$. Die Menge M der zulässigen Lösungen für das lineare Programm (1.3) ist

somit der Durchschnitt der fünf Halbebenen

$$M := \bigcap_{j=1}^{5} HE_j,$$

j	1	2	3	4	5
α_j	-1	-3	1	1	0
β_j	-3	-1	-1	0	1
γ_j	13	15	3	0	0



Graphische Lösung des linearen Progamms (1.3)

Die Graphik zeigt, dass $M \neq \emptyset$ gilt. Somit ist das lineare Programm (1.3) zulässig. Darüber hinaus muss im Fall von zwei Unbekannten stets ein **konvexer Bereich** M entstehen. Eine Lösung des linearen Programms muss nun ein Punkt in M oder auf dem Rand von M sein.

Wir untersuchen die Zielfunktion $z = -x_1 - x_2$. Ihre *Niveaulinien* $z = \text{const}$ bilden eine Schar paralleler Geraden $G_z := \{(x_1, x_2)^T : -x_1 - x_2 - z = 0\}$. In der Graphik sind die Niveaulinien $z = -1$ sowie $z = -3$ eingezeichnet, und der Pfeil gibt die Richtung an, in welcher der Wert von z abnimmt. Offenbar nimmt die Zielfunktion auf der Menge M ihren minimalen Wert im Punkt $(x_1, x_2)^T = (4, 3)^T \in M$ an, denn dieser Punkt hat den größten Abstand von den eingezeichneten Niveaulinien in Pfeilrichtung.

Der **Lösungspunkt** $(x_1, x_2)^T = (4, 3)^T$ des linearen Programms (1.3) ist eine **Ecke** des zulässigen Bereichs M.

BSP. (15.1.2) **Diätproblem.** Ein Mensch benötigt Vitamine $V_1, V_2, \ldots, V_m$, und zwar von jedem Vitamin V_j mindestens b_j Einheiten pro Tag. Zur Versorgung stehen Lebensmittel $L_1, L_2, \ldots, L_n$ zur Verfügung. Das k–te Lebensmittel L_k enthalte a_{jk} Einheiten des Vitamins V_j, und eine Einheit von L_k koste c_k DM. Ein *Diätplan* besteht in der Angabe eines Vektors $\vec{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbf{R}^n$, in welchem die Komponente x_k angibt, dass x_k Einheiten vom Lebensmittel L_k zu nehmen sind. Ein **zulässiger** Diätplan erfordert daher

$$\sum_{k=1}^{n} a_{jk} x_k \geq b_j \;\; \forall\, j = 1, 2, \ldots, m \;\; \text{und} \;\; x_k \geq 0 \;\; \forall\, k = 1, 2, \ldots, n.$$

Die Kosten eines Diätplans betragen

$$z := \sum_{k=1}^{n} c_k x_k = \langle \vec{c}, \vec{x} \rangle, \qquad \vec{c} = (c_1, c_2, \ldots, c_n)^T \in \mathbf{R}^n.$$

Setzen wir wieder $A := (a_{jk}) \in \mathbf{R}^{(m,n)}$, $\vec{b} := (b_1, b_2, \ldots, b_m)^T \in \mathbf{R}^m$, so ist das Ziel eine ausreichende tägliche Vitaminversorgung bei minimalen Kosten. Das Diätproblem ist ein (LOP) in der Form

$$\boxed{\text{Minimiere } \langle \vec{c}, \vec{x} \rangle \text{ auf der Menge } M := \{\vec{x} \in \mathbf{R}^n \,:\, \vec{x} \geq \vec{0},\ A\vec{x} \geq \vec{b}\}.} \qquad (\mathrm{P}_0)$$

Zur Herstellung der Normalform ist auch hier die Einführung von Schlupfvariablen $\vec{y} = (y_1, y_2, \ldots, y_m)^T \in \mathbf{R}^m$ erforderlich, und es folgt ganz analog wie in BSP. (15.1.1):

$$\boxed{\begin{array}{l} \text{Minimiere } \left\langle \begin{bmatrix} \vec{c} \\ \vec{0} \end{bmatrix}, \begin{bmatrix} \vec{x} \\ \vec{y} \end{bmatrix} \right\rangle \text{ auf der Menge} \\ \qquad M := \left\{ \begin{bmatrix} \vec{x} \\ \vec{y} \end{bmatrix} \in \mathbf{R}^{n+m} \,:\, \begin{bmatrix} \vec{x} \\ \vec{y} \end{bmatrix} \geq \begin{bmatrix} \vec{0} \\ \vec{0} \end{bmatrix}, \quad (A, -Id) \begin{bmatrix} \vec{x} \\ \vec{y} \end{bmatrix} = \vec{b} \right\}. \end{array}} \qquad (\mathrm{P})$$

Hierfür kann das folgende Schema erstellt werden:

$$\boxed{\begin{array}{l} y_j = \displaystyle\sum_{k=1}^{n} a_{jk} x_k - b_j \geq 0, \quad j = 1, 2, \ldots, m, \\ x_k \geq 0, \quad k = 1, 2, \ldots, n, \\ z = \displaystyle\sum_{k=1}^{n} c_k x_k \stackrel{!}{=} \text{Min.} \end{array}}$$

Zahlenbeispiel und seine graphische Lösung. Es sollen $n, m, A, \vec{b}, \vec{c}$ in (P_0) in der folgenden Weise spezifiziert werden:

$$n = 2,\ m = 4, \quad A = \begin{bmatrix} 12 & 9 \\ 6 & 15 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \vec{b} = \begin{bmatrix} 108 \\ 90 \\ 2 \\ 3 \end{bmatrix}, \quad \vec{c} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

Es ergibt sich somit das folgende lineare Programm in der Normalform:

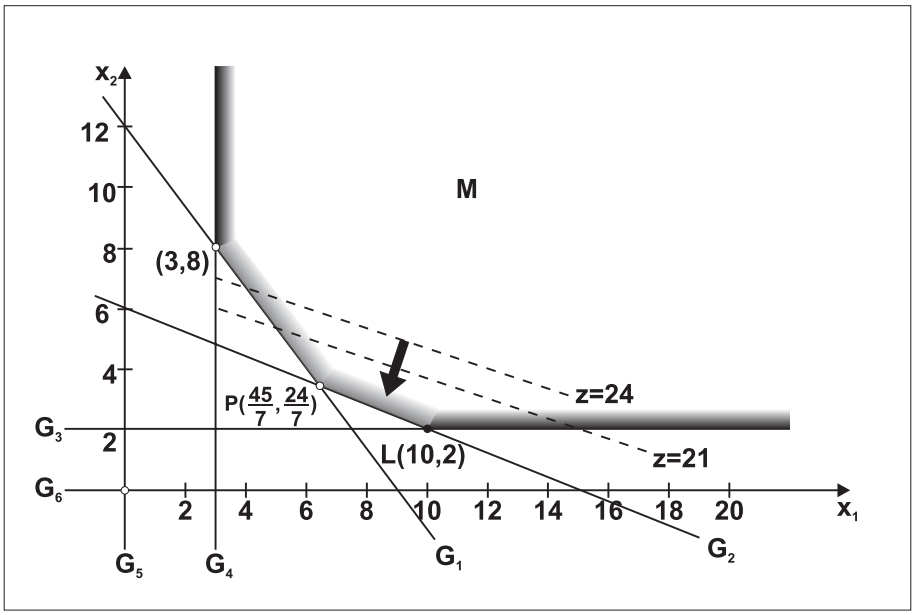
$$\boxed{\begin{array}{rrcrcrcl} y_1 = & 12x_1 & + & 9x_2 & - & 108 & \geq & 0, \\ y_2 = & 6x_1 & + & 15x_2 & - & 90 & \geq & 0, \\ y_3 = & & & x_2 & - & 2 & \geq & 0, \\ y_4 = & x_1 & & & - & 3 & \geq & 0, \\ & x_1 & & & & & \geq & 0, \\ & & & x_2 & & & \geq & 0, \\ z = & x_1 & + & 3x_2 & & & \stackrel{!}{=} & \text{Min.} \end{array}} \qquad (1.4)$$

Die Menge M der zulässigen Lösungen ist wiederum ein **konvexer Bereich**

$$M := \bigcap_{j=1}^{6} HE_j, \quad HE_j := \{(x_1, x_2)^T \,:\, \alpha_j x_1 + \beta_j x_2 + \gamma_j \geq 0\},$$

worin die Zahlen $\alpha_j, \beta_j, \gamma_j$ der folgenden Tabelle zu entnehmen sind:

j	1	2	3	4	5	6
α_j	12	6	0	1	1	0
β_j	9	15	1	0	0	1
γ_j	-108	-90	-2	-3	0	0



Graphische Lösung des linearen Progamms (1.4)

Wegen $M \neq \emptyset$ liegt ein *zulässiges* lineares Programm vor. In der obigen Graphik sind die zwei Niveaulinien $z = 21$ und $z = 24$ der Zielfunktion $z = x_1 + 3x_2$ eingetragen sowie die Richtung abnehmender Werte von z. Offensichtlich wird z in der Ecke $L(10, 2)$ minimal, so dass das lineare Programm (1.4) die optimale Lösung $(x_1, x_2)^T = (10, 2)^T$ hat. Das Kostenfunktional hat in diesem Punkt den Wert $z_{\min} = 16$.

Bemerkung 15.2 In den beiden Beispielen (15.1.1) und (15.1.2) ist die optimale Lösung jeweils **eindeutig** bestimmt, und sie wird durch eine **Ecke** des Polygonbereichs M festgelegt. Hätte in BSP. (15.1.2) die Zielfunktion z die Form

$$z = 2x_1 + 5x_2,$$

so wären die Niveaulinien $z = \text{const}$ Parallelen zur Randlinie G_2. Das Minimum von z würde in **jedem** Punkt $(x_1, x_2)^T \in G_2 \cap M$ erreicht, also auf einer ganzen **Kante** des polygonalen Randes von M. Die Lösung des linearen Programms wäre **nicht eindeutig**; jedoch sind die beiden Endpunkte $P\left(\frac{45}{7}, \frac{24}{7}\right)$ und $L(10, 2)$ der Kante als Ecken des Polygons zwei spezielle Lösungen. Jeder weitere Punkt der Kante ist eine **konvexe Linearkombination** dieser beiden Lösungen: □

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1-\lambda}{7} \begin{bmatrix} 45 \\ 24 \end{bmatrix} + \lambda \begin{bmatrix} 10 \\ 2 \end{bmatrix}, \quad 0 < \lambda < 1.$$

15.2 Der Simplexalgorithmus

15.2.1 Geometrische Grundlagen

Wir betrachten die **Normalform** eines linearen Programmes

$$\boxed{\text{Minimiere } \langle \vec{c}, \vec{x} \rangle \text{ auf der Menge } M := \{\vec{x} \in \mathbf{R}^n \,:\, \vec{x} \geq \vec{0},\ A\vec{x} = \vec{b}\},} \qquad \text{(P)}$$

worin $A \in \mathbf{R}^{(m,n)}, \vec{b} \in \mathbf{R}^m$ und $\vec{c} \in \mathbf{R}^n$ die Vorgaben sind. Die wesentliche Idee für einen Algorithmus zur numerischen Lösung von (P) ist bereits durch die beiden Beispiele (15.1.1) und (15.1.2) aufgezeigt worden:

(A) Untersuche die Menge M der zulässigen Punkte. Ist $M \neq \emptyset$, so bildet M ein konvexes Polyeder.

(B) Bestimme die **Ecke** (Regelfall) oder **Kante** (Ausnahmefall) des Polyeders, in der die Zielfunktion minimal wird.

Die folgenden Fälle sind dabei zu unterscheiden:

(a) M ist leer. Dann existiert keine Lösung des linearen Programms (P).

(b) M ist eine unbeschränkte Menge. Dann ergeben sich zwei Möglichkeiten:

(i) $\langle \vec{c}, \vec{x} \rangle$ ist auf M nach unten **nicht** beschränkt. Dann existiert keine Lösung.

(ii) $\langle \vec{c}, \vec{x} \rangle$ ist auf M nach unten beschränkt. Dann gibt es eine Lösung.

(c) $\emptyset \neq M \subset \mathbf{R}^n$ ist beschränkt. Dann existiert stets eine Lösung.

Es sollen nun die theoretischen Grundlagen für diese Statements untersucht werden. Dazu beginnen wir mit der Erklärung der erforderlichen Begriffsbildungen.

Definition 15.3 *Eine Teilmenge $\emptyset \neq C \subset \mathbf{R}^n$ heiße* **konvex** *genau dann, wenn mit je zwei Punkten $\vec{x}, \vec{y} \in C$ auch die Strecke zwischen $\vec{x}$ und $\vec{y}$ ganz in C liegt:*

$$\boxed{\forall\, \vec{x}, \vec{y} \in C : (1-\lambda)\vec{x} + \lambda\vec{y} \in C \ \forall\, \lambda \in [0,1].}$$

BSP. (15.2.1) Die Menge $M := \{\vec{x} \in \mathbf{R}^n : \vec{x} \geq \vec{0},\ A\vec{x} = \vec{b}\}$ der zulässigen Lösungen von (P) ist konvex. Denn für $\vec{x}, \vec{y} \in M$ gelten $\vec{x} \geq \vec{0},\ \vec{y} \geq \vec{0}$ sowie $A\vec{x} = \vec{b}$ und $A\vec{y} = \vec{b}$. Daraus folgt für jede Zahl $\lambda \in [0,1]$:

$$(1-\lambda)\vec{x} + \lambda\vec{y} \geq \vec{0}, \quad A[(1-\lambda)\vec{x} + \lambda\vec{y}] = (1-\lambda)A\vec{x} + \lambda A\vec{y} = (1-\lambda)\vec{b} + \lambda\vec{b} = \vec{b},$$

also $(1-\lambda)\vec{x} + \lambda\vec{y} \in M$.

Definition 15.4 *Ein Vektor $\vec{x} \in \mathbf{R}^n$ heiße* **konvexe Linearkombination** *(konvexe LK) von $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N \in \mathbf{R}^n$, wenn gilt:*

$$\boxed{\vec{x} = \sum_{k=1}^{N} \lambda_k \vec{x}_k \quad \text{mit} \quad \lambda_k \geq 0 \ \forall\, k = 1, 2, \ldots, N \quad \text{und} \quad \sum_{k=1}^{N} \lambda_k = 1.}$$

Definition 15.5 (a) *Ein Punkt $\vec{x} \in C$ heiße eine* **Ecke** *der konvexen Menge $C \subset \mathbf{R}^n$ genau dann, wenn $\vec{x}$* **keine** *konvexe LK von zwei anderen Punkten aus C ist.*

(b) *Für festes $\vec{0} \neq \vec{n} \in \mathbf{R}^n$ und $\alpha \in \mathbf{R}$ heißen $H := \{\vec{x} \in \mathbf{R}^n : \langle \vec{x}, \vec{n} \rangle = \alpha\}$ eine* **Hyperebene** *in $\mathbf{R}^n$ und $H^- := \{\vec{x} \in \mathbf{R}^n : \langle \vec{x}, \vec{n} \rangle \leq \alpha\}$ ein zugehöriger* **abgeschlossener Halbraum**. *Dieser ist offenbar konvex.*

(c) *Die Schnittmenge endlich vieler abgeschlossener Halbräume heiße ein* **Polyeder**. *Dieses ist als Durchschnitt konvexer Mengen wieder konvex.*

(d) *Ein nichtleeres* **beschränktes** *Polyeder heiße ein* **Polytop**.

In dem folgenden Satz geben wir an, wann die Menge M der zulässigen Lösungen eines linearen Programms (P) ein Polytop ist.

Satz 15.3 *Die Menge $M := \{\vec{x} \in \mathbf{R}^n : \vec{x} \geq \vec{0},\ A\vec{x} = \vec{b}\}$ sei nichtleer. Genau dann ist M ein Polytop, wenn gilt:*

$$\boxed{\vec{y} \in \mathbf{R}^n \setminus \{\vec{0}\} \quad \text{und} \quad \vec{y} \geq \vec{0} \quad \Rightarrow \quad A\vec{y} \neq \vec{0}.} \tag{2.1}$$

Begründung: (a) Würde ein Vektor $\vec{0} \neq \vec{y} \geq \vec{0}$ mit $A\vec{y} = \vec{0}$ existieren, so hätten wir für jedes $\vec{x} \in M$ auch $\vec{x} + t\vec{y} \in M \ \forall\ t \geq 0$. Somit wäre M unbeschränkt, da $\|\vec{x} + t\vec{y}\|_2 \geq t\|\vec{y}\|_2 - \|\vec{x}\|_2$ für $t \to \infty$ beliebig groß werden kann.

(b) Ist M unbeschränkt, so muss es eine Folge $(\vec{x}_k)_{k\in\mathbf{N}} \subset M$ mit $\|\vec{x}_k\|_2 \to \infty$ geben. Die beschränkte Menge $\{\vec{y}_k \in \mathbf{R}^n : \vec{y}_k := \vec{x}_k/\|\vec{x}_k\|_2\}$ besitzt dann einen Häufungspunkt $\vec{y} \neq \vec{0}$ (Satz 13.9 von HEINE–BOREL) mit $\vec{y} \geq \vec{0}$ und

$$A\vec{y} = \lim_{k\to\infty} \frac{A\vec{x}_k}{\|\vec{x}_k\|_2} = \lim_{k\to\infty} \frac{\vec{b}}{\|\vec{x}_k\|_2} = \vec{0}.$$

Also kann (2.1) nicht gelten. □

Bemerkung 15.3 Die Bedingung (2.1) ist eher für theoretische Aussagen geeignet als zur Herleitung eines nachprüfbaren Kriteriums für die Beschränktheit der Menge M der zulässigen Lösungen. Trotzdem kann man häufig durch Lösung des linearen Gleichungssystems $A\vec{y} = \vec{0}$ eine einfache Entscheidung darüber treffen, dass **keine** Lösungen $\vec{0} \neq \vec{y} \geq \vec{0}$ existieren, wie in den folgenden Beispielen gezeigt wird. □

BSP. (15.2.2) (a) Wir betrachten die Matrix

$$(A,\, Id) = \begin{bmatrix} 1 & 3 & 1 & 0 & 0 \\ 3 & 1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

aus BSP. (15.1.1). Mit Hilfe des GAUSS–Algorithmus erzeugen wir die Stufenform

$$(A,\, Id\,|\,\vec{0}) \quad \rightleftharpoons \quad \left[\begin{array}{ccccc|c} 1 & 3 & 1 & 0 & 0 & 0 \\ 3 & 1 & 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 1 & 0 \end{array}\right] \quad \rightleftharpoons \quad \left[\begin{array}{ccccc|c} 1 & 3 & 1 & 0 & 0 & 0 \\ 0 & 4 & 1 & 0 & 1 & 0 \\ 0 & 0 & -1 & \boxed{1} & \boxed{2} & 0 \end{array}\right]$$

Es resultiert $\dim \mathrm{Kern}\,(A,\, Id) = 2$, und eine Basis von $\mathrm{Kern}\,(A,\, Id)$ lautet zum Beispiel:

$$\mathrm{Kern}\,(A,\, Id) = \mathrm{span}\,\{(-1,-1,4,4,0)^T, (1,-3,8,0,4)^T\}.$$

Es wäre somit $(A,\, Id)\vec{y} = \vec{0}$ genau für

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = C_1 \begin{bmatrix} -1 \\ -1 \\ 4 \\ 4 \\ 0 \end{bmatrix} + C_2 \begin{bmatrix} 1 \\ -3 \\ 8 \\ 0 \\ 4 \end{bmatrix},$$

und wegen $y_4 = 4C_1 \geq 0, y_5 = 4C_2 \geq 0$ müssten $C_1 \geq 0, C_2 \geq 0, C_1^2 + C_2^2 > 0$, gelten. Dann wäre aber $y_2 = -C_1 - 3C_2 < 0$, so dass eine Lösung $\vec{0} \neq \vec{y} \in \mathrm{Kern}\,(A,\, Id)$ mit $\vec{y} \geq \vec{0}$ nicht existieren kann. Also ist die Menge M der zulässigen Lösungen beschränkt.

(b) Wir betrachten die Matrix

$$(A, -Id) = \begin{bmatrix} 12 & 9 & -1 & 0 & 0 & 0 \\ 6 & 15 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}$$

aus BSP. (15.1.2). Eine Transformation auf Stufenform erbringt:

$$(A, -Id\,|\,\vec{0}) \quad \rightleftharpoons \quad \left(\begin{array}{cccccc|c} 12 & 9 & -1 & 0 & 0 & 0 & 0 \\ 6 & 15 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 & 0 \end{array}\right) \quad \rightleftharpoons \quad \left(\begin{array}{cccccc|c} 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 9 & 12 & 0 \\ 0 & 0 & 0 & -1 & \boxed{15} & \boxed{6} & 0 \end{array}\right)$$

Wir haben wiederum $\dim \mathrm{Kern}\,(A, -Id) = 2$, und eine Basis von $\mathrm{Kern}\,(A, -Id)$ lautet hier:

$$\mathrm{Kern}\,(A, -Id) = \mathrm{span}\,\{(1, 0, 12, 6, 0, 1)^T, (0, 1, 9, 15, 1, 0)^T\}.$$

Eine Lösung $\vec{y}$ von $(A, -Id)\vec{y} = \vec{0}$ hat somit die Form

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = C_1 \begin{bmatrix} 1 \\ 0 \\ 12 \\ 6 \\ 0 \\ 1 \end{bmatrix} + C_2 \begin{bmatrix} 0 \\ 1 \\ 9 \\ 15 \\ 1 \\ 0 \end{bmatrix},$$

und wir haben $\vec{0} \neq \vec{y} \geq \vec{0}$ für alle $C_1 \geq 0, C_2 \geq 0$ mit $C_1^2 + C_2^2 > 0$. In der Tat ist die Menge M der zulässigen Lösungen unbeschränkt.

Wir geben nun eine *algebraische* Charakterisierung der Ecken von M an.

Satz 15.4 *Die Menge $M := \{\vec{x} \in \mathbf{R}^n : \vec{x} \geq \vec{0},\ A\vec{x} = \vec{b}\}$ der zulässigen Lösungen sei nichtleer, und es gelte $A = (\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n)$.*
(a) *Genau dann ist $\vec{x} = (x_1, x_2, \ldots, x_n)^T \in M$ eine Ecke von M, wenn die zu positiven Komponenten $x_j > 0$ gehörigen Spaltenvektoren der Matrix A linear unabhängig sind: Für die Indexmenge $B(\vec{x}) := \{j \in \mathbf{N} : 1 \leq j \leq n \ \text{ und } \ x_j > 0\}$ ist das folgende Vektorsystem linear unabhängig:*

$$\boxed{\{\vec{a}_j : j \in B(\vec{x})\}.} \tag{2.2}$$

(b) *M besitzt mindestens eine, aber höchstens endlich viele Ecken.*

Begründung: (a) Wäre das System (2.2) linear abhängig, so gäbe es Zahlen λ_j, nicht alle Null, mit

$$\sum_{j \in B(\vec{x})} \lambda_j \vec{a}_j = \vec{0}.$$

Wegen $x_j > 0$ finden wir ein $\delta > 0$ derart, dass $x_j \pm \delta\lambda_j \geq 0\ \forall\, j \in B(\vec{x})$ gilt. Wir setzen nun

$$x_j^+ := \begin{cases} x_j + \delta\lambda_j & \text{für } j \in B(\vec{x}), \\ 0 & \text{für } j \notin B(\vec{x}), \end{cases} \qquad x_j^- := \begin{cases} x_j - \delta\lambda_j & \text{für } j \in B(\vec{x}), \\ 0 & \text{für } j \notin B(\vec{x}). \end{cases}$$

Es gilt dann unter Beachtung von $x_j = 0$, $j \notin B(\vec{x})$:

$$A\vec{x}^{\pm} = \sum_{j \in B(\vec{x})} x_j \vec{a}_j \pm \delta \sum_{j \in B(\vec{x})} \lambda_j \vec{a}_j = \sum_{j=1}^{n} x_j \vec{a}_j = \vec{b},$$

und somit sind $\vec{x}^{\pm} \in M$ zwei verschiedene Punkte, deren Mittelpunkt $\vec{x} = \frac{1}{2}(\vec{x}^+ + \vec{x}^-)$ wegen der Konvexität keine Ecke sein kann.

Ist das Vektorsystem (2.2) linear unabhängig, so nehmen wir an, es gebe eine konvexe LK $\vec{x} = (1-\lambda)\vec{x}^{\,+} + \lambda\vec{x}^{\,-}$ mit zwei Punkten $\vec{0} \le \vec{x}^{\,\pm} \in M$ und $\lambda \in (0,1)$. Wegen $x_j = 0 \;\forall\; j \notin B(\vec{x})$ muss auch $x_j^+ = x_j^- = 0 \;\forall\; j \notin B(\vec{x})$ gelten. Somit erschließen wir

$$\vec{0} = \vec{b} - \vec{b} = A(\vec{x}^{\,+} - \vec{x}^{\,-}) = \sum_{j \in B(\vec{x})} (x_j^+ - x_j^-)\vec{a}_j,$$

und wegen der linearen Unabhängigkeit folgt $x_j^+ - x_j^- = 0$, also $\vec{x}^{\,+} = \vec{x}^{\,-}$. Es existiert somit keine konvexe LK für $\vec{x}$. Das heißt, $\vec{x}$ ist eine Ecke von M.

(b) Es ist klar, aus den n Spaltenvektoren $\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n$ können nur auf endlich viele Arten linear unabhängige Teilsysteme kombiniert werden. Also existieren nur endlich viele Ecken. Die Existenz von mindestens einer Ecke erhält man aus folgender Überlegung. Es existiert sicher ein $\vec{x}^{\,*} \in M$ mit einer minimalen Anzahl *positiver* Komponenten. Deren Indexmenge sei $B(\vec{x}^{\,*}) \neq \emptyset$. Wäre diese Menge leer, so wäre $\vec{x}^{\,*} = \vec{0} \in M$, und dieser Punkt ist eine Ecke. Das System $\{\vec{a}_j \,:\, j \in B(\vec{x}^{\,*})\}$ ist linear unabhängig, denn andernfalls könnten Punkte $\vec{x}^{\,+}$ oder $\vec{x}^{\,-}$ so konstruiert werden (man wähle analog (a) ein $\delta > 0$ derart, dass $x_j + \delta\lambda_j = 0$ oder $x_j - \delta\lambda_j = 0$ für ein $j \in B(\vec{x}^{\,*})$ gilt), dass diese weniger positive Komponenten haben. Dies wäre ein Widerspruch zur Wahl von $\vec{x}^{\,*}$. Also ist $\vec{x}^{\,*}$ eine Ecke von M. □

Wir zeigen in einem nächsten Satz, dass jeder Punkt in der Menge M der zulässigen Lösungen eine Konvexkombination der endlich vielen Ecken von M ist.

Satz 15.5 *Die Menge $M := \{\vec{x} \in \mathbf{R}^n \,:\, \vec{x} \ge \vec{0},\; A\vec{x} = \vec{b}\}$ der zulässigen Lösungen sei nichtleer, und es gelte $A = (\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n)$. Es sei $\{\vec{v}_i \,:\, i \in I\}$ die nichtleere endliche Menge der Ecken von M. Dann gilt für jeden Punkt $\vec{x} \in M$ eine Darstellung*

$$\boxed{\vec{x} = \sum_{k \in I} \lambda_k \vec{v}_k + \vec{d} \;\; mit \;\; \lambda_k \ge 0, \quad \sum_{k \in I} \lambda_k = 1,} \tag{2.3}$$

und einem Vektor $\vec{d} \ge \vec{0},\; A\vec{d} = \vec{0}$.

Bemerkung 15.4 Ist M beschränkt, also ein Polytop, so ist $\vec{d} = \vec{0}$ gemäß Satz 15.3 der einzige Vektor mit $\vec{d} \ge \vec{0}$ und $A\vec{d} = \vec{0}$. In diesem Fall ist jeder Punkt $\vec{x} \in M$ gemäß (2.3) eine Konvexkombination der Ecken von M. □

Begründung: Wir führen vollständige Induktion nach der Anzahl $p \le n$ *positiver* Komponenten von $\vec{x} \in M$ durch.

Induktionsverankerung: Ist $p = 0$, so gilt $\vec{x} = \vec{0}$, und dieser Punkt bildet eine Ecke von M. Die Darstellung (2.3) ist trivial.

Vererbung: Nun sei $p \ge 1$ sowie $\vec{x} \in M$. Ist $\vec{x}$ eine Ecke, so ist (2.3) wiederum trivial. Wir nehmen deshalb an, $\vec{x}$ sei keine Ecke von M. Wir setzen $B(\vec{x}) := \{j \in \mathbf{N} \,:\, 1 \le j \le n \;\text{ und }\; x_j > 0\}$ mit $\operatorname{card} B(\vec{x}) = p$. Gemäß Satz 15.4(a) muss das System der Spaltenvektoren $\{\vec{a}_j \,:\, j \in B(\vec{x})\}$ linear abhängig sein, so dass das homogene lineare Gleichungssystem $A\vec{w} =: \sum_{j=1}^{n} w_j \vec{a}_j = \vec{0}$ eine nichttriviale Lösung $\vec{0} \neq \vec{w} \in \mathbf{R}^n$ mit $w_j = 0 \;\forall\; j \notin B(\vec{x})$ zulässt. Wir treffen drei Fallunterscheidungen gemäß $\vec{w} \ge \vec{0}$, $\vec{w} \le \vec{0}$ bzw. $\vec{w}$ hat Komponenten beiderlei Vorzeichens.

Fall (i): $\vec{w}$ hat positive *und* negative Komponenten. Wir definieren $\delta_1 > 0$ und $\delta_2 > 0$ gemäß

$$\delta_1 := \min\Big\{\frac{x_j}{w_j} \,:\, j \in B(\vec{x}) \;\text{ und }\; w_j > 0\Big\}, \quad \delta_2 := \min\Big\{-\frac{x_j}{w_j} \,:\, j \in B(\vec{x}) \;\text{ und }\; w_j < 0\Big\}.$$

Setzen wir $\vec{x}^{\,1} := \vec{x} - \delta_1\vec{w},\; \vec{x}^{\,2} := \vec{x} + \delta_2\vec{w}$, so folgt $\vec{x}^{\,i} \ge \vec{0},\; A\vec{x}^{\,i} = A\vec{x} \pm \delta_i A\vec{w} = \vec{b}$, und somit $\vec{x}^{\,i} \in M$, $i = 1, 2$. Ferner haben die Punkte $\vec{x}^{\,i}$ höchstens $p-1$ positive Komponenten, und es gilt

$$\vec{x} = (1-\mu)\vec{x}^{\,1} + \mu\vec{x}^{\,2} \;\text{ mit }\; \mu := \frac{\delta_1}{\delta_1 + \delta_2} \in (0,1). \tag{2.4}$$

Nach Induktionsvoraussetzung gelten für $\vec{x}^{\,i}$ Darstellungen von der Form (2.3), nämlich

$$\vec{x}^{\,i} = \sum_{k\in I} \lambda_k^i \vec{v}_k + \vec{d}^{\,i} \quad \text{mit} \quad \lambda_k^i \geq 0, \quad \sum_{k\in I} \lambda_k^i = 1, \quad i = 1, 2.$$

Wir setzen

$$\lambda_k := (1-\mu)\lambda_k^1 + \mu\lambda_k^2, \quad k \in I, \quad \vec{d} := (1-\mu)\vec{d}^{\,1} + \mu\vec{d}^{\,2}.$$

Dann ergibt sich aus (2.4) schon die behauptete Darstellung

$$\vec{x} = \sum_{k\in I} \lambda_k \vec{v}_k + \vec{d} \ \text{ mit } \ \lambda_k \geq 0, \quad \sum_{k\in I} \lambda_k = 1.$$

Fall (ii): $\vec{w} \geq \vec{0}$. Wie im Fall (i) sind $\delta_1 > 0$ und $\vec{x}^{\,1} := \vec{x} - \delta_1\vec{w} \in M$ wohldefiniert, und auf $\vec{x}^{\,1}$ trifft nach Induktionsvoraussetzung die Darstellung (2.3) zu. Wegen $\vec{x} = \vec{x}^{\,1} + \delta_1\vec{w}$ folgt daraus die behauptete Darstellung (2.3) für $\vec{x}$.

Fall (iii): $\vec{w} \leq \vec{0}$. Diesen behandelt man entsprechend dem Fall (ii). Damit ist der Induktionsbeweis abgeschlossen. □

Nach diesen vorbereitenden Sätzen sind wir nun in der Lage, das zentrale Resultat für die Anwendung des Simplexverfahrens zu beweisen.

Satz 15.6 *Für das lineare Programm in der Normalform*

$$\boxed{\textit{Minimiere } \langle \vec{c}, \vec{x}\rangle \textit{ auf der Menge } M := \{\vec{x} \in \mathbf{R}^n \,:\, \vec{x} \geq \vec{0},\ A\vec{x} = \vec{b}\}} \qquad \text{(P)}$$

gelte $A = (\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n) \in \mathbf{R}^{(m,n)}$, $\vec{b} \in \mathbf{R}^m$, $\vec{c} \in \mathbf{R}^n$ *sowie* $M \neq \emptyset$. *Dann ist genau eine der beiden Aussagen* (a), (b) *wahr:*

(a) *Das Problem* (P) *besitzt eine der endlich vielen Ecken von* M *als Lösung.*

(b) *Es ist* $\inf\{\langle \vec{c}, \vec{x}\rangle \,:\, \vec{x} \in M\} = -\infty$, *d.h. die Zielfunktion von* (P) *ist auf der Menge* M *der zulässigen Lösungen nach unten nicht beschränkt.*

Begründung: (i) Existiert ein $\vec{d} \geq \vec{0}$ mit $A\vec{d} = \vec{0}$ und $\langle \vec{c}, \vec{d}\rangle < 0$, so gilt notwendig $\inf\limits_{\vec{x}\in M} \langle \vec{c}, \vec{x}\rangle = -\infty$.

Denn in diesem Fall ist M gemäß Satz 15.3 unbeschränkt, so dass mit jedem $\vec{x}^{\,*} \in M$ auch $\vec{x}^{\,*} + t\vec{d} \in M \ \forall\, t \geq 0$ folgt. Wir haben demgemäß

$$\inf_{\vec{x}\in M} \langle \vec{c}, \vec{x}\rangle \leq \liminf_{t\to+\infty} \langle \vec{c}, \vec{x}^{\,*} + t\vec{d}\rangle = \langle \vec{c}, \vec{x}^{\,*}\rangle + \liminf_{t\to+\infty} t\,\langle \vec{c}, \vec{d}\rangle = -\infty.$$

Dies ist genau die behauptete Aussage (b).

(ii) Sei nun $\langle \vec{c}, \vec{d}\rangle \geq 0$ für alle $\vec{d} \geq \vec{0}$ mit $A\vec{d} = \vec{0}$. Wie vorher bezeichne $\{\vec{v}_i \,:\, i \in I\}$ die nichtleere endliche Menge der Ecken von M. Dann gilt für jedes $\vec{x} \in M$ die Darstellung (2.3), und somit

$$\langle \vec{c}, \vec{x}\rangle = \Big\langle \vec{c}, \sum_{k\in I} \lambda_k \vec{v}_k + \vec{d}\Big\rangle = \sum_{k\in I} \lambda_k \langle \vec{c}, \vec{v}_k\rangle + \underbrace{\langle \vec{c}, \vec{d}\rangle}_{\geq 0} \geq \min_{i\in I} \langle \vec{c}, \vec{v}_i\rangle.$$

Dies ist aber genau die behauptete Aussage (a). □

Bemerkung 15.5 (a) Hat das lineare Programm (P) eine Lösung $\vec{x}$, so folgt aus Satz 15.6, dass eine der endlich vielen Ecken von M ebenfalls Lösung von (P) ist. Im Prinzip könnte man also durch Berechnung aller Ecken von M diejenige mit minimalem Kostenfunktional berechnen. Da M jedoch häufig sehr viele Ecken besitzt, verbietet sich eine solche Vorgehensweise.

(b) Für zulässige lineare Programme (P) (also für $M \neq \emptyset$) liefert Satz 15.6 eine Existenzaussage:

- *Gelte* $M \neq \emptyset$ *sowie* $\inf\limits_{\vec{x}\in M} \langle \vec{c}, \vec{x}\rangle > -\infty$. *Dann besitzt das lineare Programm* (P) *eine Lösung.* □

Für die weitere Untersuchung des linearen Programms (P) stellen wir die folgende Rangbedingung an die Matrix $A \in \mathbf{R}^{(m,n)}$:

$$\boxed{\operatorname{Rang} A = m.} \tag{R}$$

Bemerkung 15.6 Theoretisch bedeutet (R) keine Einschränkung der Allgemeinheit. Falls (R) nicht gilt, so ist entweder das lineare Gleichungssystem $A\vec{x} = \vec{b}$ unlösbar (und somit $M = \emptyset$), oder es treten in dem System $A\vec{x} = \vec{b}$ linear abhängige Gleichungen auf. Da diese lediglich redundante Informationen enthalten, kann man sie ohne Beschränkung der Allgemeinheit entfernen. Die Rangbedingung (R) ist stets dann erfüllt, wenn die Matrix A wie in den Beispielen (15.1.1) und (15.1.2) durch Einführung von m Schlupfvariablen zu einer Matrix $(A, \pm Id)$, $Id \in \mathbf{R}^{(m,m)}$, erweitert wird. □

Mit der folgenden Definition wird eine der Grundlagen für das Simplexverfahren geschaffen.

Definition 15.6 *Es seien* $A = (\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n) \in \mathbf{R}^{(m,n)}$ *mit* $\operatorname{Rang} A = m$ *sowie* $\vec{b} \in \mathbf{R}^m$ *gegeben. Es bezeichne* $B \subset \{1, 2, \ldots, n\}$, $\operatorname{card} B = m$, *die Menge derjenigen Indizes, die zu einem* m*–dimensionalen System linear unabhängiger Spaltenvektoren* $\{\vec{a}_j : j \in B\}$ *gehören. Ein Vektor* $\vec{x} \in \mathbf{R}^n$ *heiße* **Basislösung von** $A\vec{x} = \vec{b}$ **zur Basis** B, *wenn gilt:*

$$x_j = 0 \ \forall\, j \notin B \quad \text{und} \quad \sum_{j\in B} x_j \vec{a}_j = \vec{b}.$$

Eine Basislösung $\vec{x}$ *zur Basis* B *heiße* **zulässig**, *wenn gilt:* $\vec{x} \geq \vec{0}$. *Eine zulässige Basislösung* $\vec{x}$ *zur Basis* B *heiße* **nichtentartet**, *wenn gilt:* $\vec{x}_j > 0 \ \forall\, j \in B$. *Andernfalls heiße sie* **entartet**.

Mit dem folgenden Satz wird eine Verbindung aufgezeigt zwischen dem geometrischen Begriff einer **Ecke** und dem algebraischen Begriff einer zulässigen **Basislösung**.

Satz 15.7 *Die Menge* $M := \{\vec{x} \in \mathbf{R}^n : \vec{x} \geq \vec{0},\ A\vec{x} = \vec{b}\}$ *sei nichtleer, und es gelte* $A = (\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n) \in \mathbf{R}^{(m,n)}$ *mit* $\operatorname{Rang} A = m$. *Genau dann ist* $\vec{x} \in M$ *eine Ecke von* M, *wenn* $\vec{x}$ *zulässige Basislösung von* $A\vec{x} = \vec{b}$ *zu einer geeigneten Basis* $B \subset \{1, 2, \ldots, n\}$ *ist.*

Begründung: (i) Es sei $\vec{x} \in M$ eine Ecke von M. Dann sind die zu positiven Komponenten $x_j > 0$ gehörenden Spaltenvektoren $\vec{a}_j$ gemäß Satz 15.4(a) linear unabhängig. Sind es weniger als m Spaltenvektoren, so kann ihre Anzahl durch Hinzunahme weiterer Spaltenvektoren zu einer m–dimensionalen Basis B so ergänzt werden, dass $\operatorname{span}\{\vec{a}_j : j \in B\} = \operatorname{Bild} A$ gilt.

(ii) Ist umgekehrt $\vec{x} \in M$ eine zulässige Basislösung von $A\vec{x} = \vec{b}$ zur Basis B, so ist das Vektorsystem $\{\vec{a}_j : j \in B\}$ linear unabhängig. Gemäß Satz 15.4 ist dann $\vec{x}$ eine Ecke von M. □

Bemerkung 15.7 Zu einer nichtentarteten zulässigen Basislösung $\vec{x}$ von $A\vec{x} = \vec{b}$ gehört genau eine Basis B, nämlich die Menge derjenigen Indizes $j \in \{1, 2, \ldots, n\}$ mit $x_j > 0$.

Eine entartete zulässige Basislösung $\vec{x}$ von $A\vec{x} = \vec{b}$ mit $p < m$ positiven Komponenten kann hingegen mehrere verschiedene Basen B besitzen, nämlich genau $\binom{m-p}{n-m}$ Stück! □

15.2.2 Die Zweiphasenmethode: Phase II

Wir betrachten wieder die **Normalform** eines linearen Programmes

$$\boxed{\text{Minimiere } \langle \vec{c}, \vec{x} \rangle \text{ auf der Menge } M := \{\vec{x} \in \mathbf{R}^n \,:\, \vec{x} \geq \vec{0},\ A\vec{x} = \vec{b}\},} \tag{P}$$

worin $A \in \mathbf{R}^{(m,n)}, \vec{b} \in \mathbf{R}^m$ und $\vec{c} \in \mathbf{R}^n$ die Vorgaben sind und wobei die Rangbedingung

$$\boxed{\operatorname{Rang} A = m} \tag{R}$$

erfüllt sei. In der **Phase II** des Simplexverfahrens werde vorausgesetzt, eine Ecke $\vec{x}$ des Polyeders M sei bekannt, oder gleichwertig (vgl. Satz 15.7), eine zulässige Basislösung $\vec{x}$ von $A\vec{x} = \vec{b}$ zu einer geeigneten Basis B sei gegeben.

In der **Phase I** werden dagegen Widersprüche in den Nebenbedingungen oder der Fall $M = \emptyset$ aufgedeckt; es wird die Rangbedingung (R) überprüft, gegebenenfalls werden redundante Gleichungen entfernt, und es wird eine zulässige Ausgangsbasislösung bestimmt. In der Phase I wird mit Hilfe der Phase II ein Teilproblem gelöst. Wir beginnen deshalb zunächst mit der Beschreibung der Phase II und vereinbaren zu diesem Zweck die folgenden *Bezeichnungen:*

- Eine Basis B der Länge m wird mit $B := \{p(1), p(2), \ldots, p(m)\} \subset \{1, 2, \ldots, n\}$ bezeichnet, und es gelte $N := \{1, 2, \ldots, n\} \setminus B$.
- Für $A \in \mathbf{R}^{(m,n)}$ setzen wir $A_B := (\vec{a}_{p(1)}, \vec{a}_{p(2)}, \ldots, \vec{a}_{p(m)}) \in \mathbf{R}^{(m,m)}$ sowie $A_N := (\vec{a}_k)_{k \in N} \in \mathbf{R}^{(m,n-m)}$.
- Für $\vec{z} \in \mathbf{R}^n$ setzen wir $\vec{z}_B := (z_{p(1)}, z_{p(2)}, \ldots, z_{p(m)})^T \in \mathbf{R}^m$ sowie $\vec{z}_N := (z_k)_{k \in N}^T \in \mathbf{R}^{n-m}$.

Die Phase II des Simplexverfahrens besteht nun aus geometrischer Sicht in den folgenden Schritten:

- **Step 1:** Sei $\vec{x}$ eine Ecke des Polyeders M.
- **Step 2:** Bestimme eine von $\vec{x}$ ausgehende **Abstiegskante**. Das ist eine **Kante** des Polyeders M, längs der die Zielfunktion $\langle \vec{c}, \vec{x} \rangle$ echt abnimmt. Gibt es keine solche Abstiegskante, so ist $\vec{x}$ bereits Lösung von (P). **Stop!**
- **Step 3:** Prüfe, ob die gefundene Abstiegskante **unbeschränkt** ist. Wenn ja, so ist die Zielfunktion $\langle \vec{c}, \vec{x} \rangle$ nach unten **nicht** beschränkt, und somit ist (P) unlösbar. **Stop!**
- **Step 4:** Die gefundene Abstiegskante ist beschränkt. Die Zielfunktion $\langle \vec{c}, \vec{x} \rangle$ nimmt auf dieser Abstiegskante ihr Minimum im Endpunkt $\vec{x}^+$ an. Dieser ist eine Ecke des Polyeders M, und somit definiert er eine neue Näherung.
- **Step 5:** Setze $\vec{x} := \vec{x}^+$ und gehe zurück nach **Step 2.**

Die analytische Umsetzung des hier beschriebenen geometrischen Sachverhalts erfolgt in dem folgenden Satz:

Satz 15.8 *Für das lineare Programm in der Normalform*

$$\boxed{\textit{Minimiere } \langle \vec{c}, \vec{x} \rangle \textit{ auf der Menge } M := \{\vec{x} \in \mathbf{R}^n \,:\, \vec{x} \geq \vec{0},\ A\vec{x} = \vec{b}\},} \tag{P}$$

gelte $A = (\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n) \in \mathbf{R}^{(m,n)}, \vec{b} \in \mathbf{R}^m, \vec{c} \in \mathbf{R}^n$ *sowie* $\operatorname{Rang} A = m$. *Es sei* $\vec{x} \in M$ *eine zulässige Basislösung zur Basis* $B := \{p(1), p(2), \ldots, p(m)\}$. *Das heißt, der Basisanteil von* $\vec{x}$ *ist* $\vec{x}_B := A_B^{-1}\vec{b}$ *mit zugehörigen Kosten* $z_0 := \langle \vec{c}_B, \vec{x}_B \rangle$. *Wir setzen* $\vec{y} := (A_B^{-1})^T \vec{c}_B$. *Dann gilt:*

(a) *Ist* $\vec{c}_N - A_N^T\vec{y} \geq \vec{0}$, *so ist* $\vec{x}$ *eine Lösung von* (P). *Diese ist eindeutig, wenn die strikte Ungleichung* $\vec{c}_N - A_N^T\vec{y} > \vec{0}$ *vorliegt.*

(b) *Gilt für einen Index* $k \in N$ *aber* $c_k - \langle \vec{a}_k, \vec{y}\rangle < 0$ *und* $\vec{w} := A_B^{-1}\vec{a}_k \leq \vec{0}$, *so folgt* $\inf(P) = -\infty$. *Das heißt, die Zielfunktion* $\langle \vec{c}, \vec{x}\rangle$ *ist auf* M *nach unten* **nicht** *beschränkt.*

(c) *Gelte für einen Index* $k \in N$ *wieder* $c_k - \langle \vec{a}_k, \vec{y}\rangle < 0$, *jedoch* $w_r := (A_B^{-1}\vec{a}_k)_r > 0$ *für eine Komponente* $r \in \{1, 2, \ldots, m\}$. *Dabei sei* r *bereits so bestimmt, dass gilt:*

$$\frac{(A_B^{-1}\vec{b})_r}{w_r} = \min_{i=1,\ldots,m}\Big\{\frac{(A_B^{-1}\vec{b})_i}{w_i} \,:\, w_i > 0\Big\} =: t^* \geq 0. \tag{2.5}$$

Definiert man $\vec{x}^+ \in \mathbf{R}^n$ *gemäß*

$$x_j^+ := \begin{cases} (A_B^{-1}\vec{b})_i - t^* w_i & \text{für} \quad j = p(i) \in B, \\ t^* & \text{für} \quad j = k, \\ 0 & \text{für} \quad j \neq k,\ j \in N, \end{cases} \tag{2.6}$$

so ist $\vec{x}^+$ *eine zulässige Basislösung zur Basis*

$$B^+ := \{p(1), \ldots, p(r-1), k, p(r+1), \ldots, p(m)\} =: \{p^+(1), p^+(2), \ldots, p^+(m)\},$$

mit den Kosten

$$z_0^+ := \langle \vec{c}, \vec{x}^+\rangle = \langle \vec{c}, \vec{x}\rangle + \frac{(A_B^{-1}\vec{b})_r}{w_r}\,(c_k - \langle \vec{a}_k, \vec{y}\rangle) \leq \langle \vec{c}, \vec{x}\rangle = z_0. \tag{2.7}$$

Insbesondere gilt im Falle einer **nichtentarteten** *Basislösung* $\vec{x}$ *zur Basis* B: $\vec{x}^+$ *ist zulässige Basislösung zur Basis* B^+ *mit echter Kostenminderung* $\langle \vec{c}, \vec{x}^+\rangle < \langle \vec{c}, \vec{x}\rangle$. *Ferner gilt:*

$$A_{B^+}^{-1} = \Big(Id_m - \frac{1}{w_r}\,(\vec{w} - \vec{e}_r) \otimes \vec{e}_r\Big) A_B^{-1}. \tag{2.8}$$

Begründungen: (a) Es sei $\vec{c}_N - A_N^T\vec{y} \geq \vec{0}$. Für ein beliebiges $\vec{z} \in M$ gilt $A\vec{z} = A_B\vec{z}_B + A_N\vec{z}_N = \vec{b}$, und somit $\vec{z}_B = \vec{x}_B - A_B^{-1}A_N\vec{z}_N$. Hieraus resultiert

$$\begin{aligned} \langle \vec{c}, \vec{z}\rangle &= \langle \vec{c}_B, \vec{z}_B\rangle + \langle \vec{c}_N, \vec{z}_N\rangle = \langle \vec{c}, \vec{x}\rangle - \langle \vec{c}_B, A_B^{-1}A_N\vec{z}_N\rangle + \langle \vec{c}_N, \vec{z}_N\rangle \\ &= \langle \vec{c}, \vec{x}\rangle + \langle \vec{c}_N - A_N^T\vec{y}, \vec{z}_N\rangle \geq \langle \vec{c}, \vec{x}\rangle, \end{aligned}$$

also eine Lösung $\vec{x}$ von (P).

Eindeutigkeit: Wegen $\vec{c}_N - A_N^T\vec{y} > \vec{0}$ kann $\langle \vec{c}, \vec{z}\rangle = \langle \vec{c}, \vec{x}\rangle$ nur gelten, wenn $\vec{z}_N = \vec{0}$ ist. Dann folgt $A\vec{z} = A_B\vec{z}_B = \vec{b}$, also $\vec{z}_B = A_B^{-1}\vec{b} = \vec{x}_B$, und somit $\vec{z} = \vec{x}$.

(b) Gelte nun $c_k - \langle \vec{a}_k, \vec{y}\rangle < 0$ und $\vec{w} := A_B^{-1}\vec{a}_k \leq \vec{0}$ für einen Index $k \in N$. Wir definieren $\vec{x}(t) \in \mathbf{R}^n$, $t \geq 0$, gemäß

$$x_j(t) := \begin{cases} (A_B^{-1}\vec{b})_i - t w_i & \text{für} \quad j = p(i) \in B, \\ t & \text{für} \quad j = k, \\ 0 & \text{für} \quad j \neq k,\ j \in N. \end{cases} \tag{2.9}$$

Es gilt $\vec{x}(t) \geq \vec{0}$ sowie

$$A\vec{x}(t) = A_B(\vec{x}_B - tA_B^{-1}\vec{a}_k) + t\vec{a}_k = A_B\vec{x}_B = \vec{b},$$

also $\vec{x}(t) \in M\ \forall\, t \geq 0$. Ferner folgt

$$\langle \vec{c}, \vec{x}(t)\rangle = \langle \vec{c}, \vec{x}_B - tA_B^{-1}\vec{a}_k\rangle + tc_k = \langle \vec{c}, \vec{x}\rangle + t\underbrace{(c_k - \langle \vec{a}_k, \vec{y}\rangle)}_{<0} \to -\infty \quad (t \to +\infty).$$

Das heißt, die Zielfunktion $\langle \vec{c}, \vec{x} \rangle$ ist auf M nach unten **nicht** beschränkt.

(c) Es sei nun $t^* \geq 0$ gemäß (2.5) bestimmt. Wegen (2.6) und (2.7) gilt $\vec{x}^+ = \vec{x}(t^*) \geq \vec{0}$ sowie $x^+_{p(r)} = 0$ nach Konstruktion von r. Ferner ist $A\vec{x}^+ = \vec{b}$ und

$$z_0^+ = \langle \vec{c}, \vec{x}^+ \rangle = \langle \vec{c}, \vec{x} \rangle + t^*(c_k - \langle \vec{a}_k, \vec{y} \rangle) \leq \langle \vec{c}, \vec{x} \rangle = z_0.$$

Die Basis B^+ entsteht aus der Basis B durch Austausch von $p(r)$ gegen k. Deshalb folgt:

$$\begin{aligned} A_{B^+} &= (\vec{a}_{p(1)}, \ldots, \vec{a}_{p(r-1)}, \vec{a}_k, \vec{a}_{p(r+1)}, \ldots, \vec{a}_{p(m)}) = A_B + (\vec{a}_k - \vec{a}_{p(r)}) \otimes \vec{e}_r \\ &= A_B \Big(Id_m + (A_B^{-1}\vec{a}_k - A_B^{-1}A_B\vec{e}_r) \otimes \vec{e}_r \Big) = A_B \Big(Id_m + (\vec{w} - \vec{e}_r) \otimes \vec{e}_r \Big). \end{aligned}$$

Mit Hilfe des Ansatzes $C^{-1} := Id_m - \alpha(\vec{w} - \vec{e}_r) \otimes \vec{e}_r$ bestimmen wir die Inverse von $C := Id_m + (\vec{w} - \vec{e}_r) \otimes \vec{e}_r$. Es muss gelten:

$$\begin{aligned} Id_m &= CC^{-1} = \Big(Id_m + (\vec{w} - \vec{e}_r) \otimes \vec{e}_r \Big) \Big(Id_m - \alpha(\vec{w} - \vec{e}_r) \otimes \vec{e}_r \Big) \\ &= Id_m + (1-\alpha)(\vec{w} - \vec{e}_r) \otimes \vec{e}_r - \alpha(\vec{w} - \vec{e}_r) \underbrace{\vec{e}_r^T(\vec{w} - \vec{e}_r)}_{=w_r - 1} \vec{e}_r^T \\ &= Id_m + (1 - \alpha w_r)(\vec{w} - \vec{e}_r) \otimes \vec{e}_r = C^{-1}C. \end{aligned}$$

Diese Identität ist genau für $1 - \alpha w_r = 0$ erfüllt. Wegen $w_r > 0$ existiert somit

$$A_{B^+}^{-1} = \Big(Id_m - \frac{1}{w_r} (\vec{w} - \vec{e}_r) \otimes \vec{e}_r \Big) A_B^{-1},$$

und es gilt $\operatorname{Rang} A_{B^+} = m$. Schließlich folgt $x_j^+ = 0 \ \forall \ j \notin B^+$ sowie $\vec{x}^+ \in M$, so dass $\vec{x}^+$ eine zulässige Basislösung zur Basis B^+ ist. Damit ist der Satz vollständig bewiesen. □

Die Ergebnisse dieses Satzes geben nun Anlass, die analytischen Detailschritte zum **revidierten Simplexverfahren** zusammenzufassen:

- **Step 1:** Sei $\vec{x} \in M$ zulässige Basislösung zur Basis $B = \{p(1), p(2), \ldots, p(m)\}$. Der Basisanteil von $\vec{x}$ ist $\vec{x}_B := A_B^{-1}\vec{b}$, und die zugeordneten Kosten sind $z_0 := \langle \vec{c}_B, \vec{x}_B \rangle$. Es bezeichne $N := \{1, 2, \ldots, n\} \setminus B$ die Menge der Nichtbasisindizes.
- **Step 2:** Berechne $\vec{y} := (A_B^{-1})^T \vec{c}_B$ und danach $\tilde{c}_j := c_j - \langle \vec{a}_j, \vec{y} \rangle$ für $j \in N$.
- **Step 3:** Falls $\tilde{c}_j \geq 0 \ \forall \ j \in N$, dann **Stop!** Der Vektor $\vec{x}$ ist eine Lösung von (P).
- **Step 4:** Wähle $k \in N$ mit $\tilde{c}_k < 0$. (Oft wählt man $\tilde{c}_k := \min\limits_{j \in N} \{\tilde{c}_j \ : \ \tilde{c}_j < 0\}$, obwohl diese Wahl nicht zwingend ist und auch nicht unbedingt den maximalen Abfall der Zielfunktion bewirkt.)
- **Step 5:** Berechne $\vec{w} := A_B^{-1}\vec{a}_k$. Falls $\vec{w} \leq \vec{0}$, dann **Stop!** Die Zielfunktion ist auf der Menge M nach unten **nicht** beschränkt; es gilt $\inf(P) = -\infty$.
- **Step 6:** Bestimme $r \in \{1, 2, \ldots, m\}$ mit $w_r > 0$ so, dass gilt:

$$\frac{(A_B^{-1}\vec{b})_r}{w_r} = \min_{i=1,\ldots,m} \Big\{ \frac{(A_B^{-1}\vec{b})_i}{w_i} \ : \ w_i > 0 \Big\} =: t^* \geq 0.$$

- **Step 7:** Setze $B^+ := \{p(1), \ldots, p(r-1), k, p(r+1), \ldots, p(m)\}$, $N^+ := \{1, 2, \ldots, n\} \setminus B^+$ und berechne anschließend den Basisanteil $\vec{x}^+_{B^+}$ der neuen zulässigen Basislösung $\vec{x}^+$ durch

$$x_j^+ := \begin{cases} (A_B^{-1}\vec{b})_i - t^* w_i & \text{für} \quad j = p(i), \ i \neq r, \\ t^* & \text{für} \quad j = k. \end{cases}$$

  Dann ist

$$\vec{x}^+_{B^+} = A_{B^+}^{-1}\vec{b} \quad \text{mit} \quad A_{B^+}^{-1} := \Big(Id_m - \frac{1}{w_r} (\vec{w} - \vec{e}_r) \otimes \vec{e}_r \Big) A_B^{-1},$$

  und die zugeordneten Kosten betragen $z_0^+ := z_0 + t^* \tilde{c}_k$.

- **Step 8:** Vertausche $p(r)$ und k, setze $(\vec{x}, B, N, z_0) := (\vec{x}^+, B^+, N^+, z_0^+)$ und gehe nach **Step 2**.

Bemerkung 15.8 Offenbar besteht der Hauptaufwand beim revidierten Simplexverfahren in der Berechnung von $\vec{y} := (A_B^{-1})^T \vec{c}_B \in \mathbf{R}^m$ und $\vec{w} := A_B^{-1} \vec{a}_k \in \mathbf{R}^m$. Wegen (2.8) braucht A_B^{-1} allerdings nicht in jedem Schritt neu berechnet zu werden, da sich A_B und A_{B^+} nur in einer Spalte unterscheiden. Man gewinnt deshalb $A_{B^+}^{-1}$ aus A_B^{-1} durch Multiplikation mit der GAUSS–JORDAN–Matrix

$$E := Id_m - \frac{1}{w_r} (\vec{w} - \vec{e}_r) \otimes \vec{e}_r$$

in zwei einfachen Schritten:

- Es gilt

$$\vec{e}_r^T A_{B^+}^{-1} = \Big(\vec{e}_r^T - \frac{1}{w_r} \underbrace{\vec{e}_r^T (\vec{w} - \vec{e}_r)}_{= w_r - 1} \vec{e}_r^T\Big) A_B^{-1} = \frac{1}{w_r} \vec{e}_r^T A_B^{-1}.$$

  Das heißt:

Dividiere den r–ten Zeilenvektor von A_B^{-1} durch w_r.

  Als Resultat erhält man den r–ten Zeilenvektor der neuen Matrix $A_{B^+}^{-1}$.

- Für $i = 1, 2, \ldots, m$, $i \neq r$, gilt

$$\vec{e}_i^T A_{B^+}^{-1} = \Big(\vec{e}_i^T - \frac{1}{w_r} \underbrace{\vec{e}_i^T (\vec{w} - \vec{e}_r)}_{= w_i} \vec{e}_i^T\Big) A_B^{-1} = \vec{e}_i^T A_B^{-1} - w_i \vec{e}_r^T A_{B^+}^{-1}.$$

  Das heißt:

Man erhält den i–ten Zeilenvektor $\vec{z}_i^+$ der neuen Matrix $A_{B^+}^{-1}$ aus dem i–ten Zeilenvektor $\vec{z}_i$ der alten Matrix A_B^{-1} durch die Operation $$\vec{z}_i^+ := \vec{z}_i - w_i \vec{z}_r^+, \quad i = 1, 2, \ldots, m,\ i \neq r.$$

Hat man für die Ausgangsbasis B_0 die Inverse $A_{B_0}^{-1}$ berechnet, so erhält man nun im k–ten Schritt die Inverse $A_{B_k}^{-1}$ in der Form

$$A_{B_k}^{-1} = E_k E_{k-1} \cdots E_1 A_{B_0}^{-1}$$

mit GAUSS–JORDAN–Matrizen $E_1, E_2, \ldots, E_k$. Häufig wird $A_{B_0} = Id_m$ die Einheitsmatrix sein. Dies ist stets dann der Fall, wenn durch Einführung von m Schlupfvariablen die Gleichungsrestriktionen in der Form $(A,\ Id_m)\vec{x} = \vec{b}$ vorliegen. □

BSP. (15.2.3) Wir betrachten die Optimierungsaufgabe aus BSP. (15.1.1), indem wir die Daten aus (1.3) in ein Tableau der Form

$\vec{c}^T$	
A	$\vec{b}$

$=$

-1	-1	0	0	0	
1	3	1	0	0	13
3	1	0	1	0	15
-1	1	0	0	1	3

bringen. Offensichtlich ist hier die Rangbedingung (R), nämlich $\operatorname{Rang} A = 3$, erfüllt. Wir starten mit der Basis $B_0 = \{3, 4, 5\}$, also mit $A_{B_0} = A_{B_0}^{-1} = Id_3$. Wir bringen die Startdaten des revidierten

Simplexverfahrens in ein Tableau der Form

k	$\vec{y}^T$	z_0
B	A_B^{-1}	$\vec{x}_B$

$=$

1	0	0	0	0
3	1	0	0	13
$\boxed{4}$	0	1	0	15
5	0	0	1	3

$\left.\begin{array}{l}\tilde{c}_1 = -1\\ \tilde{c}_2 = -1\end{array}\right\} \Rightarrow$ *Step 4:* Wähle $k = 1$;

Step 5: $\vec{w} = (1, 3, -1)^T$;

Step 6: $r = 2,\ t^* = \frac{15}{3}$;

Step 7: $B^+ = \{3, 1, 5\}$.

Neues Tableau mit $w_r = 3$:

2	0	$-\frac{1}{3}$	0	-5
$\boxed{3}$	1	$-\frac{1}{3}$	0	8
1	0	$\frac{1}{3}$	0	5
5	0	$\frac{1}{3}$	1	8

$\left.\begin{array}{l}\tilde{c}_2 = -\frac{2}{3}\\ \tilde{c}_4 = \frac{1}{3}\end{array}\right\} \Rightarrow$ *Step 4:* Wähle $k = 2$;

Step 5: $\vec{w} = (\frac{8}{3}, \frac{1}{3}, \frac{4}{3})^T$; *Step 6:* $r = 1,\ t^* = 3$;

Step 7: $B^+ = \{2, 1, 5\}$.

Neues Tableau mit $w_r = \frac{8}{3}$:

	$-\frac{1}{4}$	$-\frac{1}{4}$	0	-7
2	$\frac{3}{8}$	$-\frac{1}{8}$	0	3
1	$-\frac{1}{8}$	$\frac{3}{8}$	0	4
5	$-\frac{1}{2}$	$\frac{1}{2}$	1	4

$\left.\begin{array}{l}\tilde{c}_3 = \frac{1}{4} > 0\\ \tilde{c}_4 = \frac{1}{4} > 0\end{array}\right\} \Rightarrow$ **Stop!**

Wir haben eine Lösung des linearen Programms bestimmt, nämlich

$$\boxed{\vec{x} = (4, 3, 0, 0, 4)^T, \quad z_{\min} = \langle \vec{c}, \vec{x} \rangle = -7.}$$

BSP. (15.2.4) Wir behandeln BSP. (15.1.2) in ganz analoger Weise. Zuerst bringen wir die Daten aus (1.4) in das Tableau

$\vec{c}^T$	
A	$\vec{b}$

$=$

1	3	0	0	0	0	
12	9	-1	0	0	0	108
6	15	0	-1	0	0	90
0	1	0	0	-1	0	2
1	0	0	0	0	-1	3

Wegen $\operatorname{Rang} A = 4$ ist die Rangbedingung (R) erfüllt. Wir dürfen hier nicht mit der Basis $B_0 = \{3, 4, 5, 6\}$ starten, also mit $A_B = A_B^{-1} = -\mathit{Id}_4$. Diese würde die folgenden Startdaten des revidierten Simplexverfahrens liefern:

k	$\vec{y}^T$	z_0
B	A_B^{-1}	$\vec{x}_B$

$=$

	0	0	0	0	0
3	-1	0	0	0	-108
4	0	-1	0	0	-90
5	0	0	-1	0	-2
6	0	0	0	-1	-3

Es wäre also insbesondere $\vec{x} = (0, 0, -108, -90, -2, -3)^T \leq \vec{0}$ keine zulässige Basislösung. Wir starten stattdessen mit der Basis $B_0 = \{1, 2, 5, 6\}$ und erhalten die folgenden Daten des revidierten

Simplexverfahrens:

3	$-\frac{1}{42}$	$\frac{9}{42}$	0	0	$\frac{117}{7}$
1	$\frac{5}{42}$	$-\frac{1}{14}$	0	0	$\frac{45}{7}$
2	$-\frac{2}{42}$	$\frac{4}{42}$	0	0	$\frac{24}{7}$
$\boxed{5}$	$-\frac{2}{42}$	$\frac{4}{42}$	-1	0	$\frac{10}{7}$
6	$\frac{5}{42}$	$-\frac{1}{14}$	0	-1	$\frac{24}{7}$

$\left.\begin{array}{l}\tilde{c}_3 = -\frac{1}{42} \\ \tilde{c}_4 = \frac{9}{42}\end{array}\right\} \Rightarrow$ *Step 4:* Wähle $k = 3$;

Step 5: $\vec{w} = (-\frac{5}{42}, \frac{2}{42}, \frac{2}{42}, -\frac{5}{42})^T$; *Step 6:* $r = 3,\ t^* = 30$;

Step 7: $B^+ = \{1, 2, 3, 6\}$.

Neues Tableau mit $w_r = \frac{2}{42}$:

	0	$\frac{7}{42}$	$\frac{1}{2}$	0	16
1	0	$\frac{7}{42}$	$-\frac{5}{2}$	0	10
2	0	0	1	0	2
3	-1	2	-21	0	30
6	0	$\frac{7}{42}$	$-\frac{5}{2}$	-1	7

$\left.\begin{array}{l}\tilde{c}_4 = \frac{7}{42} > 0 \\ \tilde{c}_5 = \frac{1}{2} > 0\end{array}\right\} \Rightarrow$ **Stop!**

Wir haben eine Lösung des linearen Programms bestimmt, nämlich

$$\boxed{\vec{x} = (10, 2, 30, 0, 0, 7)^T, \quad z_{\min} = \langle \vec{c}, \vec{x} \rangle = 16.}$$

Für eine **Implementierung** der Phase II des revidierten Simplexverfahrens, welche sich an den Formeln der oben beschriebenenen Schritte *Step 1* bis *Step 8* orientiert, schlagen wir den Algorithmus (2.10) auf der nächsten Seite vor. In diesem Algorithmus gehen wir von der Voraussetzung aus, dass die Matrix $A \in \mathbf{R}^{(m,n)}$ die Rangbedingung $\operatorname{Rang} A = m$ erfüllt und dass die Ausgangsbasis B einer zulässigen Basislösung bereits bekannt ist. Das Einlesen der Ausgangsbasis erfolgt durch Angabe der Basisindizes $p(1), p(2), \ldots, p(m)$ derjenigen Spaltenvektoren in $A = (\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n)$, die die zulässige Basislösung bestimmen. Die Nichtbasisindizes $p(m+1), \ldots, p(n)$ werden durch das Programm bestimmt. Für die Invertierung der Matrix A_B zur Ausgangsbasis B wird das **Austauschverfahren** vorgeschlagen.

Erläuterungen zum Simplexalgorithmus (2.10): Der Vektor $\vec{x} = (x_1, x_2, \ldots, x_n)^T$ gibt nach Abschluss aller Rechnungen eine optimale Lösung an, die das Zielfunktional $z_0 = \langle \vec{c}, \vec{x} \rangle$ minimal macht. Die Variable z_0 gibt dieses Minimum an. Über mehrdeutige Lösbarkeit erteilt das Programm keine Auskunft. In Zeile 1–4 werden die Spaltenvektoren der Ausgangsbasis zur Matrix $D = (d_{jk}) \in \mathbf{R}^{(m,m)}$ zusammengefasst, und die Matrix D wird invertiert. D wird mit der Inversen überschrieben, so dass nun $D = A_B^{-1}$ gilt. In Zeile 6–11 wird die Menge $N = \{p(m+1), p(m+2), \ldots, p(n)\}$ der Nichtbasisindizes bestimmt, und es wird $x_j = 0$ für $j \in N$ gesetzt. In den Zeilen 12–16 werden der Basisanteil $\vec{x}_B$ der Ausgangslösung und die zugeordneten Anfangskosten z_0 berechnet. In den Zeilen 18–28 werden der Hilfsvektor $\vec{y}$, die Hilfsgrößen $\tilde{c}_j$ sowie das Minimum $\tilde{c}_k := \min_{j \in N} \tilde{c}_j, k := p(q)$, bestimmt, sofern dieses Minimum negativ ist. Andernfalls ist bereits eine optimale Lösung gefunden. In den Zeilen 29–35 wird der Hilfsvektor $\vec{w}$ berechnet. Seine positiven Komponenten werden in dem Indexvektor $(i(1), i(2), \ldots, i(l))$ registriert. Für $l = 0$ gibt es keine positiven Komponenten; die Zielfunktion ist auf M nach unten nicht beschränkt. In den Zeilen 36–39 werden der Index $r \in \{i(1), i(2), \ldots, i(l)\}$ und das Minimum $t^* := (A_B^{-1}\vec{b})_r / w_r$ bestimmt. Schließlich werden in den Zeilen 40, 41 und 45 die neue zulässige Basislösung $\vec{x}^+$ sowie die zugeordneten Kosten z_0^+ berechnet, während in den Zeilen 42–44 die neue Matrix $D^+ := A_{B^+}^{-1}$ gemäß den in Bemerkung 15.8 beschriebenen Zeilenoperationen

```
0:  Einlesen von n, m = Rang A; A = (a_jk) ∈ R^(m,n); b⃗ = (b_j)^T ∈ R^m; c⃗ = (c_j)^T ∈ R^n;
    Indexvektor p(j), j = 1,2,...,m, der Ausgangsbasis; maximale Iterationszahl N_max.
    Es wird D := A_B gesetzt.
1:  für j := 1,2,...,m :
2:    für k := 1,2,...,m :
3:      d_jk := a_jp(k) ; (Ende k, Ende j)
4:  Verwende AT–Verfahren zur Berechnung von A_B^-1 ⇌ (d_jk);
5:  it := 0; l := m+1; z_0 := 0;
6:  für j := 1,2,...,n :
7:    k := 1; bn := 0;
8:    wiederhole:
9:      falls (p(k) = j) dann bn := 1 sonst k := k+1; (Ende falls)
10:   bis (bn = 1) oder (k > m);
11:   falls (bn = 0) dann p(l) := j; x_j := 0; l := l+1; (Ende falls, Ende j)
12: für j := 1,2,...,m :
13:   s := 0;
14:   für k := 1,2,...,m :
15:     s := s + d_jk * b_k ; (Ende k)
16:   x_p(j) := s; z_0 := z_0 + s * c_p(j) ; (Ende j)
17: wiederhole:
18:   min := 0;
19:   für j := 1,2,...,m :
20:     y_j := 0;
21:     für k := 1,2,...,m :
22:       y_j := y_j + d_kj * c_p(k) ; (Ende k, Ende j)
23:   für j := m+1, m+2,...,n :
24:     s := c_p(j) ;
25:     für k := 1,2,...,m :
26:       s := s − a_kp(j) * y_k ; (Ende k)
27:     falls (s < min) dann q := j; min := s; (Ende falls, Ende j)
28:   falls (min = 0) dann Stop! (Lösung gefunden) (Ende falls)
29:   l := 0;
30:   für j := 1,2,...,m :
31:     w_j := 0;
32:     für k := 1,2,...,m :
33:       w_j := w_j + d_jk * a_kp(q) ; (Ende k)
34:     falls (w_j > 0) dann l := l+1; i(l) := j; (Ende falls, Ende j)
35:   falls (l = 0) dann Stop! (Zielfunktion unbeschränkt) (Ende falls)
36:   r := i(1); t := x_p(r)/w_r ;
37:   für k := 2,3,...,l :
38:     j := i(k); s := x_p(j)/w_j ;
39:     falls (s < t) dann t := s; r := j; (Ende falls, Ende k)
40:   für k := 1,2,...,m :
41:     falls (k ≠ r) dann x_p(k) := x_p(k) − t * w_k sonst x_p(j) := 0; (Ende falls)
42:     d_rk := d_rk/w_r ;
43:     für j := 1,2,...,m :
44:       falls (j ≠ r) dann d_jk := d_jk − w_j * d_rk ; (Ende falls, Ende j, Ende k)
45:   x_p(q) := t; z_0 := z_0 + t * min;
46:   j := p(r); p(r) := p(q); p(q) := j; it := it+1 ;
47: bis (it > N_max).
```
(2.10)

bestimmt wird. Die alte Basis B wird durch die neue Basis B^+ ersetzt. Der Zähler it zählt die durchlaufenen Basiswechsel. Das Verfahren wird abgebrochen, falls eine vorgegebene Schranke $N_{\max}$ überschritten wird. Eine Begründung für diese Vorsichtsmaßregel geben wir in der folgenden

Bemerkung 15.9 Ist $\vec{x}$ eine **nichtentartete** zulässige Basislösung zur Ausgangsbasis B, so wird in Zeile 39 eine echt positive Zahl $t^* > 0$ bestimmt, und die der neuen Basislösung $\vec{x}^+$ zugeordneten Kosten z_0^+ sind echt kleiner als z_0. In jedem weiteren Schritt des Simplexverfahrens werden die Kosten zumindest nicht vergrößert, so dass das Verfahren nicht zur Ausgangslösung $\vec{x}$ zurückkehren kann. Sind sogar alle im Ablauf des Simplexverfahrens berechneten zulässigen Basislösungen nichtentartet, so muss das Verfahren nach endlich vielen Schritten abbrechen, und zwar entweder mit einer zulässigen Basislösung oder mit der Information, dass die Zielfunktion nach unten nicht beschränkt ist.

Ist $\vec{x}$ hingegen eine **entartete** zulässige Basislösung zur Basis B, und gilt

$$\left\{ j \in \{i(1), i(2), \ldots, i(l)\} : (A_B^{-1}\vec{b})_j = 0 \right\} \neq \emptyset,$$

so ist das in Zeile 39 bestimmte Minimum $t^* = (A_B^{-1}\vec{b})_r / w_r = 0$. Es folgt $\vec{x}^+ = \vec{x}$. Das heißt, das Verfahren bleibt in der Ecke $\vec{x}$ stehen; lediglich die Basisdarstellung von $\vec{x}^+$ ist eine andere. Theoretisch kann es sogar vorkommen, dass in jedem weiteren Schritt die Ecke $\vec{x}$ stationär bleibt, während lediglich die Basis ausgetauscht wird – bis zur Rückkehr auf die Ausgangsbasis. In diesem Fall spricht man von einem **Zyklus** im Simplexverfahren; man vergleiche das nachfolgende Beispiel. □

BSP. (15.2.5) Die Daten eines linearen Programms (P) in der Normalform mögen auf das folgende Tableau führen:

$\vec{c}^T$	
A	$\vec{b}$

$=$

-0.75	20	-0.5	6	0	0	0	
0.25	-8	-1	9	1	0	0	0
0.5	-12	-0.5	3	0	1	0	0
0	0	1	0	0	0	1	1

Startet man den oben angegebenen Simplexalgorithmus mit der Ausgangsbasis $B_0 := \{5, 6, 7\}$ (die auf die *entartete* zulässige Basislösung $\vec{x} = (0, 0, 1)^T$ führt), so stellt sich der folgende Zyklus von Basisindizes ein:

Basisindizes der Basen B nach 10 Iterationen:

$it = 0$	$it = 1$	$it = 2$	$it = 3$	$it = 4$	$it = 5$	$it = 6$	$it = 7$	$it = 8$	$it = 9$	$it = 10$
5	1	1	3	3	5	5	1	1	3	3
6	6	2	2	4	4	6	6	2	2	4
7	7	7	7	7	7	7	7	7	7	7

Basisvektor $\vec{x}$ nach der letzten Iteration:

$$\vec{x} = \begin{bmatrix} 0.000\,000\mathrm{E}^{+00} \\ 0.000\,000\mathrm{E}^{+00} \\ 0.000\,000\mathrm{E}^{+00} \\ 0.000\,000\mathrm{E}^{+00} \\ 0.000\,000\mathrm{E}^{+00} \\ 0.000\,000\mathrm{E}^{+00} \\ 1.000\,000\mathrm{E}^{+00} \end{bmatrix}$$

Zugeordnete Kosten: $z_0 = 0.000\,000\mathrm{E}^{+00}$.

Der Simplexalgorithmus in der vorliegenden Form liefert **kein** Resultat.

Das Auftreten von Zyklen verhindert, dass der Simplexalgorithmus nach endlich vielen Schritten zu einem Resultat gelangt. Deshalb gibt es **Zusatzregeln** zur Vermeidung von Zyklen. Die einfachste

Zusatzregel stammt von R.G. BLAND ("New finite pivoting rules for the simplex method", Math. of Operations Research **2** (1977), 103–107).

> **Regel von BLAND.** Gibt es mehrere Möglichkeiten, den Index q (Zeile 27) und den Index r (Zeile 39) zu wählen, so wähle man q und r stets so, dass $p(q)$ und $p(r)$ kleinstmöglich sind.

Eine Implementierung der BLANDschen Zusatzregel würde folgende Änderungen des Simplexalgorithmus (2.10) erforderlich machen:

```
  ⋮
18:   min := 0; ind := n;
  ⋮
27:     falls (s < 0) und (p(j) ≤ ind) dann q := j; ind := p(j); min := s;  (Ende falls, j)
  ⋮
39:     falls (s ≤ t) dann:
39_1:     falls (s < t) oder (p(j) < p(r)) dann t := s; r := j;  (Ende falls, falls, k)
  ⋮
```

Mit dieser Zusatzregel erhalten wir nun für das obige BSP. (15.2.5) die folgende korrekte Lösung:

Basisindizes der Basen B nach 6 Iterationen:

$it=0$	$it=1$	$it=2$	$it=3$	$it=4$	$it=5$	$it=6$
5	1	1	3	3	3	3
6	6	2	2	4	4	5
7	7	7	7	7	1	1

Lösungsvektor $\vec{x}$:

$$\vec{x} = \begin{bmatrix} 1.000\,000\mathrm{E}^{+00} \\ 0.000\,000\mathrm{E}^{+00} \\ 1.000\,000\mathrm{E}^{+00} \\ 0.000\,000\mathrm{E}^{+00} \\ 7.500\,000\mathrm{E}^{-01} \\ 0.000\,000\mathrm{E}^{+00} \\ 0.000\,000\mathrm{E}^{+00} \end{bmatrix}$$

Minimum der Zielfunktion:

$$\boxed{z_{\min} = -1.250\,000\mathrm{E}^{+00}.}$$

15.2.3 Die Zweiphasenmethode: Phase I

Wie in Abschnitt 15.2.2 betrachten wir die **Normalform** eines linearen Programmes, nämlich

$$\boxed{\text{Minimiere } \langle \vec{c}, \vec{x} \rangle \text{ auf der Menge } M := \{\vec{x} \in \mathbf{R}^n \,:\, \vec{x} \geq \vec{0},\ A\vec{x} = \vec{b}\},} \qquad \text{(P)}$$

worin $A = (\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n) \in \mathbf{R}^{(m,n)}$, $\vec{b} = (b_1, b_2, \ldots, b_m)^T \in \mathbf{R}^m$ und $\vec{c} = (c_1, c_2, \ldots, c_n)^T \in \mathbf{R}^n$ die Vorgaben seien. Gleichungs–Restriktionen können notfalls mit -1 multipliziert werden, so dass ohne Beschränkung der Allgemeinheit

$$\boxed{\vec{b} \geq \vec{0}}$$

angenommen werden darf. In diesem Abschnitt sollen die folgenden Fragen beantwortet werden:

- Ist (P) zulässig, das heißt, gilt $M \neq \emptyset$?
- Hat A Maximalrang m, oder sind redundante Gleichungen zu entfernen?
- Wie berechnet man eine Ausgangsbasis B, d.h. eine zulässige Basislösung von (P), um die Phase II des Simplexalgorithmus starten zu können?

Wir werden ein geeignetes Hilfsprogramm zur Beantwortung dieser Fragen installieren. Wir unterscheiden:

Trivialer Fall: Die Matrix $A = (\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_n)$ enthalte in den Spalten bereits die m Einheitsvektoren $\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_m$ der Standardbasis des $\mathbf{R}^m$. Es sei B die Indexmenge der entsprechenden Spaltenvektoren. Dann kann sofort die Phase II des Simplexverfahrens mit der Ausgangsbasis B und der zulässigen Basislösung $\vec{x}_B := \vec{b}$ gestartet werden.

Nichttrivialer Fall: Die Matrix A enthalte keine oder nicht alle Standardbasisvektoren. Im ersten Fall führen wir **künstliche Variable** $(h_1, h_2, \ldots, h_m)^T$ in Form eines Hilfsvektors $\vec{h} \in \mathbf{R}^m$ ein und betrachten das folgende lineare Programm in Normalform:

$$\boxed{\text{Minimiere } \langle \vec{e}, \vec{h} \rangle \text{ auf der Menge } M_h := \left\{ \begin{bmatrix} \vec{x} \\ \vec{h} \end{bmatrix} \in \mathbf{R}^{n+m} \ : \ \begin{bmatrix} \vec{x} \\ \vec{h} \end{bmatrix} \geq \vec{0}, \ H \begin{bmatrix} \vec{x} \\ \vec{h} \end{bmatrix} = \vec{b} \right\},} \qquad (\mathrm{P}_h)$$

worin $\vec{e} := (1, 1, \ldots, 1)^T \in \mathbf{R}^m$ sowie $H := (A, Id_m) = (\vec{a}_1, \ldots, \vec{a}_n, \vec{e}_1, \ldots, \vec{e}_m) \in \mathbf{R}^{(m,n+m)}$ gelten. Mit der Ausgangsbasis $B_0 := \{n+1, n+2, \ldots, n+m\}$ kann die Phase II des Simplexalgorithmus gestartet werden. Wegen $\vec{e} \geq \vec{0}$ ist die Zielfunktion des Hilfsproblems (P_h) auf der Menge M_h nach unten durch 0 beschränkt. Außerdem gilt $M_h \neq \emptyset$, denn M_h enthält mindestens den Vektor $\begin{bmatrix} \vec{x} \\ \vec{h} \end{bmatrix} := \begin{bmatrix} \vec{0} \\ \vec{b} \end{bmatrix}$. Gemäß Satz 15.6 besitzt dann das Problem (P_h) eine zulässige Basislösung zu einer Basis $B = \{p(1), p(2), \ldots, p(m)\} \subset \{1, 2, \ldots, n+m\}$. Diese kann mit dem Simplexalgorithmus (2.10) berechnet werden. Die Basisindizes $p(j) \in \{n+1, n+2, \ldots, n+m\} \cap B$ nennen wir **künstliche** Basisindizes. Wir treffen zwei Fallunterscheidungen:

Fall 1: Es gilt $\boxed{\min\langle \vec{e}, \vec{h} \rangle > 0.}$ Das heißt, das Problem (P_h) besitzt **keine** zulässige Lösung in der Form $\begin{bmatrix} \vec{x} \\ \vec{h} \end{bmatrix} = \begin{bmatrix} \vec{x} \\ \vec{0} \end{bmatrix}$. Somit besitzt das Problem (P) auch keine zulässige Lösung $\vec{x}$. Man bricht den Simplexalgorithmus mit einer entsprechenden Meldung ab.

Fall 2: Es gilt $\boxed{\min\langle \vec{e}, \vec{h} \rangle = 0,}$ also $\vec{h} = \vec{0}$. Dann ist die Existenz einer zulässigen Basislösung $\vec{x}$ zum Problem (P) gesichert. Es sei $r \in \{1, 2, \ldots, m\}$ so bestimmt, dass gilt

$$p(r) = \max_{j=1,\ldots,m} p(j). \qquad (2.11)$$

Ist $\boxed{p(r) \leq n,}$ so haben wir mit $B = \{p(1), p(2), \ldots, p(m)\}$ bereits die erforderliche Ausgangsbasis zum Start der Phase II des Simplexalgorithmus (2.10) vorliegen. Darüber hinaus sind mit

$$\vec{x}_B = H_B^{-1}\vec{b} = A_B^{-1}\vec{b} \quad \text{und} \quad H_B^{-1} = A_B^{-1}$$

bereits der Basisanteil $\vec{x}_B$ der Startnäherung sowie die invertierte Basismatrix A_B^{-1} gegeben. (Das heißt, die Zeilen 1–16 im Algorithmus (2.10) brauchen nicht mehr durchlaufen zu werden. Lediglich die Anfangskosten $z_0 = \langle \vec{c}_B, \vec{x}_B \rangle$ sind zu berechnen.)

Gilt hingegen $\boxed{p(r) > n,}$ so enthält die Basis B mindestens einen künstlichen Basisindex, und die gewonnene Basislösung $\begin{bmatrix} \vec{x} \\ \vec{h} \end{bmatrix}$ ist notwendig entartet: alle Komponenten mit künstlichem Basisindex

$p(j) > n$ gehören zum Vektor $\vec{h} = \vec{0}$. Insbesondere gilt auch $(H_B^{-1}\vec{b})_r = 0$. Es soll nun versucht werden, den künstlichen Basisindex $p(r) > n$ gegen einen Index $k \in \{1, 2, \ldots, n\} \setminus B$ auszutauschen oder festzustellen, dass eine der Gleichungen in den Nebenbedingungen $A\vec{x} = \vec{b}$ von den übrigen Gleichungen linear abhängig ist. Dann gilt nämlich Rang $A \leq m-1$, und die entsprechende Gleichung ist redundant. Demgemäß treffen wir eine weitere Fallunterscheidung, nämlich:

(a) Es existiere ein Index $k \in \{1, 2, \ldots, n\} \setminus B$ mit $\boxed{(H_B^{-1}\vec{a}_k)_r \neq 0.}$ In diesem Fall setze man $\vec{w} := H_B^{-1}\vec{a}_k$ sowie $B^+ := \{p(1), \ldots, p(r-1), k, p(r+1), \ldots, p(m)\}$, und man berechne gemäß *Step 7:*

$$H_{B^+}^{-1} := \Big(Id_m - \frac{1}{w_r}(\vec{w} - \vec{e}_r) \otimes \vec{e}_r \Big) H_B^{-1}.$$

An der bereits gewonnenen Basislösung ändert sich nichts; es ist $(H_{B^+}^{-1}\vec{b})_r = (H_B^{-1}\vec{b})_r$, da lediglich eine Nullkomponente gegen eine Nullkomponente ausgetauscht wurde. Schließlich setze man $B := B^+$ und prüfe erneut, ob noch ein weiterer Index $r \in \{1, 2, \ldots, m\}$ existiert mit $p(r) > n$.

(b) Es gelte für alle $k \in \{1, 2, \ldots, n\} \setminus B$ die Bedingung $\boxed{(H_B^{-1}\vec{a}_k)_r = 0.}$ Klar, wegen $H_B^{-1}H_B = Id_m$ gilt $H_B^{-1}\vec{a}_j = \vec{e}_q$ für alle $j = p(q) \in \{1, 2, \ldots, n\} \cap B$, und somit $(H_B^{-1}\vec{a}_j)_r = 0$ (beachte: $q \neq r$). Insgesamt ist somit die r–te Zeile von $H_B^{-1}A$ eine Nullzeile:

$$\vec{e}_r^{\,T} H_B^{-1} A = \vec{0} \quad \Leftrightarrow \quad \vec{0}^{\,T} = A^T (H_B^{-1})^T \vec{e}_r.$$

An dieser Gleichung ändert sich nichts, wenn die Identität $Id_m = \sum_{j=1}^{m} (\vec{e}_j \otimes \vec{e}_j)$ eingefügt wird:

$$\vec{0}^{\,T} = A^T\, Id_m (H_B^{-1})^T \vec{e}_r = \sum_{j=1}^{m} A^T (\vec{e}_j \otimes \vec{e}_j)(H_B^{-1})^T \vec{e}_r = \sum_{j=1}^{m} \underbrace{\langle \vec{e}_j, (H_B^{-1})^T \vec{e}_r \rangle}_{=:\lambda_j} A^T \vec{e}_j = \sum_{j=1}^{m} \lambda_j A^T \vec{e}_j.$$

Das heißt, die Zeilen der Matrix A sind linear abhängig und somit folgt Rang $A \leq m-1$. Setzen wir $q := p(r) - n \in \{1, 2, \ldots, m\}$, so gilt nun für den Koeffizienten λ_q:

$$\lambda_q = \langle H_B^{-1}\vec{e}_q, \vec{e}_r \rangle = \langle \vec{e}_r, \vec{e}_r \rangle = 1, \quad \text{also} \quad A^T \vec{e}_q = -\sum_{\substack{j=1 \\ j \neq q}}^{m} \lambda_j A^T \vec{e}_j.$$

Daher ist die q–te Zeile in A linear abhängig von den übrigen Zeilen, und somit ist die q–te Gleichung in $A\vec{x} = \vec{b}$ redundant. Wir streichen diese Gleichung und darüber hinaus in der Basismatrix H_B^{-1} die r–te Zeile sowie im Basisanteil $H_B^{-1}\vec{b}$ die r–te Komponente (diese war Null). Anschließend wird

$$B := \{p(1), \ldots, p(r-1), p(r+1), \ldots, p(m)\}, \quad m := m - 1,$$

gesetzt. Nun wird erneut geprüft, ob noch ein weiterer Index $r \in \{1, 2, \ldots, m\}$ existiert mit $p(r) > n$.

Nach höchstens $m-1$ Schritten hat man auf diese Weise entweder $M = \emptyset$ gezeigt, oder man hat eine zulässige Basislösung $\vec{x}$ zur Basis B konstruiert und dazu die Basismatrix A_B^{-1} berechnet.

Bei entsprechender Modifizierung der voranstehenden Überlegungen kann auch der Fall miteinbezogen werden, wenn die Matrix A in den Spalten bereits einige Einheitsvektoren der Standardbasis des $\mathbf{R}^m$ enthält.

Auf den beiden nachfolgenden Seiten haben wir einen Algorithmus vorgeschlagen, in welchem beide Phasen I und II des revidierten Simplexverfahrens implementiert sind. In diesem Algorithmus wird nur die Normalform (P) eines linearen Programms vorausgesetzt. Der Vektor $\vec{b}$ der Gleichungs–Restriktionen $A\vec{x} = \vec{b}$ darf allerdings nicht negativ sei: $\vec{b} \geq \vec{0}$. An den Rang der Matrix $A \in \mathbf{R}^{(m,n)}$

```
0:    Einlesen von Spaltenzahl n_0, Zeilenzahl m; A = (a_jk) ∈ R^(m,n_0); b⃗ = (b_j)^T ∈ R^m; c⃗^(0) =
      (c_j^(0))^T ∈ R^(n_0); Anzahl m_0 bereits vorhandener Standardbasisvektoren (m_0 = 0: keine),
      Indexvektor p(j), j = 1,2,...,m_0, der zugeordneten Teilausgangsbasis; maximale Iteri-
      onszahl N_max; Maschinengenauigkeit δ. Es wird D := A_B^{-1} gesetzt.
1:    n := n_0; z_1 := 0; l := 1; init := 1; it := 0;
2:    für j := 1,2,...,n :
3:      c_j := 0;  (Ende j)
4:    für j := 1,2,...,m :
5:      i(j) := 0;
6:      für k := 1,2,...,m_0 :
7:        falls (a_jp(k) = 1) dann d_kj := 1; i(j) := k sonst d_kj := 0;  (Ende falls, k)
8:      für k := m_0 + 1, m_0 + 2,...,m :
9:        d_kj := 0; a_{j,n+k-m_0} := 0;  (Ende k)
10:     falls (i(j) = 0) dann
11:       k := m_0 + l; p(k) := l + n; l := l + 1;
12:       d_kj := 1; a_jp(k) := 1; c_p(k) := 1; i(j) := k;  (Ende falls)
13:     k := p(i(j)); x_k := b_j;
14:     falls (j > m_0) dann z_1 := z_1 + x_k;  (Ende falls, j)
15:   n := n_0 + m - m_0;
16: M1 : l := m + 1; z_0 := 0;
17:   für j := 1,2,...,n_0 :
18:     k := 1; bn := 0;
19:     wiederhole:
20:       falls (p(k) = j) dann bn := 1; z_0 := z_0 + x_j * c_j sonst k := k + 1;  (Ende falls)
21:     bis (bn = 1) oder (k > m_0);
22:     falls (bn = 0) dann p(l) := j; x_j := 0; l := l + 1;  (Ende falls, j)
23:   wiederhole:
24:     min := 0; ind := n;
25:     für j := 1,2,...,m :
26:       y_j := 0;
27:       für k := 1,2,...,m :
28:         y_j := y_j + d_kj * c_p(k);  (Ende k, j)
29:     für j := m + 1, m + 2,...,n :
30:       s := c_p(j);
31:       für k := 1,2,...,m :
32:         s := s - a_kp(j) * y_k;  (Ende k)
33:       falls (s < -δ) und (p(j) ≤ ind) dann ind := p(j); q := j; min := s;  (Ende falls, j)
34:     falls (min = 0) dann Fallstudie:
35:       init = 0 : Stop! (Lösung gefunden)
36:       init = 1 : falls (|z_1| ≤ δ) dann init := 0;
37:         für j := 1,2,...,n_0 :
38:           c_j := c_j^(0);  (Ende j)
39:         goto M2
40:       sonst Stop! (Es existiert keine zulässige Lösung) (Ende falls)
41:     sonst goto M4;  (Ende falls)
42: M2 :  r := 1; max := p(r);
43:     für k := 2,3,...,m :
44:       falls (p(k) > max) dann max := p(k); r := k;  (Ende falls, k)
45:     falls (max ≤ n_0) dann n := n_0; m_0 := m; goto M1;  (Ende falls)
46:     it := it + 1; q := 1; s := 0;
```

werden keine Voraussetzungen gestellt. Die linear unabhängigen Standardbasisvektoren in den Spalten der Matrix A können **optional** angegeben werden: ihre Anzahl m_0 ist einzulesen und ihre Spaltenindizes sind (in beliebiger Reihenfolge) durch die Parameter $p(1), p(2), \ldots, p(m_0)$, $m_0 \leq m$, festzulegen. Wird $m_0 = 0$ gesetzt, so werden keine Spaltenvektoren der Matrix A in die Ausgangsbasis

```
47:        wiederhole:
48:          bn := 1;
49:          für j := 1, 2, ..., m :
50:            falls (p(j) = q) dann bn := 0; (Ende falls, j)
51:          falls (bn = 1) dann s := 0;
52:            für k := 1, 2, ..., m :
53:              s := s + d_rk * a_kq; (Ende k, falls)
54:          q := q + 1;
55:        bis (|s| > δ) oder (q > n_0);
56:        falls (q > n_0) dann goto M3 sonst w_r := s;
57:          für k := 1, 2, ..., m :
58:            d_rk := d_rk / s; (Ende k)
59:          für j := 1, 2, ..., m :
60:            falls (j ≠ r) dann w_j := 0;
61:              für k := 1, 2, ..., m :
62:                w_j := w_j + d_jk * a_{k,q-1}; (Ende k)
63:              für k := 1, 2, ..., m :
64:                d_jk := d_jk - w_j * d_rk; (Ende k, falls, j)
65:          p(r) := q - 1; goto M2; (Ende falls)
66:  M3 :  q := p(r) - n_0;
67:        für j := q + 1, q + 2, ..., m :
68:          b_{j-1} := b_j;
69:          für k := 1, 2, ..., n_0 :
70:            a_{j-1,k} := a_jk; (Ende k, j)
71:        für j := r + 1, r + 2, ..., m :
72:          x_{p(j-1)} := x_{p(j)}; p(j - 1) := p(j);
73:          für k := 1, 2, ..., m :
74:            d_{j-1,k} := d_jk; (Ende k, j)
75:        m := m - 1; goto M2;
76:  M4 :  l := 0;
77:        für j := 1, 2, ..., m :
78:          w_j := 0;
79:          für k := 1, 2, ..., m :
80:            w_j := w_j + d_jk * a_{kp(q)}; (Ende k)
81:          falls (w_j > δ) dann l := l + 1; i(l) := j; (Ende falls, j)
82:        falls (l = 0) dann Stop! (Zielfunktion unbeschränkt) (Ende falls)
83:        r := i(1); t := x_{p(r)} / w_r;
84:        für k := 2, 3, ..., l :
85:          j := i(k); s := x_{p(j)} / w_j;
86:          falls (s ≤ t) dann
87:            falls (s < t) oder (p(j) < p(r)) dann t := s; r := j; (Ende falls, falls, k)
88:        für k := 1, 2, ..., m :
89:          falls (k ≠ r) dann x_{p(k)} := x_{p(k)} - t * w_k sonst x_{p(k)} := 0; (Ende falls)
90:          d_rk := d_rk / w_r;
91:          für j := 1, 2, ..., m :
92:            falls (j ≠ r) dann d_jk := d_jk - w_j * d_rk; (Ende falls, j, k)
93:        x_{p(q)} := t;
94:        falls (init = 0) dann z_0 := z_0 + t * min sonst z_1 := 0;
95:          für j := n_0 + 1, n_0 + 2, ..., n :
96:            z_1 := z_1 + x_j; (Ende j, falls)
97:        j := p(r); p(r) := p(q); p(q) := j; it := it + 1;
98:      bis (it > N_max).
```

einbezogen. Die fehlenden Indizes $p(m_0+1),\ldots,p(m)$ der Ausgangsbasis sowie die Nichtbasisindizes $p(m+1),\ldots,p(n+m-m_0)$ werden durch den Algorithmus automatisch bestimmt.

Erläuterungen zum zweiphasigen revidierten Simplexalgorithmus: Die feste Spaltenzahl der Ma-

trix A wird hier mit n_0 bezeichnet, der Kostenvektor mit $\vec{c}^{(0)}$. Der Vektor $\vec{x} = (x_1, x_1, \ldots, x_{n_0})^T$ gibt nach erfolgter Rechnung die optimale Lösung an, die das Zielfunktional $z_0 = \langle \vec{c}^{(0)}, \vec{x} \rangle$ minimal macht. Die Variable z_0 gibt dieses Minimum an. Über mehrdeutige Lösbarkeit erteilt der Algorithmus keine Auskunft. Der Zeiger *init* zeigt an, in welchem Status das Programm sich befindet. Es gilt $init = 1$, solange eine zulässige Basislösung $\vec{x}_B$ von (P) zu einer Ausgangsbasis B gesucht wird. Befindet sich das Programm in der Phase II des Simplexalgorithmus, so wird dieser Status durch $init = 0$ angezeigt. In Zeile 2–14 wird die Matrix A durch die (noch fehlenden) Einheitsvektoren der Standardbasis des $\mathbf{R}^m$ ergänzt. Die resultierende Basismatrix A_B ist orthogonal; ihre Inverse A_B^T wird in der Matrix $D = (d_{jk})$ erzeugt. Der Hilfskostenvektor $\vec{c} := \vec{e}$ mit $c_j = 1$ für $j \geq m_0 + 1$ wird initialisiert, und es werden die zugeordneten Anfangskosten $z_1 = \langle \vec{c}, \vec{h} \rangle$ berechnet. In Zeile 17–22 werden die Nichtbasisindizes $p(m+1), p(m+2), \ldots, p(n_0)$ bestimmt. Die Zeilen 23–35 sowie 76–98 enthalten die bereits weiter oben beschriebenen Teile der Phase II des revidierten Simplexalgorithmus (2.10). In ihnen ist die BLANDsche Zusatzregel (Zeile 33, Zeile 86–87) inkorporiert. In den Zeilen 42–46 wird geprüft, ob die gefundene Ausgangsbasis B künstliche Basisindizes $p(j) > n_0$ enthält. Wenn ja, so sind gemäß der oben beschriebenen Fallstudie die Unterfälle **Fall 2(a)** und **Fall 2(b)** zu unterscheiden. Die Situation von **Fall 2(a)** wird in den Zeilen 46–65 algorithmisch analysiert, während sich der **Fall 2(b)** in den Zeilen 66–75 widerspiegelt.

Mit $\delta \geq 0$ haben wir einen Parameter der Maschinengenauigkeit eingeführt: es sei δ die kleinste positive Zahl, für die der verwendete Rechner das Additionsergebnis $1 + \delta = 1$ liefert. Mit der Größe δ wird das Auftreten von Rundungsfehlern berücksichtigt, die insbesondere dann von Gewicht sind, wenn fast identische Zahlen voneinander subtrahiert werden. Der oben beschriebene Algorithmus reagiert besonders dann empfindlich gegenüber Rundungsfehlern, wenn das Gleichungssytem $A\vec{x} = \vec{b}$ mehrere linear abhängige Gleichungen enthält. Durch Elimination linear abhängiger Zeilen können Parameterwerte in der Größenordnung der Maschinengenauigkeit entstehen, die bei exakter Rechnung Null sein sollten. Sensible Abfragen vom Typ $s < 0$? (Zeile 33), $z_1 = 0$? (Zeile 36), $s > 0$? (Zeile 55) oder $w_j > 0$? (Zeile 81) können dadurch erheblich verfälscht werden und zu sinnlosen Resultaten führen. Wir demonstrieren die Wirkung solcher "Schmutzeffekte" in dem folgenden Beispiel.

BSP. (15.2.6) Die Gleichungsrestriktionen des linearen Programms (P) aus BSP. (15.2.5) erweitern wir um drei Gleichungen, die von den drei bereits vorhandenen Gleichungen linear abhängig sind. Wir starten den oben angegebenen zweiphasigen Simplexalgorithmus mit der Vorgabe einer Maschinengenauigkeit $\delta = 0$. Da in der Phase I das Abfragekriterium $z_1 = 0$? (Zeile 36) nicht exakt erreichbar ist, erhält man die (falsche) Abbruchinformation

Es existiert keine zulässige Lösung.

Bei Vorgabe einer Maschinengenauigkeit $0 \leq \delta \leq 3.4 \cdot 10^{-18}$ wird das Abfragekriterium $w_j > 0$? (Zeile 81) nicht exakt realisierbar. Man erhält nun die (falsche) Abbruchinformation

Zielfunktion unbeschränkt.

$\vec{c}^T$	
A	$\vec{b}$

$=$

-0.75	20	-0.5	6	0	0	0	
0.25	-8	-1	9	1	0	0	0
0.5	-12	-0.5	3	0	1	0	0
0	0	1	0	0	0	1	1
0.75	-20	-1.5	12	1	1	0	0
0.5	-12	0.5	3	0	1	1	1
0.25	-8	0	9	1	0	1	1

Gibt man hingegen eine Maschinengenauigkeit $\delta \geq 3.5 \cdot 10^{-18}$ vor, so liefert der Simplexalgorithmus die korrekte Lösung, wie das folgende Rechenprotokoll belegt:

Maschinengenauigkeit: $\delta = 3.5\mathrm{E}^{-18}$, **Anfangsiterationen:** 6

Basisindizes der Basen B

$it=1$	$it=2$	$it=3$	$it=4$	$it=5$	$it=6$	$it=7$	$it=8$	$it=9$	$it=10$
8	1	1	3	3	3	3	3	3	3
9	9	2	2	4	4	4	4	4	5
10	10	10	10	10	1	1	1	1	1
11	11	11	11	11	11	11	11	0	0
12	12	12	12	12	12	12	0	0	0
13	13	13	13	13	13	0	0	0	0

Lösungsvektor $\vec{x}$**:**

$$\vec{x} = \begin{bmatrix} 1.000\,000\mathrm{E}^{+00} \\ 0.000\,000\mathrm{E}^{+00} \\ 1.000\,000\mathrm{E}^{+00} \\ 0.000\,000\mathrm{E}^{+00} \\ 7.500\,000\mathrm{E}^{-01} \\ 0.000\,000\mathrm{E}^{+00} \\ 0.000\,000\mathrm{E}^{+00} \end{bmatrix}$$

Minimum der Zielfunktion:

$$\boxed{z_{\min} = -1.250\,000\mathrm{E}^{+00}}$$

Das Protokoll der Basisindizes belegt, dass nach 6 Iterationsschritten eine zulässige Basislösung $\vec{x}_B$ und eine Ausgangsbasis $B = \{3, 4, 1, 11, 12, 13\}$ aus dem Hilfsprogramm (P_h) berechnet wurden. Allerdings enthält B die künstlichen Basisindizes $11, 12, 13$. Diese werden in drei weiteren Iterationsschritten durch Streichung der drei linear abhängigen Zeilen des Gleichungssytems $A\vec{x} = \vec{b}$ eliminiert. Die reduzierte Ausgangsbasis ist $B = \{3, 4, 1\}$. Nach einem weiteren Iterationsschritt ist die optimale Lösung $\vec{x}_B$ des Problems (P) zur Basis $B = \{3, 5, 1\}$ gefunden. Die hier dokumentierten Resultate wurden auf einem AT–Rechner 80286 mit eingebautem Co–Prozessor erzielt. Andere Rechner können unterhalb der Maschinengenauigkeit ganz andere Ergebnisse liefern.

Kapitel 16

Gewöhnliche Differentialgleichungen

16.1 Vorbetrachtungen, Problemstellung

In Abschnitt 14.4 wurden Gleichungen $\vec{F}(\vec{x},\vec{y})=\vec{0}$ studiert, die unter geeigneten Auflösbarkeitsbedingungen implizit eine Funktion $\vec{y}=\vec{f}(\vec{x})$ definieren. Treten in der Funktion $\vec{F}$ außer der gesuchten Funktion $\vec{y}(\vec{x})$ auch noch deren (partielle) Ableitungen nach den Koordinaten $x_1,x_2,\ldots,x_n$ auf, so heiße die Gleichung $\vec{F}=\vec{0}$ eine **Differentialgleichung** (DGl). Ist $x\in\mathbf{R}$ eine eindimensionale Variable, so liegt eine **gewöhnliche Differentialgleichung** vor; andernfalls spricht man von einer **partiellen Differentialgleichung**. In diesem Kapitel beschäftigen wir uns ausschließlich mit **gewöhnlichen** DGln.

Definition 16.1 (a) *Ist* $F\in\mathrm{Abb}\,(\mathbf{R}^{n+1},\mathbf{R})$ *eine gegebene skalare Funktion, so heiße die Gleichung*

$$\boxed{F(x,y,y',\ldots,y^{(n)})=0} \tag{1.1}$$

eine **implizite DGl** n**–ter Ordnung** *für eine gesuchte Funktion* $y=y(x)$. (*Ein Beispiel ist die* Euler*sche DGl* n*–ter Ordnung*

$$F(x,y,\ldots,y^{(n)}):=\sum_{k=0}^{n}a_kx^ky^{(k)}-RS(x)=0,\qquad a_n=1.)$$

(b) *Ist* $f\in\mathrm{Abb}\,(\mathbf{R}^n,\mathbf{R})$ *eine gegebene skalare Funktion, so heiße die Gleichung*

$$\boxed{y^{(n)}=f(x,y,y',\ldots,y^{(n-1)})} \tag{1.2}$$

eine **explizite DGl** n**–ter Ordnung** *für die gesuchte Funktion* $y=y(x)$. (*Ein Beispiel ist die obige* Euler*sche DGl* n*–ter Ordnung*

$$y^{(n)}=f(x,y,\ldots,y^{(n-1)}):=\frac{1}{x^n}\,RS(x)-\sum_{k=0}^{n-1}a_kx^{k-n}y^{(k)},\qquad x\neq 0.)$$

(c) *Ist* $\vec{F}\in\mathrm{Abb}\,(\mathbf{R}\times\mathbf{R}^{mn},\mathbf{R}^l)$ *eine gegebene vektorwertige Funktion, so heiße die Gleichung*

$$\boxed{\vec{F}(x,\vec{y},\vec{y}',\ldots,\vec{y}^{(n)})=\vec{0}} \tag{1.3}$$

ein **implizites DGl–System** n**–ter Ordnung** *für eine gesuchte Vektorfunktion* $\vec{y}=\vec{y}(x)\in\mathrm{Abb}\,(\mathbf{R},\mathbf{R}^m)$. *Ganz analog spricht man bei Gleichungen der Form*

$$\boxed{\vec{y}^{\,(n)}=\vec{f}(x,\vec{y},\vec{y}',\ldots,\vec{y}^{\,(n-1)})} \tag{1.4}$$

von einem **expliziten DGl–System** n**–ter Ordnung** *für die gesuchte Funktion* $\vec{y} = \vec{y}(x)$.

Beispiele von DGl–Systemen des Typs (1.4) sind die **linearen Systeme 1.Ordnung** mit konstanten Koeffizienten $\vec{y}\,' = A\vec{y} + \vec{g}(x)$, die wir bereits in Abschnitt 11.7 studiert haben.

In der klassischen Theorie der gewöhnlichen DGln wird der folgende Lösungsbegriff zugrunde gelegt:

Definition 16.2 *Eine Funktion* $y = y(x)$ *heiße auf einem Intervall* $I \subset \mathbf{R}$ *eine* **klassische Lösung** *der gewöhnlichen DGl* (1.1) *bzw.* (1.2)*, wenn* $y \in C^n(I)$ *gilt und wenn die Gleichungen* (1.1) *bzw.* (1.2) *nach Einsetzen von* $y(x), y'(x), \ldots, y^{(n)}(x)$ *für alle* $x \in I$ *erfüllt sind.*

Bemerkung 16.1 (a) Andere Lösungsbegriffe (*verallgemeinerte Lösung, distributionelle Lösung*) werden in speziellen Lösungstheorien verwendet.

(b) Die Bestimmung aller Lösungen einer DGl heißt die **Integration der DGl**.

(c) Eine DGl ist im allgemeinen nicht eindeutig lösbar. Wir erinnern an die linearen DGln mit konstanten Koeffizienten, die in Abschnitt 10.4 behandelt wurden. Zum Beispiel kann gefordert werden, dass die gesuchte Lösung $y = y(x)$ an einer Anfangsstelle $a \in I$ die n **Anfangswerte**

$$y(a) = C_1,\ y'(a) = C_2, \ldots, y^{(n-1)}(a) = C_n \tag{1.5}$$

annimmt. Diese Nebendingungen sind mindestens mit der expliziten DGl (1.2) kompatibel, weil nämlich aus (1.2) eine Bestimmungsgleichung $y^{(n)}(a) = f(a, C_1, \ldots, C_n)$ für den Funktionswert $y^{(n)}(a)$ folgt. Erfüllt die Funktion in der Gleichung (1.1) die Auflösbarkeitsbedingung $\frac{\partial F}{\partial y^{(n)}} \neq 0$, so kann die Gleichung (1.1) unter geeigneten Stetigkeitsannahmen in die explizite Form (1.2) überführt werden. Gilt hingegen $\frac{\partial F}{\partial y^{(n)}} = 0$, so können **singuläre Lösungen** der DGl (1.1) auftreten. □

Die Lösungsgesamtheit einer expliziten DGl n–ter Ordnung verfügt also über n **Freiheitsgrade**.

Definition 16.3 *Eine Lösung* $y = y(x, C_1, \ldots, C_n)$ *heiße* **allgemeine Lösung** *der DGl* n*–ter Ordnung* (1.1) *oder* (1.2)*, wenn jede spezielle Lösung durch geeignete Wahl der Konstanten* $C_1, C_2, \ldots, C_n$ *aus der Lösung* $y = y(x, C_1, \ldots, C_n)$ *konstruiert werden kann. Jede Lösung, die auf diese Weise für spezielle Werte der freien Konstanten* $C_1, C_2, \ldots, C_n$ *aus der allgemeinen Lösung gewonnen wurde, heiße eine* **partikuläre Lösung** *der DGl.*

Die explizite DGl (1.2) hat unter sehr allgemeinen Voraussetzungen an die Funktion f stets eine allgemeine Lösung. Dies wird in Abschnitt 16.5 zu zeigen sein. Der allgemeine Fall (1.1) soll hier nicht behandelt werden. Darüber hinaus schränken wir unsere Betrachtungen auf einige spezielle Typenklassen ein, für die eine vollständige Lösungstheorie existiert. Das Anwendungsfeld für Differentialgleichungen ist sehr weiträumig; DGln treten u.a. in der Geometrie, der Physik, der Biomathematik, den technischen und ingenieurwissenschaftlichen Disziplinen auf. In diesen Bereichen ist man in der Regel nur an partikulären Lösungen der relevanten DGln interessiert, die zum Beispiel durch den Anfangszustand eines physikalischen Systems oder durch Bedingungen an den Rändern des Betrachtungsintervalls aus der allgemeinen Lösung selektiert werden. Demgemäß hat man zu unterscheiden:

Definition 16.4 (a) **Anfangswertaufgaben:** *Zu festem* $a \in I$ *und zu vorgegebenen Zahlen* $y_0, y_1, \ldots, y_{n-1}$ *ist eine Lösung* $y \in C^n(I)$ *gesucht mit*

(AWA)
$$\boxed{\begin{aligned} &y^{(n)} = f(x, y, y', \ldots, y^{(n-1)}) \quad \text{für} \quad x \in I; \\ &y(a) = y_0, \;\; y'(a) = y_1, \;\; \ldots \;\; , y^{(n-1)}(a) = y_{n-1}. \end{aligned}}$$

(b) **Randwertaufgaben:** *Auf einem Intervall $I := [a, b]$ ist eine Lösung $y \in C^n(I)$ gesucht, deren Funktionswerte in den Randpunkten $a, b \in I$ vorgegeben sind:*

$$y(a),\, y(b), \;\; y'(a),\, y'(b), \;\; \ldots \;\; , y^{(n-1)}(a),\, y^{(n-1)}(b)$$

(*oder Linearkombinationen oder nichtlineare Relationen dieser Funktionswerte*).

Die Theorie der **Anfangswertaufgaben** soll hier in einem hinreichend allgemeinen Rahmen behandelt werden; Randwertaufgaben – diese sind nur in speziellen Fällen lösbar – können erst im Rahmen weiterführender Lehrveranstaltungen behandelt werden.

Die zentralen Fragen, die im Kontext von Anfangs– und Randwertaufgaben zu beantworten sind, lauten:

(a) Existiert eine Lösung?

(b) Ist diese Lösung eindeutig?

(c) Hängt die Lösung stetig von den Parametern der Aufgabe (Anfangsdaten, Randdaten usw.) ab?

(d) Wie bestimmt man die Lösung?

Probleme, bei denen die Fragen (a)–(c) bejaht werden können, heißen **korrekt gestellt**. Oft gelingt es, die allgemeine Lösung einer DGl explizit zu bestimmen. Dann können die Fragen (a)–(d) simultan beantwortet werden. Dieser einfache Fall soll im folgenden Abschnitt 16.2 behandelt werden.

16.2 Lösungsverfahren für explizite Differentialgleichungen 1.Ordnung

Die explizite DGl 1.Ordnung kann sowohl in der Form

$$\boxed{\frac{dy}{dx} =: y' = f(x, y)} \tag{2.1}$$

als auch in der Form

$$\boxed{P(x, y)\, dx + Q(x, y)\, dy = 0} \tag{2.2}$$

vorgelegt sein. Beide Formen sind äquivalent, wenn vom trivialen Fall $Q = 0$ abgesehen wird. Die Implikation (2.1) $\Rightarrow$ (2.2) ergibt sich aus den Spezifikationen $P(x, y) := f(x, y)$ und $Q(x, y) := -1$, während die Implikation (2.2) $\Rightarrow$ (2.1) durch Wahl von $f(x, y) := -\frac{P(x,y)}{Q(x,y)}$ folgt. Eine allgemeine Existenzaussage für die Gleichung (2.1) werden wir in Abschnitt 16.5 treffen. Im vorliegenden Abschnitt beschränken wir uns auf spezielle rechte Seiten $f(x, y)$. Die Gleichung (2.2) wird in Abschnitt 16.3 behandelt.

Typ (A) Differentialgleichungen mit getrennten Veränderlichen. Es seien stetige Funktionen $f, g \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R})$ gegeben. Zur Lösung der DGl

$$\boxed{y' = f(x) \cdot g(y)} \tag{2.3}$$

verwendet man stets die wichtige Lösungsmethode der

Trennung der Veränderlichen (TdV).

Man setzt in (2.3) $y' = \frac{dy}{dx}$ und trennt nun nach Funktionen in der Variablen y bzw. in der Variablen x allein:

$$\boxed{\frac{dy}{g(y)} = f(x)\,dx \quad \overset{\text{unbest. Integration}}{\Rightarrow} \quad \int \frac{dy}{g(y)} = \int f(x)\,dx + C.} \tag{2.4}$$

Folgerung 16.1 *Auf Intervallen I mit $g(y) \neq 0\ \forall\, y \in I$ sind die Lösungen der DGl* (2.3) *implizit durch die Relation* (2.4) *bestimmt.*

Begründung: Mit $H(y) := \int \frac{dy}{g(y)}$ liegt eine Funktion $H \in C^1(I)$ vor, die die Bedingung $H'(y) = \frac{1}{g(y)} \neq 0$ erfüllt. Somit ist H monoton, und es existiert die Umkehrfunktion $y(x) = H^{-1}(\int f(x)\,dx + C)$ als C^1–Funktion. □

BSP. (16.2.1) Die DGl $y' = \frac{y}{x}$ ist vom Typ (2.3) mit $f(x) := \frac{1}{x}$, $x \neq 0$, und $g(y) := y$. Für $y \neq 0$ erhalten wir durch TdV:

$$\frac{dy}{y} = \frac{dx}{x} \quad \Rightarrow \quad \ln|y| = \int \frac{dy}{y} = \int \frac{dx}{x} + C = \ln|x| + \ln|C^*|, \quad C^* \neq 0,$$

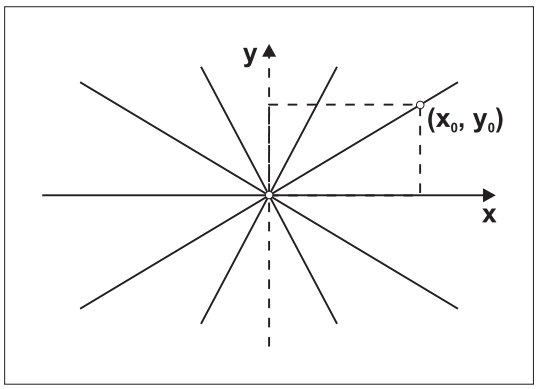
und somit die allgemeine Lösung

$$\boxed{y(x) = C^* x, \qquad C^* \neq 0.}$$

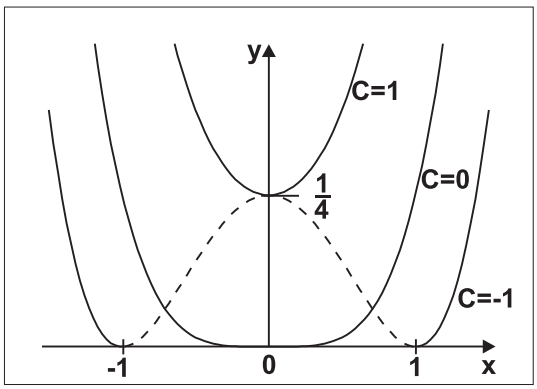
Die Lösungskurven sind Geradenbüschel durch den Ursprung; man verifiziert noch, dass auch $y = 0$ eine Lösung ist. Lediglich die Koordinatenachse $x = 0$ gehört nicht zur Lösungsschar. Durch jeden Punkt (x_0, y_0), $x_0 \neq 0$, verläuft somit genau eine Lösungskurve, und die Anfangswertaufgabe

$$y(x_0) = y_0, \qquad x_0 \neq 0, \tag{2.5}$$

hat stets eine eindeutige Lösung, nämlich $y(x) = \frac{y_0}{x_0}\,x$.



Die Lösungsschar der DGl $y' = \frac{y}{x}$



Die Lösungsschar der DGl $y' = 2x\sqrt{y}$

Allgemein gilt für die DGl (2.3) der folgende Existenzsatz:

Satz 16.1 *Es seien stetige Funktionen* $f, g \in \mathrm{Abb}(\mathbf{R}, \mathbf{R})$ *sowie ein innerer Punkt* $y_0 \in D(g)$ *mit* $g(y_0) \neq 0$ *gegeben. Dann hat die* AWA

$$\boxed{y' = f(x) \cdot g(y), \qquad y(x_0) = y_0,} \tag{2.6}$$

für jeden Punkt $x_0 \in D(f)$ *stets genau eine Lösung* $y(x)$*, welche implizit durch die folgende Gleichung definiert ist:*

$$\boxed{\int_{y_0}^{y} \frac{dt}{g(t)} = \int_{x_0}^{x} f(s)\, ds.}$$

Begründung: Da aus Stetigkeitsgründen $g(y) \neq 0$ in einer Umgebung von y_0 gilt, folgt die Behauptung unmittelbar aus (2.4). □

BSP. (16.2.2) Die DGl $y' = 2x\sqrt{y}$ ist ebenfalls vom Typ (2.3) mit $f(x) := 2x$ und $g(y) := \sqrt{y}$, $y > 0$. Beide Funktionen sind stetig, und somit folgt durch TdV:

$$\frac{dy}{\sqrt{y}} = 2x\, dx \quad \Rightarrow \quad 2\sqrt{y} = \int \frac{dy}{\sqrt{y}} = \int 2x\, dx + C = x^2 + C.$$

Da die linke Gleichungsseite positiv sein muss, resultiert die allgemeine Lösung

$$\boxed{y(x) = \tfrac{1}{4}\,(x^2 + C)^2, \qquad x^2 + C \geq 0.}$$

Der in der obigen Skizze gestrichelt gezeichnete Kurventeil gehört **nicht** zur Lösungsschar. Denn dann wäre z.B. die Lösung der AWA $y(0) = \frac{1}{4}$ **nicht eindeutig**, im Widerspruch zu Satz 16.1. Die eindeutige Lösung bestimmt man jedoch nach der Vorschrift des Satzes 16.1:

$$2\sqrt{y} - 2\sqrt{\frac{1}{4}} = \int_{\frac{1}{4}}^{y} \frac{dt}{\sqrt{t}} = \int_0^x 2s\, ds = x^2 \quad \Rightarrow \quad y(x) = \frac{1}{4}\,(x^2 + 1)^2.$$

Bemerkung 16.2 (a) Lösungen der AWA (2.6) sind in der Regel **lokale Lösungen**: Sie existieren lediglich auf einem Intervall $I(x_0)$ in der Umgebung des Anfangspunktes x_0, nämlich dort, wo $g(y(x)) \neq 0 \ \forall\, x \in I(x_0)$ gilt. In den seltensten Fällen hat man $I(x_0) = \mathbf{R}$ vorliegen.

(b) Was passiert mit der AWA (2.6) im Ausnahmefall $g(y_0) = 0$? Offensichtlich ist in diesem Fall eine **Lösung** der AWA durch die konstante Funktion $y^*(x) := y_0$ gegeben. Existieren Berührungspunkte (x, y_0) der Geraden $y^*(x)$ mit den Lösungskurven (2.4) – also Punkte mit gemeinsamer Tangente –, so kann man in (x, y_0) von dieser Geraden stetig differenzierbar in eine andere Lösungskurve überwechseln. Die AWA ist **mehrdeutig lösbar**. Ein solcher Fall tritt in BSP. (16.2.2) auf. Es gilt dort $g(y_0) = 0$ genau für $y_0 = 0$. Die Gerade $y^*(x) := 0$ hat Berührungspunkte $(x_C := \pm\sqrt{|C|}, 0)$ mit jeder der Lösungskurven $y(x) = \frac{1}{4}\,(x^2 + C)^2$, $C \leq 0$. Das heißt, die AWA $y(x_0) = 0$ ist in jedem Anfangspunkt x_0 unendlich vieldeutig. Der mathematische Grund liegt in der Tatsache, dass das uneigentliche Integral

$$\lim_{\epsilon \to 0(\pm)} \int_{y_0+\epsilon}^{y} \frac{dt}{g(t)} \quad \Big(= \int_{x_0}^{x} f(s)\, ds \Big)$$

existiert. Dies ist ein **hinreichendes Eindeutigkeitskriterium:** □

Satz 16.2 *Gegeben seien stetige Funktionen* $f, g \in \mathrm{Abb}(\mathbf{R}, \mathbf{R})$ *sowie ein Punkt* $y_0 \in D(g)$ *mit* $g(y_0) = 0$*. Hinreichend für die eindeutige Lösbarkeit der* AWA (2.6) *ist, dass das folgende uneigentliche Integral* **nicht existiert***:*

$$\boxed{\lim_{\epsilon \to 0(\pm)} \int_{y_0+\epsilon}^{y} \frac{dt}{g(t)}.} \tag{2.7}$$

BSP. (16.2.3) Es sei $y' = \frac{y}{x}$ die DGl aus BSP. (16.2.1). Hier gilt $g(y) := y$ und somit $g(y_0) = 0$ genau für $y_0 = 0$. Das uneigentliche Integral

$$\lim_{\epsilon \to 0} \int_{\epsilon}^{y} \frac{dt}{t}$$

existiert nicht, in Übereinstimmung mit der Tatsache, dass die AWA (2.6) im Punkt $(x_0, 0)$, $x_0 \neq 0$, die eindeutige Lösung $y(x) := 0$ besitzt.

Die folgenden DGln vom Typ (B) und (C) lassen sich durch Transformation auf den Typ (2.3) einer DGl mit getrennten Variablen zurückführen.

Typ (B) Die homogene Differentialgleichung. Das ist die DGl

$$\boxed{y' = g\Big(\frac{y}{x}\Big), \qquad x \neq 0,} \tag{2.8}$$

worin $g \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R})$ eine **stetige** Funktion sei. Die homogene DGl (2.8) wird stets mit dem **Ansatz**

$$\boxed{y(x) =: x \cdot u(x), \quad y'(x) = x \cdot u'(x) + u(x) \quad \Rightarrow \quad u' = \frac{g(u) - u}{x}, \quad x \neq 0,}$$

in eine DGl mit getrennten Variablen für die neue Funktion $u(x)$ transformiert.

Bemerkung 16.3 (a) Zur Erläuterung der Bezeichnung definieren wir: Eine skalare Funktion $f \in \mathrm{Abb}\,(\mathbf{R}^n, \mathbf{R})$ heiße **homogen vom Grade** $p \in \mathbf{R}$, wenn gilt:

$$f(\lambda\,\vec{x}) = \lambda^p f(\vec{x}) \;\; \forall\, 0 \neq \lambda \in \mathbf{R} \;\; \forall\, \vec{x} \in D(f).$$

Ist $f \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R})$ homogen vom Grade Null, so gilt also $f(x, y) = f(\lambda x, \lambda y) \;\; \forall\; \lambda \neq 0$. Für eine solche Funktion heiße die DGl (2.1) eine **homogene Differentialgleichung**. Mit der Spezifikation $\lambda := \frac{1}{x}$, $x \neq 0$, resultiert $f(x, y) = f(1, \frac{y}{x}) =: g(\frac{y}{x})$. Deshalb heißt die DGl (2.8) auch die **Normalform** einer homogenen Differentialgleichung.

(b) Eine Variablentransformation $z := \lambda\, x$ führt auf

$$y'(z) = \frac{d}{dx}\Big(\frac{1}{\lambda}\, y(\lambda x)\Big) = g\Big(\frac{\frac{1}{\lambda}\, y}{z}\Big).$$

Das heißt, ist $y_1(x)$ eine Lösung der DGl (2.8), so trifft dies auch auf die Funktion $y_2(x) := \frac{1}{\lambda}\, y_1(\lambda\, x)$ für jedes $\lambda \neq 0$ zu. Die allgemeine Lösung $y(x, C) := \frac{1}{C}\, y_p(Cx)$ gewinnt man somit aus jeder beliebigen partikulären Lösung $y_p(x)$. □

BSP. (16.2.4) Die Normalform der homogenen DGl

$$x^2 y' = a^2 x^2 + y^2 + xy, \qquad x \neq 0 \neq a,$$

erhält man nach Division durch x^2: $y' = a^2 + (\frac{y}{x})^2 + \frac{y}{x}$. Eine Transformation auf die neue Variable $u = \frac{y}{x}$ führt über die Relation $y' = xu' + u$ auf eine DGl mit getrennten Variablen: $u' = \frac{a^2 + u^2}{x}$. Wir integrieren mit dem Verfahren der TdV:

$$\frac{du}{a^2 + u^2} = \frac{dx}{x} \quad \Rightarrow \quad \frac{1}{a} \operatorname{arc\,tan}_H \frac{u}{a} = \int \frac{du}{a^2 + u^2} = \int \frac{dx}{x} = \ln|x| + \ln|C|, \quad C \neq 0.$$

Man kann explizit nach $u(x)$ auflösen, und durch Rücktransformation $y(x) = xu(x)$ erhält man die allgemeine Lösung

$$\boxed{y(x) = ax \tan\,(a\, \ln|Cx|), \qquad x \neq 0, \;\; C \neq 0,}$$

die tatsächlich die oben prognostizierte Form $y(x,C) = \frac{1}{C}\, y_p(Cx)$ hat, wenn man $y_p(x) := ax\tan(a\,\ln|x|)$ definiert.

BSP. (16.2.5) Zu bestimmen ist die Lösung der AWA

$$y' = \frac{y}{x} - \Big(\frac{y}{x}\Big)^2, \quad x \neq 0, \qquad y(1) = y_0.$$

Da hier die Normalform einer homogenen DGl vorliegt, stellen wir wieder mit der Transformation $u = \frac{y}{x}$ eine DGl mit getrennten Variablen her: $u' = -\frac{u^2}{x} =: f(x)\cdot g(u),\ g(u) := u^2$. Die Anfangsbedingung wird gemäß $u(1) = y_0$ transformiert. Wir haben $g(y_0) = 0$ genau für $y_0 = 0$, so dass dieser Anfangswert gesonderter Aufmerksamkeit bedarf. Das uneigentliche Integral

$$\lim_{\epsilon\to 0}\int_\epsilon^u \frac{dt}{g(t)}$$

ist jedoch divergent, so dass die AWA auch für den Anfangswert $y_0 = 0$ eine eindeutige Lösung besitzt, nämlich ganz offensichtlich die Lösung $y(x) := 0$. Für $y_0 \neq 0$ integrieren wir mit dem Verfahren der TdV:

$$\frac{du}{u^2} = -\frac{dx}{x} \quad \Rightarrow \quad \frac{1}{y_0} - \frac{1}{u} = \int_{y_0}^u \frac{dt}{t^2} = -\int_1^x \frac{ds}{s} = -\ln|x|, \quad x \neq 0.$$

Man kann wiederum nach $u(x)$ auflösen und findet somit durch Rücktransformation $y(x) = xu(x)$ die gesuchte Lösung der AWA:

$$\boxed{y(x) = \begin{cases} \dfrac{y_0 x}{1 + y_0\ln|x|} & : \ y_0 \neq 0, \\ 0 & : \ y_0 = 0. \end{cases}}$$

Typ (C) Für feste Zahlen $a, b, c \in \mathbf{R},\ b \neq 0$, und für eine stetige Funktion $f \in \mathrm{Abb}(\mathbf{R},\mathbf{R})$ betrachten wir die DGl

$$\boxed{y' = f(ax + by + c),} \tag{2.9}$$

die mit Hilfe des **Ansatzes**

$$\boxed{u(x) := ax + by(x) + c, \quad u'(x) = a + by'(x) \quad \Rightarrow \quad u' = a + bf(u)}$$

in eine DGl mit getrennten Variablen transformiert wird.

BSP. (16.2.6) Zu bestimmen ist die Lösung der AWA

$$y' = (x - y)^2, \qquad y(0) = y_0.$$

Der obige Ansatz hat hier die Form $u(x) = x - y(x)$, und er führt auf die AWA $u' = 1 - u^2 =: g(u),\ \ u(0) = -y_0$. Wegen $g(y_0) = 0$ genau für $y_0 = \pm 1$, müssen die beiden Anfangswerte $y_0 = \pm 1$ wieder einer gesonderten Betrachtung unterzogen werden. Das uneigentliche Integral

$$\lim_{\epsilon\to 0}\int_{\mp 1+\epsilon}^u \frac{dt}{1 - t^2}$$

ist jedoch divergent; zu diesen Anfangswerten gibt es eindeutig bestimmte Lösungen $u(x) = \mp 1$ bzw. $y(x) = x \pm 1$. Im Fall $y_0 \neq \pm 1$ integrieren wir wiederum mit dem Verfahren der TdV:

$$\frac{du}{1 - u^2} = dx \quad \Rightarrow \quad x = \int_0^x dx = \int_{-y_0}^u = \begin{cases} \operatorname{Ar}\tanh u + \operatorname{Ar}\tanh y_0 & : \ |y_0| < 1, \\ \operatorname{Ar}\coth u + \operatorname{Ar}\coth y_0 & : \ |y_0| > 1. \end{cases}$$

Hieraus erhält man die Lösung in der expliziten Form

$$\boxed{y(x) = \begin{cases} x - \tanh(x - C_1) & : \; C_1 := \operatorname{Ar}\tanh y_0, \quad |y_0| < 1, \\ x - \coth(x - C_2) & : \; C_2 := \operatorname{Ar}\coth y_0, \quad |y_0| > 1, \\ x \pm 1 & : \; y_0 = \pm 1. \end{cases}}$$

Typ (D) Die lineare Differentialgleichung 1.Ordnung. Für stetige Funktionen $p, q \in \operatorname{Abb}(\mathbf{R}, \mathbf{K})$ betrachten wir die DGl

$$\boxed{y' + p(x)\, y = q(x),} \tag{2.10}$$

die als Sonderfall $n = 1$ der bereits in Abschnitt 10.2 betrachteten allgemeinen linearen DGl n–ter Ordnung auftritt. Wir hatten dort zur Lösungskonstruktion nicht konkret Stellung bezogen. Wir verwenden hier wie in Abschnitt 10.2 die Bezeichnung

$$L_1 y := y' + p(x)\, y = \Big(\frac{d}{dx} + p(x)\Big) y.$$

Satz 16.3 *Gegeben seien ein Intervall $I \subset \mathbf{R}$ und Funktionen $p, q \in \operatorname{Abb}(\mathbf{R}, \mathbf{K})$ mit $p, q \in C(I)$.*

(a) *Die Lösungen $y_h \in C^1(I)$ der homogenen DGl $L_1 y = 0$ bilden einen Unterraum* $\operatorname{Kern} L_1 \subset C^1(I)$*; das heißt*

$$\boxed{L_1 y_1 = 0 = L_1 y_2 \quad \text{implizieren} \quad L_1(\lambda\, y_1 + \mu\, y_2) = 0 \;\; \forall\, \lambda, \mu \in \mathbf{K}.}$$

(b) *Ist $y_p \in C^1(I)$ eine partikuläre Lösung der inhomogenen DGl $L_1 y = q(x)$, so ist die allgemeine Lösung der DGl* (2.10) *der affine Unterraum*

$$\boxed{\mathcal{L}(DGl) = y_p + \operatorname{Kern} L_1 = \{y \in C^1(I) \; : \; y(x) = y_p(x) + y_h(x), \;\; y_h \in \operatorname{Kern} L_1\}.}$$

Dies ist lediglich eine Wiederholung des Satzes 10.1 für den Fall $n = 1$. Satz 16.3 gibt uns die Information, dass die Lösungskonstruktion für die DGl (2.10) in die folgenden zwei Teilaufgaben (H) und (P) zerfällt:

(H) Bestimme den Unterraum $\operatorname{Kern} L_1 \subset C^1(I)$, das heißt, die Lösungsgesamtheit der homogenen DGl $L_1 y := y' + p(x)\, y = 0$.

(P) Bestimme eine partikuläre Lösung $y_p \in C^1(I)$ der inhomogenen DGl $L_1 y := y' + p(x)\, y = q(x)$.

Satz 16.4 *Für gegebenes $p \in C(I)$ hat die homogene DGl $L_1 y = 0$ genau die Lösungen*

$$\boxed{y_h(x) := C e^{-P(x)} \quad \text{mit} \quad P(x) := \int p(x)\, dx, \quad x \in I.} \tag{2.11}$$

Begründung: Da die Funktion $e^{P(x)}$ auf dem Intervall I nullstellenfrei ist, gilt äquivalent mit der Gleichung $L_1 y = 0$:

$$0 = e^{P(x)} L_1 y = e^{P(x)}(y' + p(x)\, y) = \frac{d}{dx}(e^{P(x)} y), \qquad x \in I.$$

Aus Satz 7.13 folgt nun $e^{P(x)} y(x) = C = const \;\; \forall\, x \in I$, und dies führt schon auf die behauptete Relation (2.11). □

Bemerkung 16.4 Die Teilaufgabe (H) hat also genau die Lösung (2.11): Der Unterraum $\operatorname{Kern} L_1$ ist **eindimensional**. Man erhält (2.11) in gleicher Weise mit dem Verfahren der TdV:

$$\frac{dy}{y} = -p(x)\, dx \quad \Rightarrow \quad \ln|y| = \int \frac{dy}{y} = -\int p(x)\, dx + \ln|C|.$$

Auflösen nach $y = y(x)$ ergibt wiederum (2.11). □

Die **Teilaufgabe (P)** ist bei Kenntnis der allgemeinen Lösung $y_h(x)$ der homogenen DGl stets konstruktiv lösbar. Man bedient sich dazu des D'ALEMBERTschen Verfahrens der **Variation der Konstanten** (VdK). Dazu wird in der Darstellung (2.11) die Integrationskonstante C als **differenzierbare Funktion von** x aufgefasst:

$$\left.\begin{array}{l|l} y_p(x) = C(x)e^{-\int p(x)\,dx} & \cdot p(x) \\ y_p'(x) = C'(x)e^{-\int p(x)\,dx} - C(x)p(x)e^{-\int p(x)\,dx} & \cdot 1 \end{array}\right\} (+)$$

Durch Einsetzen in die DGl (2.10) resultiert eine Differentialgleichung für die Unbekannte $C(x)$, nämlich

$$C'(x)e^{-\int p(x)\,dx} = q(x),$$

die aber sofort direkt integriert werden kann:

$$C(x) = \int q(x)e^{\int^x p(t)\,dt}\,dx, \qquad x \in I.$$

Daraus erhalten wir eine partikuläre Lösung der inhomogenen DGl (2.10) in der Form

$$\boxed{y_p(x) = e^{-P(x)} \int^x q(t)e^{P(t)}\,dt, \qquad x \in I.} \tag{2.12}$$

Satz 16.5 *Gegeben seien ein Intervall $I \subset \mathbf{R}$ und Funktionen $p, q \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K})$ mit $p, q \in C(I)$. Es sei $P(x) := \int p(x)\,dx$ eine Stammfunktion von p. Dann hat die lineare DGL*

$$L_1 y := y' + p(x)\,y = q(x), \qquad x \in I,$$

die allgemeine Lösung

$$\boxed{y(x) = y_h(x) + y_p(x) = e^{-P(x)}\Big(C + \int^x q(t)e^{P(t)}\,dt\Big), \qquad x \in I, \quad C = const.} \tag{2.13}$$

Die Anfangswertaufgabe

$$L_1 y = q(x), \quad x \in I, \qquad y(x_0) = y_0 \quad \text{mit} \quad x_0 \in I \tag{2.14}$$

ist stets eindeutig lösbar mit der Lösung

$$\boxed{y(x) = e^{-P_0(x)}\Big(y_0 + \int_{x_0}^x q(t)e^{P_0(t)}\,dt\Big), \qquad x \in I, \quad P_0(x) := \int_{x_0}^x p(s)\,ds.} \tag{2.15}$$

BSP. (16.2.7) In der linearen DGl 1.Ordnung

$$y' - 2\Big(x + \frac{1}{x}\Big)y = 1, \qquad I := (0, +\infty)$$

gilt mit der Spezifikation $p(x) := -2(x + \frac{1}{x})$ und $q(x) := 1$ sicher $p, q \in C(I)$. Somit besitzt p eine Stammfunktion, nämlich $P(x) = \int p(x)\,dx = -(x^2 + \ln x^2)$, und wir erhalten gemäß Satz 16.4 die allgemeine Lösung der homogenen DGl

$$y_h(x) = Ce^{x^2 + \ln x^2} = Cx^2e^{x^2}, \qquad x \in I.$$

Die Formel (2.12) liefert eine partikuläre Lösung der inhomogenen DGl:

$$\begin{aligned} y_p(x) &= x^2e^{x^2}\int^x \frac{1}{t^2}\,e^{-t^2}\,dt \overset{\text{part. Int.}}{=} -x^2e^{x^2}\Big(\frac{1}{x}\,e^{-x^2} + 2\int^x e^{-t^2}\,dt\Big) \\ &= -x - x^2e^{x^2}\Big(\sqrt{\pi}\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt + \tilde{C}\Big). \end{aligned}$$

Hier tritt das Gausssche Fehlerintegral

$$\operatorname{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt, \qquad x \geq 0,$$

auf, vgl. BSP. (8.3.4). Somit resultiert die allgemeine Lösung

$$\boxed{y(x) = y_h(x) + y_p(x) = -x + x^2 e^{x^2} (C - \sqrt{\pi} \operatorname{erf}(x)), \quad x \in I.}$$

BSP. (16.2.8) Ein Kapital wird in gleicher Höhe K_0 sowohl bei der Stadt– und Kreissparkasse zum festen Jahreszins $z_1 := 7.75\%$ als auch bei der Deutschen Bank bei laufender Verzinsung zum Zinssatz $z_2 := 7.5\%$ angelegt. Welche Anlageform hat nach einem Jahr den höheren Ertrag gebracht?

Lösung: (A) *Feste Verzinsung.* Das Kapital K beträgt nach einem Jahr:

$$K = K_0 + z_1 K_0 = 1.0775\, K_0.$$

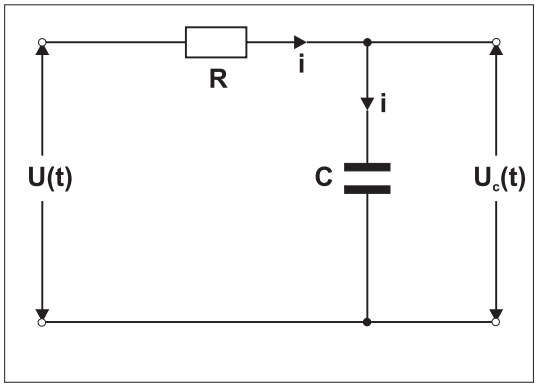
(B) *Laufende Verzinsung.* Das Kapital K erhält man als Lösung der AWA

$$\frac{dK}{dt} = z_2 K, \qquad K(0) = K_0$$

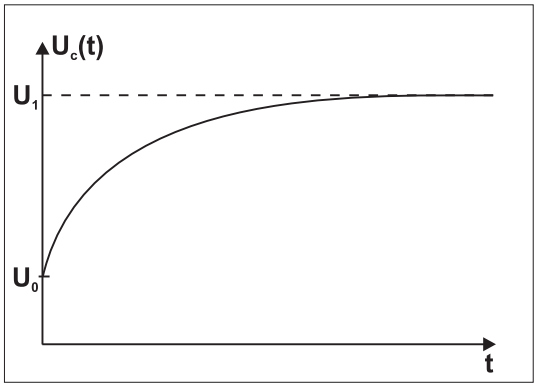
zum Zeitpunkt $t = 1$. Da diese AWA eindeutig durch $K(t) = K_0 e^{z_2 t}$ gelöst wird, resultiert

$$K = K(1) = e^{0.075} \cdot K_0 \doteq 1.07788\, K_0.$$

Das heißt, die Anlageform (B) erweist sich als die günstigere.



RC–**Glied, bestehend aus einem Widerstand** R **und einer Kapazität** C



Ladevorgang eines Kondensators

BSP. (16.2.9) Beim oben skizzierten *RC*–Glied – der Hintereinanderschaltung von Ohmschem Widerstand R und Kapazität C – ist die Spannung am Kondensator $u_C(t)$ in Abhängigkeit von der Eingangsspannung $u(t)$ zu bestimmen.

Lösung: Nach den Kirchhoffschen Gesetzen hat man

$$u = u_R + u_C, \qquad u_R = iR, \qquad u_C = \frac{Q}{C}, \qquad i = \frac{dQ}{dt}.$$

Darin bezeichnet Q die Ladung am Kondensator. Durch Elimination von Q, u_R und i erhält man die lineare inhomogene DGl 1.Ordnung mit konstanten Koeffizienten

$$\dot{u}_C + \frac{1}{RC}\, u_C = \frac{1}{RC}\, u(t).$$

Der in der Lösung der homogenen DGl

$$\boxed{u_h(t) = u_h(0)e^{-\frac{t}{T}}, \qquad T := RC,}$$

auftretende Faktor $T = RC$ heißt die **Zeitkonstante** des RC–Gliedes, das ist diejenige Zeit, in der die Spannung $u_C(t) := u_h(t)$ ohne äußeren Einfluss auf den Bruchteil e^{-1} abfällt:

$$\frac{u_C(t+T)}{u_C(t)} = e^{-1}.$$

Als **Sonderfälle** der inhomogenen DGl betrachten wir:

(α) **Ladevorgang des Kondensators** bei konstanter Eingangsspannung $u(t) := u_1 = const.$ Da $u_p(t) := u_1$ eine partikuläre Lösung der inhomogenen DGl ist, wird die AWA bei $t = 0$ zum Anfangswert $u_C(0) = u_0$ eindeutig durch die folgende Funktion gelöst:

$$\boxed{u_C(t) = u_1 + (u_0 - u_1)e^{-\frac{t}{T}}.}$$

(β) **Wirkung als Integrierglied** gegenüber periodischer Eingangsspannung $u(t) = u(t+\omega)$: Wir bestimmen eine partikuläre Lösung $u_p(t)$ mit dem Ansatz der VdK:

$$\left.\begin{array}{l|l} u_p(t) = K(t)e^{-\frac{t}{T}} & \cdot\frac{1}{T} \\ u_p'(t) = K'(t)e^{-\frac{t}{T}} - K(t)\frac{1}{T}e^{-\frac{t}{T}} & \cdot 1 \end{array}\right\} (+)$$

Durch Einsetzen in die inhomogene DGl erhält man $K'(t) = \frac{1}{T}e^{\frac{t}{T}}u(t)$, und somit

$$u_p(t) = \frac{1}{T}\int_0^t e^{\frac{1}{T}(s-t)}\,u(s)\,ds.$$

Wir zeigen nun, dass aus der allgemeinen Lösung

$$u_C(t) = Ke^{-\frac{t}{T}} + \frac{1}{T}\int_0^t e^{\frac{1}{T}(s-t)}\,u(s)\,ds$$

eine ω–periodische partikuläre Lösung $u_\omega(t)$ konstruiert werden kann, sofern $\omega \neq T$ gilt. Dazu muss **notwendig** $u_\omega(0) = u_\omega(\omega)$ erfüllt sein:

$$K = Ke^{-\frac{\omega}{T}} + \frac{1}{T}\int_0^\omega e^{\frac{1}{T}(s-\omega)}\,u(s)\,ds \quad \Rightarrow \quad K = \frac{1}{T(1-e^{-\omega/T})}\int_0^\omega e^{\frac{1}{T}(s-\omega)}\,u(s)\,ds.$$

Es resultiert nach einigen elementaren Rechenschritten

$$u_\omega(t) = \frac{e^{-t/T}}{T(1-e^{-\omega/T})}\Big(\int_t^\omega e^{\frac{1}{T}(s-\omega)}\,u(s)\,ds + \int_0^t e^{\frac{s}{T}}\,u(s)\,ds\Big),$$

und schließlich durch Substitution $s - \omega \mapsto s$ im ersten Integral:

$$\boxed{u_\omega(t) = \frac{1}{T(1-e^{-\omega/T})}\int_{t-\omega}^t e^{\frac{1}{T}(s-t)}\,u(s)\,ds.}$$

Dass die Bedingung $u_\omega(0) = u_\omega(\omega)$ auch **hinreichend** für die ω–Periodizität $u_\omega(t+\omega) = u_\omega(t)$ war, überprüft man nun durch Rechnung an der konstruierten Lösung. Die DGl hat somit die folgende allgemeine Lösung

$$\boxed{u_C(t) = u_h(t) + u_\omega(t) = Ke^{-\frac{t}{T}} + \frac{1}{T(1-e^{-\omega/T})}\int_{t-\omega}^t e^{\frac{1}{T}(s-t)}\,u(s)\,ds.}$$

Für $t \gg 1$ verhält sich diese Lösung asymptotisch wie $u_\omega(t)$:

$$u_\infty(t) \approx u_\omega(t) = \frac{1}{T(1-e^{-\omega/T})} \int_{-\omega}^{0} e^{\frac{s}{T}}\, u(s+t)\, ds.$$

Wird noch $\frac{\omega}{T} \ll 1$ angenommen, so gelten die Näherungen $1 - e^{-\omega/T} \approx \frac{\omega}{T}$ und $e^{s/T} \approx 1$. Es folgt

$$u_\omega(t) \approx \frac{1}{\omega} \int_{-\omega}^{0} u(s+t)\, ds, \qquad \frac{\omega}{T} \ll 1.$$

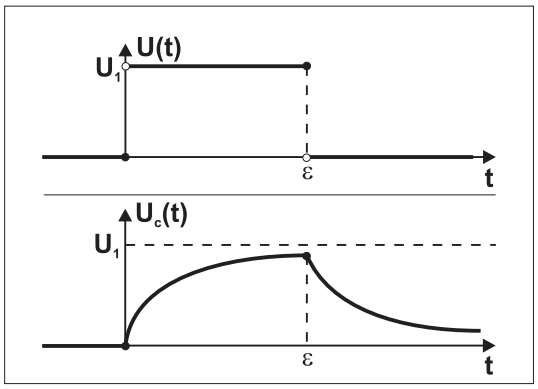
Wegen

$$\frac{d}{dt} \int_{-\omega}^{0} u(s+t)\, ds = \int_{-\omega}^{0} u'(s+t)\, ds = u(s+t)\Big|_{-\omega}^{s=0} = u(t) - u(t-\omega) = 0,$$

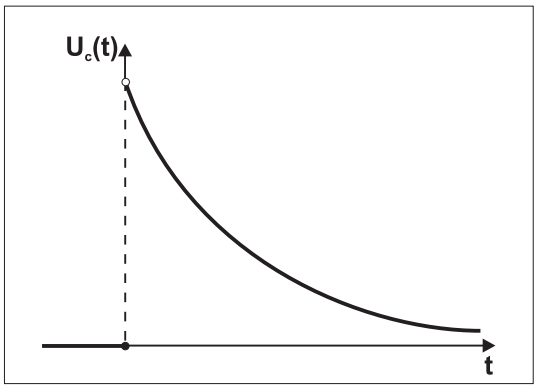
ist das Integral nun von t unabhängig, so dass schließlich resultiert:

$$u_\infty(t) \approx \frac{1}{\omega} \int_{-\omega}^{0} u(s+\omega)\, ds = \frac{1}{\omega} \int_{0}^{\omega} u(s)\, ds, \qquad \frac{\omega}{T} \ll 1, \quad t \gg 1.$$

Am Kondensator stellt sich also näherungsweise der **Integralmittelwert** der Eingangsspannung $u(t)$ ein, wenn eine hinreichend lange Zeitspanne verstrichen ist. Daher heißt das RC–Glied manchmal auch **Integrierglied**.



Sprungimpuls und Response an einem RC–Glied



Response auf einen DIRAC–Impuls

(γ) **Der Sprungimpuls**. Es sei $u(t)$ als zeitbegrenzter Sprungimpuls

$$u_\epsilon(t) := \begin{cases} 0 & : \ t \le 0, \\ u_1 & : \ 0 < t \le \epsilon, \\ 0 & : \ t > \epsilon \end{cases}$$

vorgegeben. Aus der Lösung unter (α) erhalten wir dann

$$\boxed{u_C(t) = u_1(1 - e^{-\frac{t}{T}}) \quad \text{für} \quad 0 \le t \le \epsilon.}$$

Daraus gewinnt man den Funktionswert $u_C(\epsilon) = u_1(1 - e^{-\frac{\epsilon}{T}})$ und somit für $t > \epsilon$ die Lösung

$$\boxed{u_C(t) = u_C(\epsilon)\, e^{-\frac{1}{T}(t-\epsilon)} = u_1(e^{\frac{\epsilon}{T}} - 1)\, e^{-\frac{t}{T}}, \qquad t > \epsilon.}$$

Wird speziell $u_1 := \frac{U}{\epsilon}$ angenommen, so erfährt der Kondensator einen maximalen Spannungsstoß

$$u_{C\,\max} = u_C(\epsilon) = \frac{U}{\epsilon}\,(1 - e^{-\frac{\epsilon}{T}}) \le \frac{U}{T}, \qquad \epsilon \ge 0,$$

und es existiert der Grenzwert

$$\lim_{\epsilon\to 0+} u_C(t) = \frac{U}{T}\, e^{-\frac{t}{T}}, \qquad t > 0,$$

obwohl der Grenzwert der Eingangsspannung

$$\lim_{\epsilon\to 0+} u_\epsilon(t)$$

im Funktionensinn **nicht** existiert. Man spricht von einem DIRAC–**Impuls** der Intensität U. Die Behandlung von Problemen dieser Art wird erst im Rahmen der **Distributionentheorie** angemessen ermöglicht und verständlich.

Typ (E) Die BERNOULLI–**Differentialgleichung**. Das ist die DGl

$$\boxed{y' + p(x)\, y = q(x)\, y^r, \qquad y \geq 0,} \tag{2.16}$$

worin $p, q \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R})$ **stetige** Funktionen seien und $0 \neq r \neq 1$ gelte. Für ganzzahliges r können auch Lösungen $y < 0$ sinnvoll sein. Die BERNOULLI–DGl (2.16) wird stets mit dem **Ansatz**

$$\boxed{z(x) := y^{1-r}(x), \quad z'(x) = (1-r)y^{-r}(x)\cdot y'(x) \quad \Rightarrow \quad z' + (1-r)p(x)\, z = (1-r)q(x)}$$

in eine lineare DGl 1.Ordnung für die neue Funktion $z(x)$ transformiert.

BSP. (16.2.10) Zu bestimmen ist die Lösung der AWA

$$y' + \frac{2x}{1+x^2}\, y = \frac{2x}{\sqrt{1+x^2}}\, \sqrt{|y|}\, \mathrm{sign}\, y, \qquad y(2) = \frac{9}{20}.$$

Fall (A): Der Anfangswert $y_0 := \frac{9}{20}$ liegt im Bereich $y > 0$. Wir suchen also Lösungen $y > 0$ der BERNOULLI–DGl

$$y' + \frac{2x}{1+x^2}\, y = \frac{2x}{\sqrt{1+x^2}}\, y^r, \qquad r := \frac{1}{2},$$

zum Anfangswert $y(2) = y_0$. Wir stellen mit der Transformation $z(x) := \sqrt{y(x)}$ eine AWA für eine lineare DGl 1.Ordnung her:

$$z' + \frac{x}{1+x^2}\, z = \frac{x}{\sqrt{1+x^2}}, \qquad z(2) = \frac{3}{10}\sqrt{5}.$$

Zunächst lösen wir die homogene DGl mit dem Verfahren der TdV:

$$\frac{dz}{z} = -\frac{x\,dx}{1+x^2} \quad \Rightarrow \quad \ln|z| = \int \frac{dz}{z} = -\int \frac{x\,dx}{1+x^2} = -\frac{1}{2}\,\ln(1+x^2) + \ln|C|.$$

Es resultiert die Lösung

$$z_h(x) = \frac{C}{\sqrt{1+x^2}}, \qquad x \in \mathbf{R}.$$

Eine partikuläre Lösung der inhomogenen DGl ermitteln wir mit dem Ansatz der VdK:

$$\left.\begin{array}{l|l} z_p(x) = \dfrac{C(x)}{\sqrt{1+x^2}} & \cdot\dfrac{x}{1+x^2} \\ z_p'(x) = \dfrac{C'(x)}{\sqrt{1+x^2}} - \dfrac{xC(x)}{(1+x^2)^{3/2}} & \cdot 1 \end{array}\right\} (+)$$

Durch Einsetzen in die inhomogene DGl erhält man $C'(x) = x$ und somit $z_p(x) = \frac{x^2}{2\sqrt{1+x^2}}$. Die allgemeine Lösung der transformierten DGl lautet nun

$$z(x) = z_h(x) + z_p(x) = \frac{C^* + x^2}{2\sqrt{1+x^2}}, \qquad x \in \mathbf{R}, \quad C^* := 2C.$$

Bei der Rücktransformation $z(x) = \sqrt{y(x)}$ muss jedoch auf die Bedingung $z(x) \geq 0$ oder äquivalent $x^2 \geq -C^*$ geachtet werden. Deshalb resultiert

$$y(x) = \frac{(C^* + x^2)^2}{4(1+x^2)}, \qquad x^2 \geq -C^*,$$

und die Lösung der AWA erfordert $z(2) = \frac{1}{10}\sqrt{5}(C^* + 4) \stackrel{!}{=} \frac{3}{10}\sqrt{5}$, also $C^* = -1$. Somit lautet die gesuchte Lösung der AWA

$$\boxed{y(x) = \frac{(x^2-1)^2}{4(x^2+1)}, \qquad x^2 \geq 1.}$$

Fall (B): Lösungen $y < 0$ der DGl sind für die gestellte AWA zwar nicht relevant, sie existieren aber. Setzt man nämlich $u(x) := -y(x)$, $u > 0$, so gilt für $u(x)$ dieselbe BERNOULLI–DGl wie in (A):

$$u' + \frac{2x}{1+x^2}\, u = \frac{2x}{\sqrt{1+x^2}}\, u^r, \qquad r := \frac{1}{2}.$$

Wir erhalten somit die Lösungsschar

$$\boxed{y(x) = -\frac{(C^* + x^2)^2}{4(1+x^2)}, \qquad x^2 \geq -C^*.}$$

Typ (F) Die RICCATI–**Differentialgleichung**. Das ist die DGl

$$\boxed{y' + p(x)\, y + q(x)\, y^2 = r(x),} \tag{2.17}$$

worin $p, q, r \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R})$ **stetige** Funktionen seien. Abgesehen von Spezialfällen – zum Beispiel liegt für $r(x) := 0$ eine BERNOULLI–DGl vor – ist die RICCATI–DGl (2.17) i.a. **nicht** geschlossen lösbar. In einigen Fällen kann jedoch eine partikuläre Lösung geraten werden. Mit dieser Kenntnis ist es dann möglich, die allgemeine Lösung anzugeben:

Satz 16.6 *Gegeben seien ein Intervall $I \subset \mathbf{R}$ und Funktionen p, q, $r \in C(I)$. Ist $y_1 \in C^1(I)$ eine partikuläre Lösung der* RICCATI*–DGl* (2.17)*, so findet man ihre allgemeine Lösung mit Hilfe des Ansatzes*

$$\boxed{y(x) = y_1(x) + u(x)} \tag{2.18}$$

und durch Integration der BERNOULLI*–DGl*

$$\boxed{u' + (p(x) + 2q(x)y_1(x))\, u = -q(x)\, u^2.} \tag{2.19}$$

Diese überführt man mittels der Transformation $z(x) := \frac{1}{u(x)}$ in die lineare DGl 1.Ordnung für die gesuchte Funktion $z(x)$:

$$\boxed{z' - (p(x) + 2q(x)y_1(x))\, z = q(x).} \tag{2.20}$$

Begründung: Diese erfolgt unmittelbar durch Einsetzen der angegebenen Ansätze in die DGl. □

BSP. (16.2.11) Wir betrachten die RICCATI–DGl

$$y' + 4x^3 y - 2xy^2 = 2x(x^4 + 1).$$

Als *Daumenregel* zur Ermittlung einer partikulären Lösung versuche man einen Ansatz der Form $y_1(x) = \beta x^\alpha$. Hier resultiert:

$$\beta\alpha x^{\alpha-1} + 4\beta x^{\alpha+3} - 2\beta^2 x^{2\alpha+1} \stackrel{!}{=} 2x^5 + 2x.$$

Das Paar $\alpha = 2$, $\beta = 1$ führt offenbar zum gewünschten Erfolg; das heißt, die gegebene RICCATI–DGl hat eine partikuläre Lösung $y_1(x) = x^2$. Wir können somit die lineare DGl (2.20) aufstellen:

$$z' - \underbrace{(4x^3 - 4x \cdot x^2)}_{=0} z = -2x.$$

Die elementar bestimmbare Lösung $z(x) = C - x^2$ führt nun für $x^2 \neq C$ auf die Funktion $u(x) := \frac{1}{z(x)} = \frac{1}{C-x^2}$, und vermöge (2.18) erhalten wir die allgemeine Lösung in der Form

$$\boxed{y(x) = x^2 + \frac{1}{C - x^2}, \qquad x^2 \neq C.}$$

Man beachte, dass die partikuläre Lösung $y_1(x) = x^2$ im Limes $C \to \infty$ aus der allgemeinen Lösung folgt.

16.3 Die vollständige Differentialgleichung und der integrierende Faktor

Einer skalaren Funktion $y \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R})$, die über einem Intervall $I \subset \mathbf{R}$ die Regularität $y \in C^1(I)$ besitzt, kann stets vermöge

$$\vec{x}(t) := \begin{bmatrix} t \\ y(t) \end{bmatrix}, \quad \|\dot{\vec{x}}(t)\| = \sqrt{1 + y'^2(t)} \geq 1, \qquad t \in I,$$

eine ebene reguläre Parameterkurve $\vec{x} = \vec{x}(t)$ zugeordnet werden. Die Umkehrung gilt nicht: Nicht jede ebene reguläre Parameterkurve $\vec{x} = \vec{x}(t)$ lässt eine explizite Darstellung $y = y(x)$ mit $y \in C^1(I)$ zu. *Zum Beispiel* gestattet die durch die Gleichung

$$F(x, y) := x^2 + y^2 - c^2, \qquad c > 0, \tag{3.1}$$

implizit definierte Kreisschar sehr wohl eine reguläre Parameterdarstellung $\vec{x}(t) = (c\cos t, c\sin t)^T$, $t \in [0, 2\pi)$, während eine explizite Darstellung nur **lokal** für die zwei Halbkreise

$$y_\pm(x) = \pm\sqrt{c^2 - x^2}, \qquad x \in I := [-c, c],$$

möglich ist. Diese Darstellung ist nicht einmal auf dem ganzen Intervall I C^1–regulär; in den Intervallendpunkten $x = \pm c$ existieren keine endlichen Ableitungen. Als Ursache stellen wir fest, dass durch implizites Differenzieren der Gleichung (3.1) die Beziehung $0 = F_x(x, y) + F_y(x, y) \cdot y'(x) = 2x + 2y(x) \cdot y'(x)$ resultiert und daraus die explizite DGl vom Typ (2.1)

$$y' = -\frac{x}{y} =: f(x, y), \tag{3.2}$$

in der sich die Singularität in der Ableitung der expliziten Darstellung bei $y = 0$ widerspiegelt. Da die Funktion $F(x, y)$ vollständig symmetrisch in den beiden Variablen x, y ist, erhält man durch Rollentausch ganz analog zu (3.2) die explizite DGl

$$x' = -\frac{y}{x} := \tilde{f}(x, y) \tag{3.3}$$

mit den Lösungskurven $x_\pm(y) = \pm\sqrt{c^2 - y^2}$. Diese sind in den Punkten $y = \pm c$ nicht mehr differenzierbar, also dort, wo $x = 0$ gilt und somit, wo die DGl (3.3) singulär wird. Man vermeidet die offenbar nur von der Wahl der expliziten Darstellungen abhängigen Singuläritäten in den DGln (3.2) und (3.3) durch Betrachten der symmetrischen Form

$$x\,dx + y\,dy = 0, \tag{3.4}$$

in der die Symmetrie der Funktion $F(x, y)$ in x und y angemessen berücksichtigt wird. Darüber hinaus wird durch (3.4) nicht **kanonisch** festgelegt, welche der Variablen als abhängig oder als unabhängig gelten soll. Die Gleichung (3.4) hat die Form (2.2) einer allgemeinen expliziten DGl 1.Ordnung, nämlich

$$\boxed{P(x, y)\,dx + Q(x, y)\,dy = 0,} \tag{3.5}$$

worin $P, Q \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R})$ **stetige** Funktionen seien.

Definition 16.5 *Die Gleichung* (3.5) *heiße* **Differentialform einer DGl**. (*Diese Bezeichnung stammt aus der Differentialgeometrie.*)

Wie am Anfang von Abschnitt 16.2 festgestellt wurde, besteht eine Korrespondenz zwischen den Differentialformen von DGln und den expliziten DGln 1.Ordnung. Diese Korrespondenz ist **mehrdeutig**: Wird die Gleichung (3.5) mit einem Faktor $\lambda(x, y) \neq 0$ multipliziert, so entsteht sehr wohl eine neue Differentialform einer DGl, ohne dass die zugeordnete explizite DGl 1.Ordnung geändert wird:

$$P(x, y)\,dx + Q(x, y)\,dy = 0 \quad \rightleftharpoons \quad y' = -\frac{P(x, y)}{Q(x, y)}, \quad Q(x, y) \neq 0,$$
$$\rightleftharpoons \lambda(x, y)\,P(x, y)\,dx + \lambda(x, y)\,Q(x, y)\,dy = 0.$$

Im Falle $P(x, y) \neq 0$ gilt Gleiches auch für die explizite DGl $x' = -\frac{Q(x,y)}{P(x,y)}$. In Punkten (x_0, y_0) mit $P(x_0, y_0) = 0 = Q(x_0, y_0)$ kann der Differentialform (3.5) offenbar keine explizite DGl zugeordnet werden. Diese **singulären Punkte** werden noch Gegenstand weiterer Untersuchungen sei.

Das Beispiel der Kreisgleichung (3.1) lehrt, dass die Äquipotentiallinien $F(x, y) = c$ einer differenzierbaren Funktion $F \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R})$ der folgenden Differentialform genügen:

$$F_x(x, y)\,dx + F_y(x, y)\,dy \equiv dF = 0. \tag{3.6}$$

Da hier das **vollständige Differential** dF der Funktion F auftritt, erklärt sich die folgende Definition:

Definition 16.6 (a) *Die Differentialform* (3.5) *heiße* **vollständige DGl**, *wenn es eine* **stetig differenzierbare** *Funktion* $F \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R})$ *gibt mit*

$$\boxed{F_x(x, y) = P(x, y), \quad F_y(x, y) = Q(x, y).} \tag{3.7}$$

(*Das heißt, in diesem Fall ist das Vektorfeld* $(-P(x, y), -Q(x, y))^T$ *ein Potentialfeld mit dem Potential* $F(x, y)$.)

(b) *Eine stetig differenzierbare Funktion* $F \in \mathrm{Abb}\,(\mathbf{R}^2, \mathbf{R})$ *heiße eine* **Stammfunktion** *der Differentialform* (3.5), *wenn es eine Hilfsfunktion* $\lambda(x, y) > 0$ *gibt mit*

$$\boxed{F_x(x, y) = \lambda(x, y)\,P(x, y), \quad F_y(x, y) = \lambda(x, y)\,Q(x, y).} \tag{3.8}$$

Die Funktion λ *heißt dann ein* **integrierender Faktor** *oder* EULER**scher Multiplikator** *der Differentialform* (3.5).

Folgerung 16.2 *Ist* $F \in \mathrm{Abb}(\mathbf{R}^2, \mathbf{R})$ *eine Stammfunktion der Differentialform* (3.5)*, so ist die* **allgemeine Lösung** *von* (3.5) *implizit durch die folgende Gleichung gegeben:*

$$\boxed{F(x,y) = C = const.} \tag{3.9}$$

Begründung: Klar, aus (3.9) resultiert ja

$$0 = dF = F_x(x,y)\,dx + F_y(x,y)\,dy \overset{(3.8)}{=} \lambda(x,y)(P(x,y)\,dx + Q(x,y)\,dy),$$

und wegen $\lambda > 0$ muss nun (3.5) gelten. □

BSP. (16.3.1) Wir betrachten für $a, b > 0$ die Funktion $F(x,y) := a^2x^2 + b^2y^2$. Wegen $F_x(x,y) = 2a^2x$ und $F_y(x,y) = 2b^2y$ ist sie Stammfunktion der vollständigen DGl

$$2a^2x\,dx + 2b^2y\,dy = 0,$$

deren allgemeine Lösung somit die **Ellipsenschar** $F(x,y) = a^2x^2 + b^2y^2 = C \geq 0$ ist.

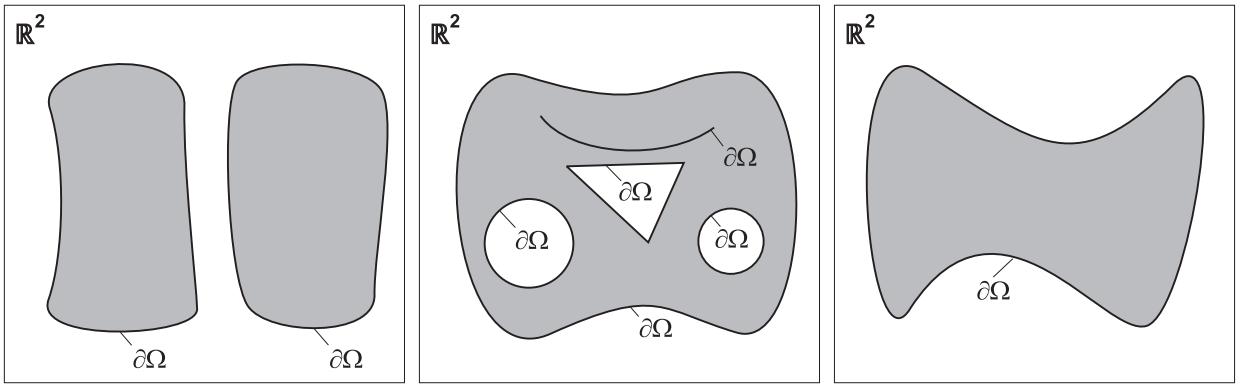
Stammfunktionen $F(x,y)$ der Differentialform (3.5) existieren immer, wenn eine vollständige DGl vorliegt. Ihre Bestimmung wirft zwei Probleme auf:

(A) Welches Kriterium gibt **ohne** Kenntnis der (Stamm–)Funktion F aus (3.7) an, dass die Differentialform vollständig ist?

(B) Welche Konstruktionsvorschrift gestattet die Bestimmung von Stammfunktionen einer vollständigen DGl?

Beide Fragen werden in dem folgenden Satz 16.7 beantwortet, zu dessen Vorbereitung wir den Gebietsbegriff noch spezialisieren müssen. Wir hatten in Definition 13.10 eine nichtleere, offene und wegzusammenhängende Teilmenge $\Omega \subset \mathbf{R}^n$ ein **Gebiet** genannt.

Definition 16.7 *Ein Gebiet* $\Omega \subset \mathbf{R}^2$ *heiße* **einfach zusammenhängend**, *wenn* **jede** *geschlossene Kurve* $\Gamma \subset \Omega$ *stetig auf einen Punkt* $\vec{x}_0 \in \Omega$ *zusammengezogen werden kann.* Ω *darf mit anderen Worten keine Löcher enthalten.*



Nicht zusammenhängende Teilmenge | **Mehrfach zusammenhängende Teilmenge** | **Einfach zusammenhängende Teilmenge**

Satz 16.7 *Es sei* $\Omega \subset \mathbf{R}^2$ *ein* **einfach zusammenhängendes** *Gebiet, und es seien Funktionen* $P, Q \in C^1(\Omega)$ *gegeben. Genau dann ist die Differentialform* (3.5) *vollständig, wenn gilt:*

$$\boxed{P_y(x,y) = Q_x(x,y) \;\; \forall\, (x,y) \in \Omega.} \tag{3.10}$$

Eine Stammfunktion $F(x,y)$ ist in diesem Fall gemäß

$$\boxed{F(x,y) := \int_{x_0}^{x} P(s,y)\,ds + \int_{y_0}^{y} Q(x_0,t)\,dt, \qquad (x,y) \in \Omega,} \tag{3.11}$$

definiert, worin $(x_0,y_0) \in \Omega$ ein beliebiger Punkt ist, und worin die Strecken $\overline{x_0x}$ und $\overline{y_0y}$ ganz in Ω verlaufen müssen.

Begründung: Ist die Differentialform (3.5) vollständig, so existiert eine Funktion $F \in C^1(\Omega)$ mit $F_x(x,y) = P(x,y)$ und $F_y(x,y) = Q(x,y)$. Da nun $P, Q \in C^1(\Omega)$ gelten, muss sogar $F \in C^2(\Omega)$ vorliegen, und aus dem SCHWARZschen Vertauschungssatz resultiert die Bedingung (3.10):

$$P_y(x,y) = F_{xy}(x,y) = F_{yx}(x,y) = Q_x(x,y) \;\; \forall\, (x,y) \in \Omega.$$

Ist umgekehrt die Relation (3.10) wahr, so sei $F(x,y)$ durch (3.11) definiert. Sicher gilt $F \in C^1(\Omega)$ sowie in jedem Punkt $(x,y) \in \Omega$:

$$\begin{aligned} F_x(x,y) &= P(x,y), \\ F_y(x,y) &= \int_{x_0}^{x} P_y(s,y)\,ds + Q(x_0,y) \overset{(3.10)}{=} \int_{x_0}^{x} Q_x(s,y)\,ds + Q(x_0,y) \\ &= Q(x,y) - Q(x_0,y) + Q(x_0,y) = Q(x,y). \end{aligned}$$

Also ist die Differentialform (3.5) vollständig. □

BSP. (16.3.2) In der Differentialform

$$(2y^2 + 6xy - x^2)\,dx + (y^2 + 4xy + 3x^2)\,dy = 0$$

gelten mit $P(x,y) := 2y^2 + 6xy - x^2$ und $Q(x,y) := y^2 + 4xy + 3x^2$ sicher die Regularitätseigenschaften $P, Q \in C^2(\mathbf{R}^2)$. Wegen

$$P_y(x,y) = 4y + 6x = Q_x(x,y)$$

liegt also eine vollständige Differentialform vor, und eine Stammfunktion $F(x,y)$ kann gemäß (3.11) berechnet werden. Wir wählen zum Beispiel $(x_0,y_0) = (0,0)$ und erhalten:

$$F(x,y) = \int_0^x (2y^2 + 6sy - s^2)\,ds + \int_0^y (t^2 + 0 \cdot t + 0)\,dt = 2y^2x + 3x^2y - \frac{x^3}{3} + \frac{y^3}{3}.$$

Die allgemeine Lösung ist nun implizit durch die Gleichung $F(x,y) = C =: \frac{1}{3}\,C^*$ gegeben, das heißt durch

$$\boxed{F^*(x,y) := y^3 + 6y^2x + 9x^2y - x^3 = C^*.}$$

(Um zu prüfen, ob hier tatsächlich eine Funktion $y = f(x)$ oder $x = g(y)$ implizit definiert wird, können zum Beispiel die Voraussetzungen zum Satz über implizite Funktionen verifiziert werden. Als Zahlenbeispiel testen wir die Umgebung des Punktes $(1,1)$. Es gilt $F^*(1,1) = 15 =: C^*$, $F_y^*(1,1) = 24 \neq 0$, $F_x^*(1,1) = 21 \neq 0$. Das heißt, in der Umgebung des Punktes $(1,1)$ existiert ein eindeutig bestimmter Zweig der Äquipotentiallinie $F^*(x,y) = 15$, der sowohl eine explizite Darstellung $y = f(x)$ als auch eine explizite Darstellung $x = g(y)$ gestattet.)

Bemerkung 16.5 Der Punkt $(x_0,y_0) \in \Omega$ liegt offenbar auf der Äquipotentiallinie $F(x,y) = 0$ der durch (3.11) definierten Funktion F. Gilt die Voraussetzung (3.10) sowie die Zusatzbedingung

$$P^2(x,y) + Q^2(x,y) > 0 \;\; \forall\, (x,y) \in \Omega, \tag{3.12}$$

so hat die Anfangswertaufgabe

- *Finde diejenige ebene Lösungskurve* Γ *der Differentialform* (3.5)*, die durch den Punkt* $(x_0, y_0) \in \Omega$ *verläuft,*

genau eine Lösung, nämlich $F(x,y) = 0$. Da wegen (3.12) nicht beide Funktionen $F_x(x,y) = P(x,y)$ und $F_y(x,y) = Q(x,y)$ gleichzeitig verschwinden können, gestattet die Gleichung $F(x,y) = 0$ in einer Umgebung des Punktes (x_0, y_0) eine eindeutige Auflösung in der Form $y = f(x)$ (falls $Q(x_0, y_0) \neq 0$) bzw. $x = g(y)$ (falls $P(x_0, y_0) \neq 0$) mit $y_0 = f(x_0)$ bzw. $x_0 = g(y_0)$. □

BSP. (16.3.3) Zu bestimmen ist die ebene Lösungskurve Γ der Differentialform

$$18x\,dx - 8y\,dy$$

durch den Punkt $(x_0, y_0) := (2, 0)$.

Lösung: Hier gelten $P(x,y) = 18x$ und $Q(x,y) = 8y$, so dass wegen $P_y(x,y) = 0 = Q_x(x,y)$ eine vollständige Differentialform auf $\Omega := \mathbf{R}^2$ vorliegt. Die Bedingung (3.12)

$$P^2(x,y) + Q^2(x,y) = (18x)^2 + (8y)^2 > 0$$

ist nur im Punkt $(0,0) \neq (2,0)$ verletzt. Deshalb ist die gesuchte Lösung Γ durch die Funktion (3.11) implizit bestimmt:

$$0 = F(x,y) = \int_2^x 18s\,ds - \int_0^y 8t\,dt = 9x^2 - 4y^2 - 36.$$

Wegen $F_x(2,0) = 36 \neq 0$ kann diese Gleichung lokal nach $x = g(y)$ aufgelöst werden. Man erhält

$$\Gamma = \{(x,y) \,:\, x = g(y) := +\frac{1}{3}\sqrt{36 + 4y^2},\ \ y \in \mathbf{R}\},$$

und das ist der rechte Zweig der Hyperbel $(\frac{x}{2})^2 - (\frac{y}{3})^2 = 1$.

Die Differentialform (3.5) ist nicht mehr vollständig, wenn die Bedingung (3.10) verletzt ist, das heißt, wenn $P_y(x,y) \neq Q_x(x,y)$ in mindestens einem Punkt $(x,y) \in \Omega$ gilt. In diesem Fall kann versucht werden, die Differentialform (3.5) mittels eines **integrierenden Faktors** auf eine vollständige Form zu bringen. Gelten P, Q, $\lambda \in C^1(\Omega)$, so ist die Bedingung $(\lambda P)_y(x,y) = (\lambda Q)_x(x,y)$ gemäß Satz 16.7 **notwendig und hinreichend** für die Vollständigkeit. Das heißt, ein integrierender Faktor $\lambda \in C^1(\Omega)$ muss Lösung der *partiellen Differentialgleichung*

$$\boxed{P(x,y)\frac{\partial \lambda}{\partial y} - Q(x,y)\frac{\partial \lambda}{\partial x} = (Q_x(x,y) - P_y(x,y))\,\lambda} \tag{3.13}$$

sein. Ihre Integration ist kein einfaches Problem, und es soll hier nicht vertieft werden. Bisweilen kann aber ein integrierender Faktor aus der Gleichung (3.13) mit Hilfe spezieller Ansätze gewonnen werden. Wir stellen hier zwei solcher Ansätze vor.

(A) Man sucht λ als Funktion der Variablen x allein: $\lambda = \lambda(x)$. Dazu muss wegen (3.13) die folgende Bedingung gelten, die dann auch zur expliziten Bestimmung von $\lambda(x)$ führt:

$$\boxed{\frac{P_y(x,y) - Q_x(x,y)}{Q(x,y)} =: h(x),\ \text{unabhängig von } y \quad \Rightarrow \quad \lambda' = h(x)\,\lambda, \quad \lambda(x) = e^{\int h(x)\,dx}.}$$

(B) Man sucht λ als Funktion der Variablen y allein: $\lambda = \lambda(y)$. Dazu muss wegen (3.13) die folgende Bedingung gelten, die dann auch zur expliziten Bestimmung von $\lambda(y)$ führt:

$$\boxed{\frac{Q_x(x,y) - P_y(x,y)}{P(x,y)} =: g(y),\ \text{unabhängig von } x \quad \Rightarrow \quad \lambda' = g(y)\,\lambda, \quad \lambda(y) = e^{\int g(y)\,dy}.}$$

Weitere Ansätze wie $\lambda = \lambda(x+y)$ oder $\lambda = \lambda(xy)$ können unter ähnlichen Überlegungen zum Erfolg führen.

BSP. (16.3.4) Der **erste Hauptsatz der Wärmelehre** hat für ein ideales Gas die spezielle Form

$$dQ = nc_v\,dT + \frac{nRT}{V}\,dV.$$

In der Physik ist es bekannt, dass dQ **kein** vollständiges Differential der Wärmemenge Q ist. Das heißt, die Differentialform

$$nc_v\,dT + \frac{nRT}{V}\,dV = 0$$

ist **nicht vollständig**. Ein integrierender Faktor kann mit dem Ansatz $\lambda = \lambda(T)$ bestimmt werden. Mit $P(T,V) := nc_v$ und $Q(T,V) := \frac{nRT}{V}$ prüfen wir die obige Bedingung (Fall (A)) nach:

$$\frac{P_V(T,V) - Q_T(T,V)}{Q(T,V)} = -\frac{nR/V}{nRT/V} = -\frac{1}{T} =: h(T).$$

Somit ist ein integrierender Faktor durch

$$\lambda(T) = e^{-\int \frac{dT}{T}} = \frac{1}{T}$$

explizit bestimmt, und wir erhalten das vollständige Differential einer Funktion $S = S(T,V)$:

$$dS := \frac{dQ}{T} = \frac{nc_v}{T}\,dT + \frac{nR}{V}\,dV.$$

Diese Funktion heißt in der Physik die **Entropie** eines idealen Gases.

BSP. (16.3.5) Für gegebene Funktionen $P(x,y) := xy^3$ und $Q(x,y) := 1 + 2x^2y^2$ ist diejenige Kurvenschar zu bestimmen, die senkrecht zu den Feldlinien des Vektorfeldes $\vec{f}(x,y) := (P(x,y), Q(x,y))^T$ verläuft.

Lösung: Ist $\vec{x}(t) := (x(t), y(t))^T$ eine Parameterdarstellung der gesuchten Kurvenschar, so muss der Tangentenvektor $\dot{\vec{x}}(t)$ in $(x,y) = (x(t), y(t))$ senkrecht auf dem Vektor $\vec{f}(x,y)$ stehen:

$$0 = \langle \vec{f}(x(t), y(t)), \dot{\vec{x}}(t) \rangle = P(x(t), y(t))\frac{dx}{dt} + Q(x(t), y(t))\frac{dy}{dt}.$$

Da diese Gleichung unabhängig in jedem Punkt $\vec{x}(t)$ erfüllt sein soll, erhält man für die gesuchte Orthogonalschar die Differentialform

$$\boxed{P(x,y)\,dx + Q(x,y)\,dy = 0,}$$

also (3.5). Sie ist wegen $P_y(x,y) = 3xy^2 \neq 4xy^2 = Q_x(x,y)$ nicht vollständig. Da sich die Orthogonalitätsbedingung und somit auch die Lösungsschar nicht ändern, wenn das Vektorfeld $\vec{f}(x,y)$ mit einem Skalar $\lambda > 0$ multipliziert wird, verändern wir durch die Verwendung eines integrierenden Faktors nichts an der Aufgabenstellung. Ein solcher Faktor kann mit dem Ansatz $\lambda = \lambda(y)$ bestimmt werden, denn es gilt

$$\frac{Q_x(x,y) - P_y(x,y)}{P(x,y)} = \frac{4xy^2 - 3xy^2}{xy^3} = \frac{1}{y} =: g(y) \quad \Rightarrow \quad \lambda(y) = e^{\int \frac{dy}{y}} = y.$$

Wir erhalten die vollständige Differentialform

$$xy^4\,dx + (y + 2x^2y^3)\,dy = 0,$$

und unter Verwendung von (3.11) ermitteln wir zum Anfangspunkt $(x_0, y_0) := (0, 0)$ eine Stammfunktion

$$F(x,y) = \int_0^x s y^4\, ds + \int_0^y t\, dt = \frac{x^2 y^4}{2} + \frac{y^2}{2}.$$

Somit ist die gesuchte Orthogonalschar implizit durch die folgende Gleichung definiert:

$$\boxed{\frac{1}{2}\, y^2(1 + x^2 y^2) = C \geq 0.}$$

16.4 Differentialgleichungen von Kurvenscharen und singuläre Lösungen

Die bisherige Erfahrung lehrt, dass durch die allgemeine Lösung einer Differentialgleichung 1.Ordnung

$$F(x, y, y') = 0, \qquad F \in \mathrm{Abb}\,(\mathbf{R}^3, \mathbf{R}), \tag{4.1}$$

eine **einparametrige Kurvenschar** explizit oder implizit definiert wird:

$$\varphi(x, y, C) = 0. \tag{4.2}$$

Umgekehrt sei eine einparametrige Kurvenschar in der Form (4.2) mit Scharparameter C und mit einer stetig differenzierbaren Funktion $\varphi \in \mathrm{Abb}\,(\mathbf{R}^3, \mathbf{R})$ vorgelegt. Durch implizite Differentiation resultiert

$$\varphi_x(x, y, C)\, dx + \varphi_y(x, y, C)\, dy = 0. \tag{4.3}$$

Gemäß dem Satz über implizite Funktionen kann der Scharparameter C unter der Voraussetzung $0 \neq \varphi_C(x, y, C)$ zumindest theoretisch aus den beiden Gleichungen (4.2) und (4.3) eliminiert werden, so dass schließlich die Relation (4.3) wieder auf eine Differentialgleichung oder wenigstens auf eine Differentialform 1.Ordnung ohne den Parameter C führt.

BSP. (16.4.1) Wir betrachten die **Parabelschar** $\varphi(x, y, p) := y^2 - 2px = 0$ mit dem Scharparameter $p \in \mathbf{R}$. Hier ist ganz offenkundig die Auflösbarkeit nach $p = \frac{y^2}{2x}$, $x \neq 0$, gewährleistet. Aus der Differentialform (4.3) kann nun p eliminiert werden:

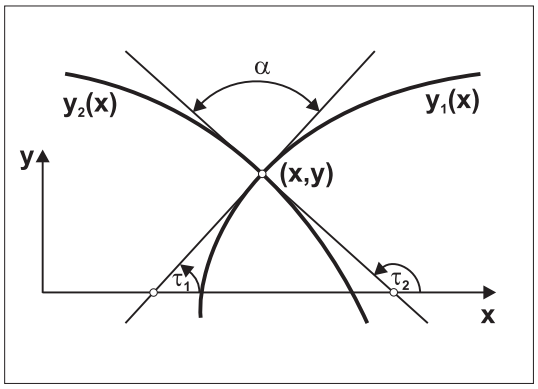
$$0 = \varphi_x(x, y, p)\, dx + \varphi_y(x, y, p)\, dy = -2p\, dx + 2y\, dy = -\frac{y^2}{x}\, dx + 2y\, dy, \qquad x \neq 0.$$

Für $x = 0$ muss $y = 0$ gesetzt werden. Die Parabelschar ist also die allgemeine Lösung der expliziten DGl 1.Ordnung

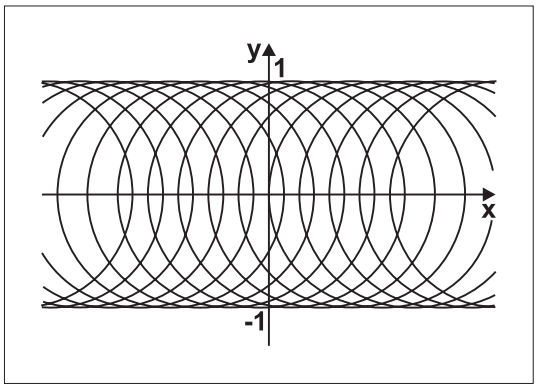
$$\boxed{y' = \frac{y}{2x}, \qquad x \neq 0, \quad y(0) = 0.}$$

Für geometrische Fragestellungen ist manchmal auch die folgende Definition von einiger Relevanz.

Definition 16.8 *Gegeben seien stetige Funktionen $\varphi, \psi \in \mathrm{Abb}\,(\mathbf{R}^3, \mathbf{R})$. Dann heiße die einparametrige Kurvenschar $\psi(x, y, K) = 0$ die* **Schar der isogonalen Trajektorien** *zur Kurvenschar $\varphi(x, y, C) = 0$, wenn sich je zwei Kurven $\varphi(x, y, C_1) = 0$ und $\psi(x, y, K_1) = 0$ in einem gemeinsamen Kurvenpunkt (x, y) stets unter demselben Winkel α schneiden. Für $\alpha = \frac{\pi}{2}$ spricht man insbesondere von der Schar der* **orthogonalen Trajektorien**.



Skizze der Winkelbedingungen für isogonale Trajektorien



Die Enveloppen der Kreisschar aus BSP. (16.4.3)

Wir geben notwendige Winkelbedingungen an, denen die Schar der isogonalen Trajektorien unterliegen muss. Unter der Voraussetzung stetiger Differenzierbarkeit muss gemäß obiger Skizze für je zwei Kurven $y_1(x)$ und $y_2(x)$ aus den beiden Kurvenscharen $\varphi(x,y,C)=0$ bzw. $\psi(x,y,K)=0$ gelten:

$$y_1'(x)=\tan\tau_1, \qquad y_2'(x)=\tan\tau_2, \qquad \alpha=\tau_2-\tau_1.$$

Aus dem Additionstheorem des Tangens folgern wir somit

$$\boxed{m:=\tan\alpha=\frac{\tan\tau_2-\tan\tau_1}{1+\tan\tau_1\tan\tau_2}, \qquad \alpha\neq(n+\tfrac{1}{2})\pi.} \tag{4.4}$$

Es sind zwei Fälle zu unterscheiden gemäß $\alpha\neq\frac{\pi}{2}$ und $\alpha=\frac{\pi}{2}$:

- *Fall* $\alpha\neq\frac{\pi}{2}$*:* Wir setzen $m:=\tan\alpha$ und lösen (4.4) nach $\tan\tau_2$ auf:

$$\tan\tau_2=\frac{m+\tan\tau_1}{1-m\tan\tau_1}.$$

- *Fall* $\alpha=\frac{\pi}{2}$*:* In diesem Fall gilt $\tan\tau_2=\tan(\tau_1+\frac{\pi}{2})=-\frac{1}{\tan\tau_1}$.

Neben diesen Winkelbedingungen muss in einem gemeinsamen Punkt (x,y) die Schnittbedingung $y_1(x)=y_2(x)\equiv y(x)$ erfüllt sein. Somit resultiert:

Satz 16.8 *Gegeben sei eine stetig differenzierbare Funktion* $\varphi\in\mathrm{Abb}\,(\mathbf{R}^3,\mathbf{R})$ *so, dass die einparametrige Kurvenschar* $\varphi(x,y,C)=0$ *allgemeine Lösung der Differentialform*

$$P(x,y)\,dx+Q(x,y)\,dy=0 \tag{4.5}$$

sei. Es sei $\psi(x,y,K)=0$ *die Schar der isogonalen Trajektorien, die die Kurvenschar* $\varphi(x,y,C)=0$ *unter konstantem Winkel* α *schneiden. Dann ist die Schar* $\psi(x,y,K)=0$ *die allgemeine Lösung der folgenden Differentialformen:*

(a) *Im Fall* $\alpha\neq\frac{\pi}{2}$ *und mit der Setzung* $m:=\tan\alpha$*:*

$$\boxed{(P(x,y)-m\cdot Q(x,y))\,dx+(m\cdot P(x,y)+Q(x,y))\,dy=0.} \tag{4.6}$$

(b) *Im Fall* $\alpha=\frac{\pi}{2}$ *der* **orthogonalen Trajektorien***:*

$$\boxed{Q(x,y)\,dx-P(x,y)\,dy=0.} \tag{4.7}$$

BSP. (16.4.2) Wir betrachten nochmals die **Parabelschar** $\varphi(x,y,p) := y^2 - 2px = 0$ aus BSP. (16.4.1), die wir als allgemeine Lösung der (hier gekürzten) Differentialform

$$\frac{dx}{2x} - \frac{dy}{y} = 0$$

erkannt hatten. Die Schar der orthogonalen Trajektorien wird nun gemäß (4.7) durch die Differentialform

$$\frac{dx}{y} + \frac{dy}{2x} = 0$$

beschrieben. Wir integrieren mit dem Verfahren der TdV:

$$y\,dy = -2x\,dx \quad \Rightarrow \quad \frac{y^2}{2} = -x^2 + C \quad \text{oder gleichwertig} \quad \frac{y^2}{2} + x^2 = C \geq 0.$$

Dies ist die Gleichung einer **Ellipsenschar** mit der Normalform $(\frac{x}{\sqrt{C}})^2 + (\frac{y}{\sqrt{2C}})^2 = 1$. Die anderen isogonalen Trajektorien sind Lösungen der homogenen DGl

$$y' = \frac{m + \frac{y}{2x}}{1 - m\frac{y}{x}}, \qquad m := \tan\alpha, \quad \alpha \neq \frac{\pi}{2},$$

die wiederum mit dem Ansatz $y(x) =: x\,u(x)$ gelöst wird.

In einigen Fällen existiert zu einer einparametrigen Kurvenschar $\varphi(x,y,C) = 0$ eine ebene Kurve Γ, die in jedem ihrer Punkte $(x_0,y_0) \in \Gamma$ eine Kurve der gegebenen Schar berührt. Wir belegen dies durch folgendes Beispiel.

BSP. (16.4.3) Die **Kreisschar** $\varphi(x,y,C) := (x-C)^2 + y^2 - 1 = 0$ mit Mittelpunkt $(C,0)$ und Radius 1 wird von den beiden Geraden $y = \pm 1$ berührt.

Definition 16.9 *Gegeben sei eine einparametrige Kurvenschar $\varphi(x,y,C) = 0$ mit glatter Funktion $\varphi \in \mathrm{Abb}\,(\mathbf{R}^3, \mathbf{R})$. Eine ebene Kurve Γ, von der jeder Punkt ein Berührungspunkt für mindestens eine Kurve der gegebenen Schar ist, und umgekehrt jedes Teilstück $\Gamma' \subset \Gamma$ von unendlich vielen Kurven der Schar $\varphi(x,y,C) = 0$ berührt wird, heiße eine* **Einhüllende** *oder* **Enveloppe** *der gegebenen Kurvenschar.*

Eine Berechnungsvorschrift für Enveloppen einer gegebenen Kurvenschar wird im folgenden Satz angegeben.

Satz 16.9 *Es seien eine Teilmenge $\Omega \subset \mathbf{R}^3$, eine skalare Funktion $\varphi \in C^2(\Omega)$ und ein innerer Punkt $(x_0,y_0,C_0) \in \Omega$ gegeben. Sind die Bedingungen*

$$\boxed{\varphi(x_0,y_0,C_0) = 0, \qquad \frac{\partial\varphi}{\partial C}(x_0,y_0,C_0) = 0, \qquad \frac{\partial^2\varphi}{\partial C^2}(x_0,y_0,C_0) \neq 0} \tag{4.8}$$

erfüllt, so besitzt die einparametrige Kurvenschar $\varphi(x,y,C) = 0$ in einer Umgebung U des Punktes (x_0,y_0) eine Enveloppe Γ in impliziter Darstellung $f(x,y) = 0$ mit $f \in C^1(U)$. Die Funktion f erhält man durch Elimination des Parameters C aus der **Enveloppenbedingung**

$$\boxed{\varphi(x,y,C) = 0, \qquad \frac{\partial\varphi}{\partial C}(x,y,C) = 0.} \tag{4.9}$$

Begründung: Der Satz über implizite Funktionen sichert wegen $\varphi_{CC}(x_0, y_0, C_0) \neq 0$ die Existenz einer Umgebung $U \subset \mathbf{R}^2$ des Punktes (x_0, y_0) und einer differenzierbaren Funktion $g : U \to \mathbf{R}$ mit

$$g(x_0, y_0) = C_0, \quad \varphi_C(x, y, g(x, y)) = 0, \qquad (x, y) \in U.$$

Die Kurve

$$f(x, y) := \varphi(x, y, g(x, y)) = 0, \qquad (x, y) \in U,$$

leistet nun das Verlangte: Es gilt in der Tat $\varphi(x_0, y_0, g(x_0, y_0)) = 0$, und man berechnet aus der Relation $\varphi_x(x_0, y_0, C_0)\, dx + \varphi_y(x_0, y_0, C_0)\, dy = 0$ im Punkt (x_0, y_0) die Tangentensteigung

$$y'(x_0) = -\varphi_x(x_0, y_0, C_0)/\varphi_y(x_0, y_0, C_0)$$

der Kurvenschar. Die Tangentensteigung der Enveloppe $f(x, y) = 0$ berechnet sich in (x_0, y_0) gemäß

$$\begin{aligned} 0 &= f_x(x_0, y_0)\, dx + f_y(x_0, y_0)\, dy \\ &= (\varphi_x(x_0, y_0, C_0) + \underbrace{\varphi_C(x_0, y_0, C_0)}_{=0} g_x(x_0, y_0))\, dx + (\varphi_y(x_0, y_0, C_0) + \underbrace{\varphi_C(x_0, y_0, C_0)}_{=0} g_y(x_0, y_0))\, dy \\ &= \varphi(x_0, y_0, C_0)\, dx + \varphi_y(x_0, y_0, C_0)\, dy, \end{aligned}$$

genau wie die Tangentensteigung der gegebenen Kurvenschar. □

Bemerkung 16.6 Man erhält aus der Enveloppenbedingung (4.9) auch dann die richtige Enveloppe, wenn die Berührungspunkte **singuläre Punkte** der Kurvenschar $\varphi(x, y, C) = 0$ sind, wenn also $\varphi_x(x_0, y_0, C_0) = 0 = \varphi_y(x_0, y_0, C_0)$ gelten. □

BSP. (16.4.4) Auf der Menge $\Omega := \mathbf{R}^3$ betrachten wir die Funktion $\varphi(x, y, C) := (C-y)^2 + (C-x)^3$. Durch $\varphi(x, y, C) = 0$ wird eine Schar NEILscher Parabeln definiert, deren Spitzen (= singuläre Punkte) in $(x_0, y_0) = (C, C)$ liegen. Die Enveloppenbedingung (4.9) führt hier auf das Gleichungspaar

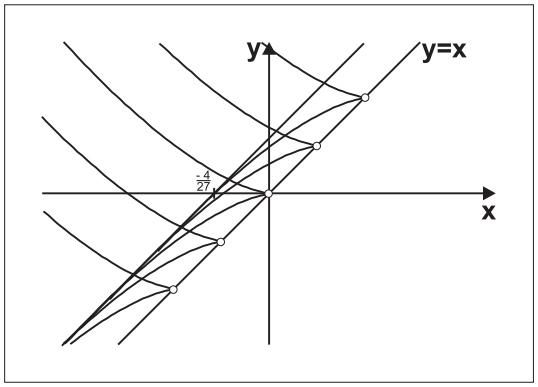
$$\varphi(x, y, C) = (C - y)^2 - (C - x)^3 = 0, \quad \frac{\partial \varphi}{\partial C}(x, y, C) = 2(C - y) - 3(C - x)^2 = 0,$$

aus dem zunächst der Term $C - y$ eliminiert werden kann:

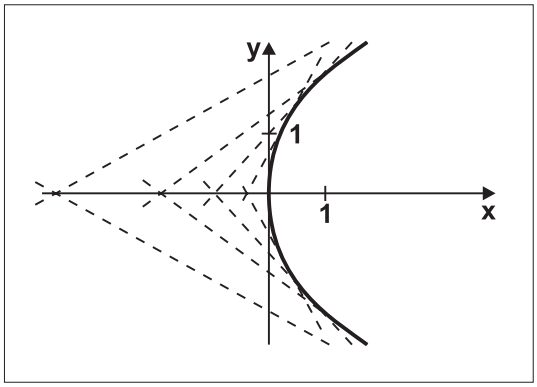
$$(C - x)^3 \Big(\frac{9}{4} (C - x) - 1 \Big) = 0.$$

Das heißt,

- entweder gilt $C = x$, und man erhält die Gerade $\boxed{y(x) = x}$ als Enveloppe; auf ihr liegen genau die singulären Punkte der NEILschen Parabeln,
- oder es gilt $C = \frac{4}{9} + x$, und man erhält die Gerade $\boxed{y(x) = x + \frac{4}{27}}$ als zweite Enveloppe mit echten Berührungspunkten.



Die Enveloppen der NEILschen Parabelschar $(C - x)^3 = (C - y)^2$



Allgemeine und singuläre Lösung der CLAIRAUT–DGl aus BSP. (16.4.6)

Ist die einparametrige Kurvenschar $\varphi(x,y,C)=0$ die allgemeine Lösung einer Differentialgleichung $F(x,y,y')=0$, so muss eine eventuell existierende Enveloppe Γ der gegebenen Kurvenschar ebenfalls Lösung dieser DGl sein: In jedem Punkt $(x,y)\in\Gamma$ hat nämlich die Enveloppe die durch die Gleichung $F(x,y,y')=0$ vorgeschriebene Tangentensteigung $y'(x)$. Offenbar gehen durch jeden Punkt der Enveloppe mindestens zwei Lösungen der DGl $F(x,y,y')=0$, so dass die Anfangswertaufgabe in diesem Punkt nicht mehr eindeutig lösbar sein darf. Die Enveloppe nimmt als Lösung also eine Sonderstellung ein:

Definition 16.10 *Auf einer Teilmenge $\Omega\subset\mathbf{R}^3$ sei eine skalare Funktion $F\in C^1(\Omega)$ gegeben. Ein* **Linienelement** *der DGl $F(x,y,y')=0$ ist ein Punkt $(x_0,y_0,y_0')\in\Omega$ mit $F(x_0,y_0,y_0')=0$.*

Ein Linienelement $(x_0,y_0,y_0')\in\Omega$ heiße **singulär**, *wenn gilt:*

$$\boxed{\frac{\partial F}{\partial y'}(x_0,y_0,y_0')=0;}$$

sonst heiße das Linienelement **regulär**.

Eine Lösung $y(x)$ der DGl $F(x,y,y')=0$ heiße **singulär** *bzw.* **regulär**, *wenn diese ausschließlich singuläre bzw. reguläre Linienelemente enthält.*

BSP. (16.4.5) Wir betrachten die implizite DGl

$$F(x,y,y'):=y'^2-(1+x^2)y^2=0.$$

Wegen $\frac{\partial F}{\partial y'}(x,y,y')=2y'$ gibt es singuläre Linienelemente höchstens für $y_0'=0$, wenn auch noch $F(x_0,y_0,0)=-(1+x_0^2)y_0^2=0$ erfüllt ist. Das trifft genau für die Punkte $(x_0,0,0)$ zu, $x_0\in\mathbf{R}$, und die Gerade $y(x)=0$ ist singuläre Lösung der DGl.

Als wichtigstes Beispiel einer Differentialgleichung mit singulären Linienelementen gilt die Clairautsche **Differentialgleichung**

$$\boxed{y=xy'+h(y'),} \tag{4.10}$$

worin $h\in\mathrm{Abb}\,(\mathbf{R},\mathbf{R})$ eine **stetig differenzierbare** Funktion sei.

Satz 16.10 *Gegeben seien ein Intervall $I\subset\mathbf{R}$ und eine skalare Funktion $h\in C^1(I)$.*

(a) *Die regulären Lösungen der* Clairautschen *DGl* (4.10) *sind genau die Geraden*

$$\boxed{y(x):=Cx+h(C),\qquad C\in I,} \tag{4.11}$$

eventuell mit Ausnahme eines einzelnen singulären Linienelements. Sie bilden die allgemeine Lösung.

(b) *Besitzt die einparametrige Geradenschar* (4.11) *eine Enveloppe Γ, so ist diese eine singuläre Lösung der DGl* (4.10). *Die Enveloppe Γ hat dann die folgende Parameterdarstellung:*

$$\boxed{\vec{x}(t)=\begin{bmatrix}-h'(t)\\ h(t)-th'(t)\end{bmatrix},\qquad t\in I.} \tag{4.12}$$

Begründungen: (a) Aus der Darstellung (4.11) der Geradenschar erhalten wir $y'(x)=C$, und somit durch Einsetzen in (4.10): $y=Cx+h(C)$. Das heißt, die Geradenschar (4.11) ist Lösung der Clairautschen DGl. Setzen wir $F(x,y,y'):=y-xy'-h(y')$, so kann auf der Geraden (4.11) wegen $F_{y'}(x,y,y')=-x-h'(y')$ höchstens für $x=-h'(C)$ ein singuläres Linienelement auftreten.

(b) Wir setzen $\varphi(x,y,C):=y-Cx-h(C)$, so dass die Enveloppenbedingung (4.9) hier in der folgenden Form vorliegt:

$$\varphi(x,y,C)=y-Cx-h(C)=0,\qquad \frac{\partial\varphi}{\partial C}(x,y,C)=-x-h'(C)=0.$$

Wählt man $t := C \in I$ als Parameter, so resultiert die behauptete Parameterdarstellung (4.12) der Enveloppe. Gemäß (a) sind alle ihre Linienelemente singulär. □

BSP. (16.4.6) Wir betrachten die CLAIRAUTsche DGl

$$y = xy' + \frac{1}{y'}, \qquad y' \neq 0.$$

Es gilt hier $h(y') := \frac{1}{y'}$, und die Voraussetzungen von Satz 16.10 sind auf jedem der beiden Intervalle $I := (0, +\infty)$ und $I := (-\infty, 0)$ erfüllt. Die Geradenschar

$$\boxed{y(x) = Cx + \frac{1}{C}, \qquad C \neq 0,}$$

bildet die allgemeine Lösung. Es existiert eine singuläre Lösung, die wir durch Elimination von C aus der Enveloppenbedingung

$$\varphi(x, y, C) := y - Cx - \frac{1}{C} = 0, \qquad \frac{\partial \varphi}{\partial C}(x, y, C) = -x + \frac{1}{C^2} = 0$$

gewinnen, nämlich

$$\boxed{y^2(x) = 4x, \qquad x \geq 0.}$$

Ein spezieller Fall singulärer Linienelemente tritt in der Differentialform

$$P(x, y)\, dx + Q(x, y)\, dy = 0 \tag{4.13}$$

an den gemeinsamen Nullstellen der Funktionen $P(x, y)$ und $Q(x, y)$ auf. Es sei daran erinnert, dass die Differentialform (4.13) in solchen Punkten nicht mehr in eine explizite DGl überführt werden kann.

Definition 16.11 *Ein Punkt $(x_0, y_0) \in D(P) \cap D(Q)$ heiße* **singulärer Punkt** *der Differentialform (4.13), wenn $P(x_0, y_0) = 0 = Q(x_0, y_0)$ gelten. Ein singulärer Punkt (x_0, y_0) heiße* **isoliert**, *wenn es eine offene δ–Kugel $B_\delta(x_0, y_0)$ gibt mit*

$$P^2(x, y) + Q^2(x, y) > 0 \;\; \forall\, (x, y) \in B_\delta(x_0, y_0) \setminus \{(x_0, y_0)\}.$$

Schreibt man die Differentialform (4.13) als implizite Gleichung $F(x, y, y') := P(x, y) + Q(x, y)\, y' = 0$ auf, so erkennt man an der Relation $F_{y'}(x, y, y') = Q(x, y)$, dass in den singulären Punkten (x_0, y_0) der Differentialform (4.13) singuläre Linienelemente (x_0, y_0, y') für jede Wahl von $y' \in \mathbf{R}$ liegen müssen. Wir diskutieren einige typische isolierte singuläre Punkte der Differentialform (4.13) im folgenden Beispiel.

BSP. (16.4.7) Wir betrachten die **gebrochen–lineare Differentialgleichung**

$$\boxed{y' = \frac{ax + by}{cx + dy}, \qquad ad - bc \neq 0.} \tag{4.14}$$

(Gilt $ad - bc = 0$, so sind Zähler und Nenner linear abhängig; durch Kürzung erhält man die DGl $y' = A = const$ mit der Lösung $y(x) = Ax + B$. Dieser Fall bedarf keiner weiteren Diskussion.)

Durch die Setzung $P(x, y) := ax + by$ und $Q(x, y) := -(cx + dy)$ kann die DGl (4.14) mit der Differentialform (4.13) identifiziert werden. Offensichtlich ist der Punkt $(x_0, y_0) = (0, 0)$ der einzige singuläre Punkt. Da die DGl (4.14) vom homogenen Typ ist, kann sie vermöge der Transformation $y(x) =: x\, u(x)$, $y'(x) = u(x) + x\, u'(x)$ wieder in eine DGl mit getrennten Variablen überführt werden:

$$\boxed{xu' = \frac{a + (b - c)u - d\, u^2}{c + d\, u}.} \tag{4.15}$$

Ihre Integration erfolgt mit dem Verfahren der TdV. Dabei wird die Anzahl der Nullstellen des quadratischen Polynoms

$$\Delta(u) := a+(b-c)u-du^2 = -d\Big(u-\frac{b-c+\sqrt{D}}{2d}\Big)\cdot\Big(u-\frac{b-c-\sqrt{D}}{2d}\Big), \qquad D := (b-c)^2+4ad, \quad (4.16)$$

eine entscheidende Rolle spielen, wobei das Vorzeichen der Diskriminante D angibt, ob reelle oder komplexe Nullstellen vorliegen. Wir setzen

$$u_\pm := \frac{1}{2d}\,(b-c\pm\sqrt{D}) \qquad (4.17)$$

und treffen drei Fallunterscheidungen gemäß $D=0,\ D>0,\ D<0$.

Fall 1: Es gelte $D=0$. Wir unterscheiden zwei Unterfälle (i) und (ii) gemäß

(i) $a=d=0$: Aus $D=0$ folgt zwangsläufig $b-c=0$ und somit $u'=0$ mit der allgemeinen Lösung $u(x)=C=\frac{y(x)}{x}=const.$ Somit ist die **Geradenschar** durch den singulären Punkt $(0,0)$

$$\boxed{y(x)=Cx, \qquad C\in\mathbf{R},}$$

die allgemeine Lösung der DGl (4.14). Der Punkt $(0,0)$ heißt in diesem Fall ein **Knoten 1.Art**.

(ii) $d\neq 0$: Das quadratische Polynom $\Delta(u)$ hat die doppelte Nullstelle $u_0 := \frac{b-c}{2d}$, und man erhält durch Partialbruchzerlegung

$$\frac{c+du}{\Delta(u)} = -\frac{1}{u-u_0}-\frac{\delta}{(u-u_0)^2}, \qquad \delta := \frac{b+c}{2d}\neq 0.$$

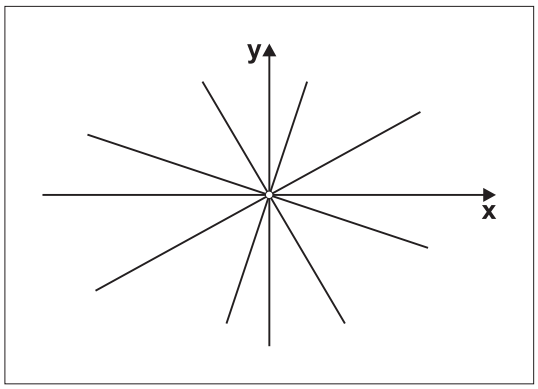
Nun lässt sich die Gleichung (4.15) mit dem Verfahren der TdV integrieren:

$$\frac{\delta}{u-u_0} = \ln|x(u-u_0)|+C \quad \text{bzw.} \quad \frac{\delta x}{y-u_0x} = \ln|y-u_0x|+C.$$

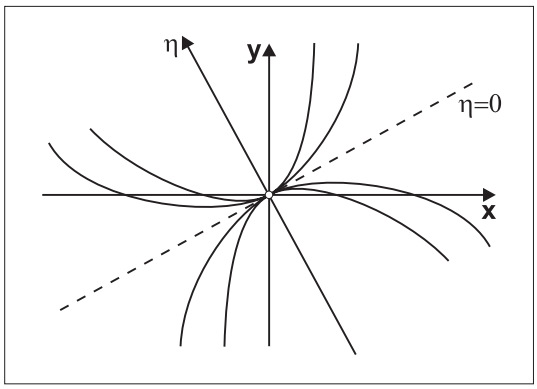
Wegen $\Delta(u)=0$ ist es klar, dass auch die Gerade $u=u_0$ eine Lösung von (4.15) ist, bzw. die Gerade $y(x)=u_0x$ eine Lösung von (4.14). Diese Lösung heiße die **Hauptgerade** der DGl (4.14). Eine *Transformation auf die Hauptgerade* wird durch Einführung einer neuen Koordinate $\eta := y-u_0x$ bewerkstelligt. Die Lösungsschar durch den singulären Punkt $(0,0)$ hat nun die Darstellung

$$\boxed{x(\eta) = \frac{\eta}{\delta}\,(\ln|\eta|+C), \qquad C\in\mathbf{R}.}$$

Der singuläre Punkt $(0,0)$ heißt in diesem Fall ein **Knoten 2.Art**



Knoten 1.Art: $D=0,\ a=0=d$



Knoten 2.Art: $D=0,\ d\neq 0$

Fall 2: Es gelte $D > 0$. Wir unterscheiden wieder zwei Unterfälle (i) und (ii) gemäß

(i) $d \neq 0$: Das quadratische Polynom $\Delta(u) = -d(u-u_+)(u-u_-)$ hat die zwei verschiedenen reellen Nullstellen $u_\pm$. Man erhält nun durch Partialbruchzerlegung

$$\frac{c+du}{\Delta(u)} = \frac{\alpha}{u-u_+} + \frac{\beta}{u-u_-}, \qquad \alpha := -\frac{1}{2} - \frac{b+c}{2\sqrt{D}}, \quad \beta := -\frac{1}{2} + \frac{b+c}{2\sqrt{D}},$$

und es gelten die Relationen

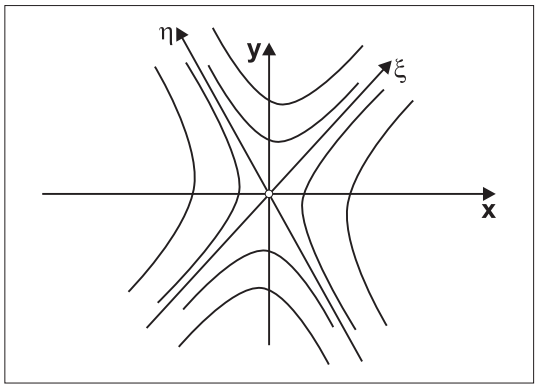
$$\alpha + \beta = -1, \qquad \alpha\beta D = ad - bc \neq 0.$$

Wir können jetzt die Gleichung (4.15) mit dem Verfahren der TdV integrieren:

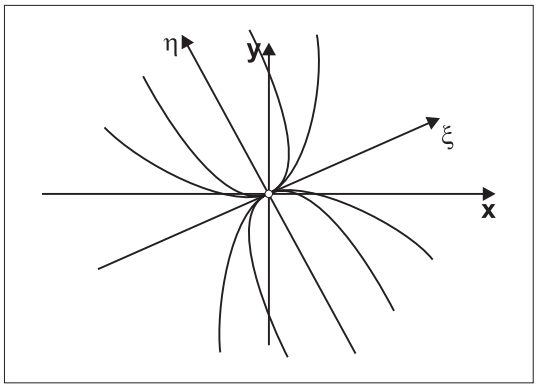
$$|u-u_+|^\alpha |u-u_-|^\beta = C|x| \quad \text{bzw.} \quad |y-u_+x|^\alpha |y-u_-x|^\beta = C > 0.$$

Wegen $\Delta(u_\pm) = 0$ treten hier die beiden **Hauptgeraden** $y(x) = u_\pm x$ ebenfalls als Lösungen der DGl (4.14) auf. Eine Transformation auf diese Hauptgeraden bewerkstelligt man vermöge der Substitutionen $\xi := y - u_+x$ und $\eta = y - u_-x$. Die Lösungsschar beim singulären Punkt $(0,0)$ hat nun die Darstellung

$$\boxed{\eta(\xi) = \begin{cases} C_1|\xi|^{-|\alpha/\beta|} & : \quad \alpha\beta D = ad-bc > 0, \quad \textbf{Sattelpunkt}, \\ C_2|\xi|^{+|\alpha/\beta|} & : \quad \alpha\beta D = ad-bc < 0, \quad \textbf{Knoten 2.Art}. \end{cases}}$$



Sattelpunkt: $D > 0,\ ad - bc > 0$



Knoten 2.Art: $D > 0,\ ad - bc < 0$

(ii) $d = 0$: Nun gilt $\Delta(u) = (b-c)(u-u_0)$ mit $u_0 = \frac{a}{c-b}$, $b - c \neq 0$. Es existiert nur eine einfache Nullstelle $u = u_0$, und dieser entspricht wieder die **Hauptgerade** $y(x) = u_0 x$ als Lösung der DGl (4.14). Darüber hinaus erhält man durch Trennung der Veränderlichen:

$$\frac{c\,du}{u-u_0} = \frac{(b-c)\,dx}{x} \quad \Rightarrow \quad |u-u_0|^c = C\,|x|^{b-c} \quad \Rightarrow \quad |y-u_0x|^c = C\,|x|^b.$$

Wir führen wiederum durch Setzen von $\eta := y - u_0 x$ eine Transformation auf die Hauptgerade durch, und erhalten schließlich die Lösungsschar

$$\boxed{\eta(x) = \begin{cases} C_1|x|^{-|b/c|} & : \quad ad-bc = -bc > 0, \quad \textbf{Sattelpunkt}, \\ C_2|x|^{+|b/c|} & : \quad ad-bc = -bc < 0, \quad \textbf{Knoten 2.Art}. \end{cases}}$$

Fall 3: Es gelte $D < 0$, und dies kann nur für $ad < 0$ eintreten. Das quadratische Polynom $\Delta(u)$ besitzt nun ein Paar konjugiert komplexer Wurzeln:

$$\Delta(u) = -d(u-u_0)(u-\overline{u_0}); \qquad u_0 := \alpha + i\,\beta, \quad \alpha := \frac{b-c}{2d}, \quad \beta := \frac{\sqrt{-D}}{2d}.$$

Wie in den vorangegangenen Fällen gewinnt man durch Partialbruchzerlegung und anschließender Integration die Lösungsschar der DGl (4.14), und zwar in der Form

$$\ln|y^2 - 2\alpha xy + (\alpha^2+\beta^2)x^2| + \frac{b+c}{\beta d}\,\mathrm{arc\,tan}_H\,\frac{y-\alpha x}{\beta x} + C = 0.$$

Gilt $b+c=0$, so resultiert eine **Ellipsenschar**

$$\boxed{y^2 - 2\alpha xy + (\alpha^2+\beta^2)x^2 = e^{-C}.}$$

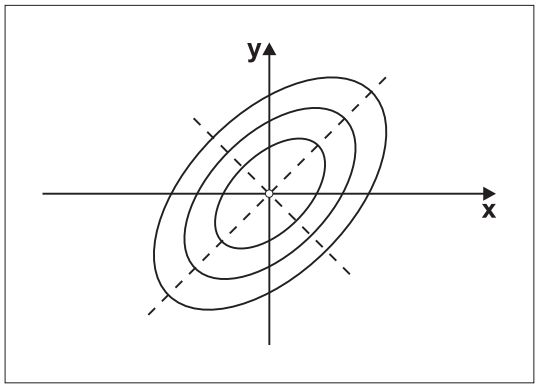
Der singuläre Punkt $(0,0)$ heiße in diesem Fall ein **Wirbelpunkt**. Für $b+c\neq 0$ transformieren wir auf neue Koordinaten $\xi := \beta x,\ \eta := y-\alpha x$:

$$\ln(\xi^2+\eta^2) + \frac{b+c}{\beta d}\,\mathrm{arc\,tan}_H\,\frac{\eta}{\xi} + C = 0.$$

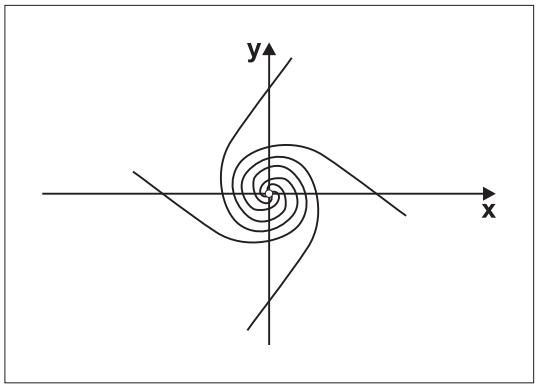
Unter Verwendung von Polarkoordinaten $\xi = r\cos\varphi,\ \eta = r\sin\varphi$ erkennt man, dass die Lösungskurven eine Schar **logarithmischer Spiralen** sind:

$$\boxed{r(\varphi) = C_1 \exp\Big(-\frac{b+c}{2\beta d}\,\varphi\Big), \qquad \varphi\in\mathbf{R}.}$$

Der singuläre Punkt $(0,0)$ heiße in diesem Fall ein **Strudelpunkt**.



Wirbelpunkt: $D<0,\ b+c=0$



Strudelpunkt: $D<0,\ b+c\neq 0$

Fall 4: Es gelte $b-c=0=d$ und somit $\Delta(u)=a$ sowie $D=0$; das heißt, das quadratische Polynom $\Delta(u)$ besitzt keine Nullstelle. Sicher muss $c\neq 0$ sein, so dass die DGl (4.15) auf die Form $xu' = \frac{a}{c}$ schrumpft mit der offensichtlichen Lösung

$$u(x) = \frac{a}{c}\,\ln|x| + C \quad\Rightarrow\quad y(x) = x\Big(\frac{a}{c}\,\ln|x| + C\Big).$$

In diesem Fall ist der singuläre Punkt $(0,0)$ wieder ein **Knoten 2.Art**.

Wir stellen nun in einer **verkürzten Übersicht** die Kriterien zusammen, die den Typ des singulären Punktes $(0,0)$ der gebrochen–linearen DGl charakterisieren:

$$\boxed{y' = \frac{ax+by}{cx+dy}, \qquad ad-bc\neq 0, \quad D := (b-c)^2 + 4ad.}$$

Typenkatalog des singulären Punktes $(0,0)$		
$D>0$	$ad-bc>0$	**Sattelpunkt**
	$ad-bc<0$	**Knotenpunkt 2.Art**
$D=0$	$a=d=0$	**Knotenpunkt 1.Art**
	$d\neq 0$ oder $a\neq 0$	**Knotenpunkt 2.Art**
$D<0$	$b+c=0$	**Wirbelpunkt**
	$b+c\neq 0$	**Strudelpunkt**

16.5 Existenz– und Eindeutigkeitsfragen

Die Frage nach der korrekten Stellung der Anfangswertaufgabe

$$y' = f(x,y), \qquad y(x_0) = y_0, \tag{5.1}$$

kann in sehr speziellen Fällen durch die Angabe einer expliziten Darstellung der Lösungsgesamtheit der DGl mittels Integralen beantwortet werden, aus der dann diejenigen Integralkurven selektiert werden, die durch den Anfangspunkt (x_0, y_0) verlaufen. Bereits in einfachen Fällen kann die Lösungsgesamtheit jedoch nicht mehr in geschlossener Form analytisch bestimmt werden. Hier sei als Beispiel die DGl $y' = x^2 + y^2$ genannt. Deshalb ist es von Interesse, allgemeine Existenzaussagen bereitzustellen, die wenigstens eine theoretische Garantie der Existenz von Lösungen geben. Zur Orientierung betrachten wir das folgende Beispiel.

BSP. (16.5.1) Es sei die DGl $y' = \frac{y}{x}$, $x \neq 0$, aus BSP. (16.2.1) vorgelegt, für die bereits gezeigt wurde, dass die Anfangswertaufgabe (5.1) stets genau eine Lösung $y(x) = \frac{y_0}{x_0}\, x$ hat, sofern $x_0 \neq 0$ vorausgesetzt wird. Für $x_0 = 0$ und $y_0 \neq 0$ existiert überhaupt keine Lösung der AWA (5.1), während für $x_0 = 0$ und $y_0 = 0$ jede Kurve $y(x) = Cx$, $C \in \mathbf{R}$, eine Lösung ist. In den Unstetigkeitspunkten $(x_0, y_0) := (0, y_0)$ der rechten Seite $f(x,y) = \frac{y}{x}$ ist also im allgemeinen weder Existenz noch Eindeutigkeit der AWA (5.1) gewährleistet.

Hingegen gilt die folgende Existenzaussage:

Satz 16.11 (Existenzsatz von Peano)
Gegeben seien für ein festes $a > 0$ ein Streifen

$$\bar{G} := \{(x,y) \,:\, x_0 \leq x \leq x_0 + a \text{ und } y \in \mathbf{R}\} \subset \mathbf{R}^2$$

sowie eine skalare Funktion $f \in C(\bar{G})$, die auf $\bar{G}$ beschränkt ist: $|f(x,y)| \leq M < \infty \;\forall\; (x,y) \in \bar{G}$. Dann hat die AWA (5.1) für jede Vorgabe $y_0 \in \mathbf{R}$ **mindestens eine Lösung** *$y = y(x)$, $x \in [x_0, x_0 + a]$.*

Wir verzichten hier auf einen Beweis dieses Satzes, da er das Problem der korrekten Stellung nur im Existenzteil beantwortet. In der Tat müssen Lösungen der AWA (5.1) unter der sehr schwachen Voraussetzung der Stetigkeit der Funktion $f(x,y)$ **nicht eindeutig** sein.

Die Funktion $f(x,y) := 2x\sqrt{|y|}$ erfüllt sicher $f \in C(\mathbf{R}^2)$, während die Anfangswertaufgabe

$$y' = f(x,y), \qquad y(0) = 0,$$

zwei Lösungen besitzt, nämlich $y_1(x) := \frac{x^4}{4}$ und $y_2(x) := 0$.

Die Regularitätsvoraussetzung an f soll nun soweit verschärft werden, dass gleichzeitig Existenz und Eindeutigkeit von Lösungen der AWA (5.1) gewährleistet sind. Wir verallgemeinern die Problemstellung, indem wir anstelle von (5.1) die folgende Anfangswertaufgabe betrachten:

$$y^{(n)} = f(x, y, y', \ldots, y^{(n-1)}), \quad y(x_0) = y_0,\; y'(x_0) = y_1, \ldots, y^{(n-1)}(x_0) = y_{n-1}. \tag{5.2}$$

Eine AWA vom Typ (5.2) kann stets auf eine Anfangswertaufgabe für ein **System von DGln** 1.Ordnung zurückgeführt werden:

Satz 16.12 *Die AWA (5.2) ist der folgenden Anfangswertaufgabe für ein System von DGln 1.Ordnung äquivalent:*

$$\boxed{\vec{y}\,'(x) = \vec{f}(x, \vec{y}(x)), \qquad \vec{y}(x_0) = \vec{y}_0.} \tag{5.3}$$

Dabei gelten die folgenden Bezeichnungen

$$\vec{y}(x) := \begin{bmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_n(x) \end{bmatrix}, \qquad \vec{y}_0 := \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix}, \qquad \vec{f}(x,\vec{y}) := \begin{bmatrix} f_1(x,\vec{y}) \\ f_2(x,\vec{y}) \\ \vdots \\ f_n(x,\vec{y}) \end{bmatrix}.$$

Die Äquivalenz versteht sich im folgenden Sinne: Jeder Lösung $y(x)$ der AWA (5.2) mit der Regularität $y \in C^n$ entspricht eine Lösung $\vec{y} \in C^1$ der AWA (5.3) (in der speziellen Form (5.4)) und umgekehrt.

Begründung: Ist $y \in C^n$ Lösung der AWA (5.2), so erhält man durch Einführen neuer abhängiger Variabler

$$\boxed{y_j(x) := y^{(j-1)}(x), \quad j = 1, 2, \ldots, n,}$$

eine spezielle Form des Systems (5.3), nämlich

$$\begin{array}{lcl} y_1' & = & y_2 =: f_1(x,\vec{y}), \\ y_2' & = & y_3 =: f_2(x,\vec{y}), \\ \vdots & & \\ y_{n-1}' & = & y_n =: f_{n-1}(x,\vec{y}), \\ y_n' & = & y^{(n)} = f(x, y_1, y_2, \ldots, y_n) =: f_n(x,\vec{y}). \end{array} \tag{5.4}$$

Darüber hinaus gilt $\vec{y}(x_0) = \vec{y}_0$. Ist umgekehrt $\vec{y} \in C^1$ eine Lösung der AWA (5.3) in der speziellen Form (5.4), so setze man $y(x) := y_1(x)$. Dann gilt $y \in C^n$, und es folgt (5.2). □

Bemerkung 16.7 Die Aufgabe (5.3) ist allgemeiner als die Aufgabe (5.2); nur in der speziellen Form (5.4) sind (5.2) und (5.3) äquivalent. □

BSP. (16.5.2) Die Anfangswertaufgabe für eine **lineare DGl** n–ter Ordnung mit konstanten Koeffizienten, nämlich

$$y^{(n)} = \sum_{k=0}^{n-1} a_k y^{(k)} + g(x), \qquad y(x_0) = y_0,\ y'(x_0) = y_1, \ldots, y^{(n-1)}(x_0) = y_{n-1},$$

ist mit der folgenden Anfangswertaufgabe für ein **lineares System** 1.Ordnung äquivalent:

$$\vec{y}\,'(x) = A\vec{y}(x) + \vec{f}(x), \quad \vec{y}(x_0) = \vec{y}_0 \quad \text{mit } A := \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & & & 0 \\ 0 & 0 & 0 & \ddots & & 0 \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & & & 1 \\ a_0 & a_1 & a_2 & \cdots & \cdots & a_{n-1} \end{bmatrix}, \quad \vec{f}(x) := \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ g(x) \end{bmatrix},$$

man vergleiche auch Abschnitt 11.1.

Die Existenz von Lösungen zur AWA (5.3) kann sicher nur unter den Mindestvoraussetzungen des PEANO–Satzes erwartet werden. Dazu betrachten wir für festes $x_0 \in \mathbf{R}$, $\vec{y}_0 \in \mathbf{R}^n$ und $a, b > 0$ den **Zylinder**

$$G_b := \{(x,\vec{y}) \,:\, x_0 \le x \le x_0 + a \text{ und } \|\vec{y} - \vec{y}_0\| < b\} \subset \mathbf{R}^{n+1}. \tag{5.5}$$

Darin darf auch $b = \infty$ zugelassen sein. Ferner gelte die Stetigkeitsvoraussetzung

$$\vec{f} \in C(G_b; \mathbf{R}^n). \tag{5.6}$$

Ist $\vec{y} \in C^1([x_0, x_0 + a]; \mathbf{R}^n)$ eine Lösung der AWA (5.3), so kann die vektorielle Differentialgleichung $\vec{y}\,'(x) = \vec{f}(x, \vec{y}(x))$ unter Beachtung der Anfangsbedingung $\vec{y}(x_0) = \vec{y}_0$ **komponentenweise** über dem Intervall $[x_0, x]$ aufintegriert werden. Zusammengefasst resultiert die vektorielle Gleichung

$$\boxed{\vec{y}(x) = \vec{y}_0 + \int_{x_0}^{x} \vec{f}(s, \vec{y}(s))\, ds, \qquad x \in [x_0, x_0 + a].} \tag{5.7}$$

Definition 16.12 *Die Gleichung* (5.7) *heiße eine* **Integralgleichung** *für die gesuchte Funktion* $\vec{y}(x)$*. Die Bestimmung einer vektorwertigen Funktion* $\vec{y} \in C([x_0, x_0 + a]; \mathbf{R}^n)$*, die die Integralgleichung* (5.7) *in jedem Punkt* $x \in [x_0, x_0 + a]$ *erfüllt, heiße das* **Integralgleichungsproblem**.

Es gilt nun der folgende Äquivalenzsatz:

Satz 16.13 *Unter der Voraussetzung* $\vec{f} \in C(G_b; \mathbf{R}^n)$ *sind das Integralgleichungsproblem* (5.7) *und die Anfangswertaufgabe* (5.3) *miteinander äquivalent.*

Begründung: Wir haben lediglich noch zu zeigen, dass **stetige** Lösungen der Integralgleichung (5.7) auch Lösungen der AWA (5.3) sind. Da $\vec{f}$ als stetig vorausgesetzt ist, muss die rechte Seite der Gleichung (5.7) sogar stetig differenzierbar sein. Somit ist wegen der Gleichheit auch die linke Seite stetig differenzierbar, und durch Differentiation beider Gleichungsseiten resultiert

$$\vec{y}\,'(x) = \vec{f}(x, \vec{y}(x)), \qquad x \in [x_0, x_0 + a].$$

Man bestätigt ferner direkt den Anfangswert $\vec{y}(x_0) = \vec{y}_0$. □

Die Lösung des Integralgleichungsproblems (5.7) gelingt uns nun mit Hilfe des Existenzsatzes 13.4 von Picard. Dazu müssen die Voraussetzungen in geeigneter Weise formuliert werden:

Satz 16.14 (Differentieller Existenzsatz von Picard)
Der Funktionenraum $M := C([x_0, x_0 + a]; \mathbf{R}^n)$ *sei mit der gewichteten Metrik*

$$d_r(\vec{y}, \vec{z}) := \max_{x \in [x_0, x_0 + a]} e^{-r(x - x_0)} \|\vec{y}(x) - \vec{z}(x)\|, \quad \vec{y}, \vec{z} \in M,$$

versehen. Gegeben seien ferner ein $\vec{y}_0 \in \mathbf{R}^n$ *und eine stetige Funktion* $\vec{f} : [x_0, x_0 + a] \times \mathbf{R}^n \to \mathbf{R}^n$*, die der* Lipschitz*–Bedingung*

$$\boxed{\exists\, L \geq 0 \;:\; \|\vec{f}(x, \vec{y}) - \vec{f}(x, \vec{z})\| \leq L\, \|\vec{y} - \vec{z}\| \;\; \forall\, x \in [x_0, x_0 + a] \;\; \forall\, \vec{y},\, \vec{z} \in \mathbf{R}^n} \tag{5.8}$$

genügt. Dann wird durch die Vorschrift

$$\boxed{(T\vec{y})(x) := \vec{y}_0 + \int_{x_0}^{x} \vec{f}(s, \vec{y}(s))\, ds, \qquad \vec{y} \in M, \quad x \in [x_0, x_0 + a],} \tag{5.9}$$

eine Abbildung $T : M \to M$ *erklärt, die genau einen Fixpunkt* $\vec{y} = T\vec{y} \in M$ *besitzt. Das heißt, die Anfangswertaufgabe*

$$\boxed{\vec{y}\,'(x) = \vec{f}(x, \vec{y}(x)), \qquad \vec{y}(x_0) = \vec{y}_0,} \tag{5.10}$$

hat für jede Vorgabe $\vec{y}_0 \in \mathbf{R}^n$ *genau eine Lösung* $\vec{y} \in C^1([x_0, x_0 + a]; \mathbf{R}^n)$.

Bemerkung 16.8 (a) Wie bereits in Bemerkung 13.2 festgestellt wurde, hat der PICARDsche Existenzsatz einen **konstruktiven Aspekt**. Durch **sukzessive Approximation** – in diesem Zusammenhang als Verfahren von PICARD–LINDELÖF in der Literatur geführt – kann der Fixpunkt $\vec{y}$ der Abbildung T durch die Iterationsvorschrift

$$\vec{y}_{k+1}(x) := (T\vec{y}_k)(x) = \vec{y}_0 + \int_{x_0}^{x} \vec{f}(s, \vec{y}_k(s))\, ds, \quad k = 0, 1, 2, \ldots, \quad \vec{y}_0(x) := \vec{y}_0, \tag{5.11}$$

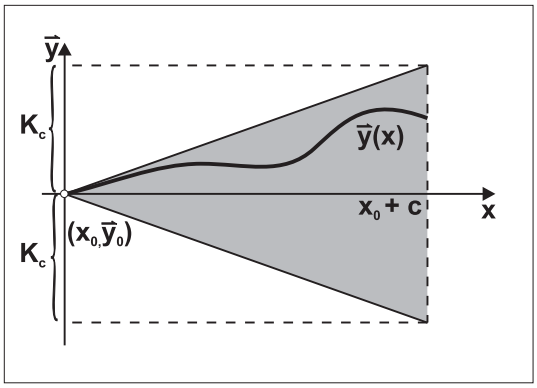
näherungsweise berechnet werden.

(b) Mit einer leicht modifizierten Version des obigen Satzes 16.14 kann man auch eine Existenz– und Eindeutigkeitsaussage für solche Funktionen $\vec{f}(x, \vec{y})$ treffen, die die LIPSCHITZbedingung (5.8) nur auf dem Zylinder G_b, $b < \infty$, erfüllen. Gelte nämlich

$$K := \max_{(x,\vec{y}) \in G_b} \|\vec{f}(x, \vec{y})\|.$$

Damit die PICARD–LINDELÖF–Iteration (5.11) im $(k+1)$–ten Schritt nicht über den Zylinder G_b hinausschießt, muss gelten:

$$\|\vec{y}_{k+1} - \vec{y}_0\| = \Big\| \int_{x_0}^{x} \vec{f}(s, \vec{y}(s))\, ds \Big\| \le K(x - x_0) \le Kc \overset{!}{\le} b, \quad x_0 \le x \le x_0 + c.$$



Das Existenzintervall beim Satz von PICARD

Das heißt, es muss $c \le \frac{b}{K}$ sichergestellt sein. Da x auch nicht über das Stetigkeitsintervall $[x_0, x_0+a]$ der Funktion $\vec{f}$ hinausreichen darf, können Lösungen der AWA (5.10) im allgemeinen nur auf dem Intervall

$$x_0 \le x \le x_0 + c, \qquad c := \min\Big\{a, \frac{b}{K}\Big\}, \tag{5.12}$$

erwartet werden. An diesen Überlegungen ändert sich nichts, wenn anstelle des Intervalls $[x_0, x_0 + c]$ das symmetrische Intervall $[x_0 - c, x_0 + c]$ zugrundegelegt wird. □

BSP. (16.5.3) Wir betrachten die Anfangswertaufgabe

$$y' = \sqrt{1 + y^2}, \qquad y(0) = 0,$$

in der also $f(x, y) := \sqrt{1 + y^2}$, $x_0 := 0$ und $y_0 := 0$ zu setzen sind. Da die Funktion f nicht von x abhängt, dürfen wir $a > 0$ beliebig wählen. Aus dem Mittelwertsatz der Differentialrechnung erhalten wir:

$$|f(x, y) - f(x, z)| \le \sup_{\eta \in \mathbf{R}} \Big(\frac{|\eta|}{\sqrt{1 + \eta^2}} \Big) |y - z| = |y - z|, \qquad y, z \in \mathbf{R}.$$

Hieraus resultiert die LIPSCHITZ–Konstante $L = 1$. Wir dürfen wegen Satz 16.14 für jedes $|x| \le a$ eine eindeutige Lösung $y(x)$ der obigen AWA erwarten. Tatsächlich ist $y(x) = \sinh x$ die gesuchte Lösung, die für alle $x \in \mathbf{R}$ existiert.

16.6 Lineare Systeme von DGln 1. Ordnung und lineare Differentialgleichungen n–ter Ordnung

Der Existenzsatz 16.14 von PICARD garantiert unter geeigneten Stetigkeitsvoraussetzungen die lokale Existenz einer Lösung der Anfangswertaufgabe

$$\boxed{\vec{y}\,'(x) = \vec{f}(x, \vec{y}(x)), \qquad \vec{y}(x_0) = \vec{y}_0.} \tag{6.1}$$

Damit ist nichts über die **Lösungsgesamtheit** des DGl–Systems $\vec{y}\,'(x) = \vec{f}(x, \vec{y}(x))$ gesagt, und zu diesem Punkt gibt es auch keine allgemeinen Aussagen, die über den Existenzsatz 16.11 von PEANO hinausgehen. Die Sachlage erweist sich wesentlich günstiger im Sonderfall der **linearen DGl–Systeme** 1.Ordnung:

$$\boxed{\vec{y}\,'(x) = A(x)\,\vec{y}(x) + \vec{g}(x)} \tag{6.2}$$

mit der gegebenen Matrixfunktion $A \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K}^{(n,n)})$ und der gegebenen Vektorfunktion $\vec{g} \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K}^n)$. Wir beschränken uns hier ausschließlich auf die Körper $\mathbf{K} := \mathbf{R}$ und $\mathbf{K} := \mathbf{C}$ der reellen bzw. der komplexen Zahlen. Bisweilen verwenden wir in (6.2) anstelle der Variablen x die unabhängige Veränderliche t, insbesondere im Zusammenhang mit zeitabhängigen physikalischen Problemen.

Die allgemeine lineare DGl n–ter Ordnung

$$\boxed{L_n y := \sum_{k=0}^{n} a_k(x) y^{(k)} = f(x)} \tag{6.3}$$

kann wie die lineare DGl n–ter Ordnung mit **konstanten Koeffizienten** als Sonderfall des Systems (6.2) aufgefasst werden. Ist nämlich $I \subset \mathbf{R}$ ein Intervall, auf dem $a_n(x) \neq 0 \;\; \forall\, x \in I$ gilt, so können neue Funktionen

$$b_k(x) := -\frac{a_k(x)}{a_n(x)}, \quad k = 0, 1, \ldots, n-1, \qquad g(x) := \frac{f(x)}{a_n(x)}, \quad x \in I,$$

eingeführt werden. Dann ergibt sich anstelle von (6.3) die explizite DGl

$$y^{(n)} = \sum_{k=0}^{n-1} b_k(x) y^{(k)} + g(x), \qquad x \in I. \tag{6.4}$$

Mit den neuen abhängigen Veränderlichen

$$\boxed{y_j(x) := y^{(j-1)}(x), \quad j = 1, 2, \ldots, n,}$$

erhalten wir nun das zu (6.4) äquivalente System

$$\begin{aligned}
y_1'(x) &= y_2(x), \\
y_2'(x) &= y_3(x), \\
&\vdots \\
y_{n-1}'(x) &= y_n(x), \\
y_n'(x) &= b_0(x)y_1(x) + b_1(x)y_2(x) + b_2(x)y_3(x) + \cdots + b_{n-1}(x)y_n(x) + g(x),
\end{aligned}$$

oder in Matrixform

$$\boxed{\vec{y}'(x) = A(x)\vec{y}(x) + \vec{g}(x)} \tag{6.5}$$

worin die folgenden Spezifikationen vorzunehmen sind:

$$\vec{y}(x) := \begin{bmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_n(x) \end{bmatrix}, \quad A(x) := \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & & & 0 \\ 0 & 0 & 0 & \ddots & & 0 \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & & & 1 \\ b_0(x) & b_1(x) & b_2(x) & \cdots & \cdots & b_{n-1}(x) \end{bmatrix}, \quad \vec{g}(x) := \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ g(x) \end{bmatrix}.$$

Die Äquivalenz der Lösungsmenge der Gleichung (6.3) mit der Lösungsmenge des Systems (6.5) wird durch Satz 16.12 sichergestellt. Es genügt also, die nachfolgenden Untersuchungen auf das System (6.2) zu beziehen; darin enthalten sind als Spezialfälle die allgemeinen linearen DGln n–ter Ordnung (6.3).

BSP. (16.6.1) Ein Teilchen mit der Ladung e und der Masse m befinde sich zum Zeitpunkt $t = 0$ im Punkt $\vec{x}_0 := (x_0, y_0, z_0)^T$ eines homogenen Magnetfeldes $\vec{H} := H_0\vec{e}_z$. Es sei $\vec{v}_0 := (\dot{x}_0, \dot{y}_0, \dot{z}_0)^T$ die Anfangsgeschwindigkeit des Teilchens. Welche DGl gilt für die Bahnkurve des Teilchens, wenn die durch das Magnetfeld ausgeübte Kraft gemäß

$$\vec{K} := \frac{e}{c}\,\vec{v} \times \vec{H}, \quad c = const,$$

bestimmt ist?

Lösung: Das NEWTONsche Kraftgesetz erfordert $m\ddot{\vec{x}} = \vec{K} = \frac{e}{c}\,\dot{\vec{x}} \times \vec{H}$, wobei gilt:

$$\dot{\vec{x}} \times \vec{H} = \begin{vmatrix} \vec{e}_x & \dot{x} & 0 \\ \vec{e}_y & \dot{y} & 0 \\ \vec{e}_z & \dot{z} & H_0 \end{vmatrix} = \begin{bmatrix} \dot{y}H_0 \\ -\dot{x}H_0 \\ 0 \end{bmatrix}.$$

Es resultiert die folgende Anfangswertaufgabe

$$\begin{aligned} m\ddot{x} &= \frac{e}{c}\,\dot{y}H_0, \\ m\ddot{y} &= -\frac{e}{c}\,\dot{x}H_0, \qquad t > 0, \quad \vec{x}(0) = (x_0, y_0, z_0)^T, \quad \dot{\vec{x}}(0) = (\dot{x}_0, \dot{y}_0, \dot{z}_0)^T. \\ m\ddot{z} &= 0, \end{aligned} \tag{6.6}$$

Führt man neue Veränderliche

$$y_1(t) := x(t),\ y_2(t) := \dot{x}(t),\ y_3(t) := y(t),\ y_4(t) := \dot{y}(t),\ y_5(t) := z(t),\ y_6(t) := \dot{z}(t)$$

ein, so kann das DGl–System 2.Ordnung (6.6) in der kanonischen Form

$$\dot{\vec{y}}(t) = A\vec{y}(t), \qquad \vec{y}(0) = \vec{y}_0 \tag{6.7}$$

als System 1.Ordnung geschrieben werden, wobei zu setzen sind:

$$A := \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & a & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & -a & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad a := \frac{eH_0}{mc}, \qquad \vec{y}_0 := \begin{bmatrix} x_0 \\ \dot{x}_0 \\ y_0 \\ \dot{y}_0 \\ z_0 \\ \dot{z}_0 \end{bmatrix}.$$

BSP. (16.6.2) Wir betrachten eine ebene elastische Vollkreismembran, deren Rand fest eingespannt sei und die in der Ruhelage einen Vollkreis in der (x,y)–Ebene mit Mittelpunkt im Ursprung einnimmt. Die zeit– und ortsabhängige Auslenkung $u(x,y,t)$ der Membran wird durch die **Wellengleichung**

$$u_{tt} - c^2 \Delta u = 0 \tag{6.8}$$

beschrieben, worin $c := \sqrt{\sigma/\rho}$, σ die mechanische Vorspannung der Membran, ρ ihre Dichte gelten. In Polarkoordinaten $x = r\cos\varphi$, $y = r\sin\varphi$ hat der LAPLACE–Operator Δ die Form

$$\Delta = \frac{1}{r}\,\frac{\partial}{\partial r} + \frac{\partial^2}{\partial r^2} + \frac{1}{r^2}\,\frac{\partial^2}{\partial \varphi^2}, \qquad r > 0.$$

Somit kann (6.8) mit $u = u(r,\varphi,t)$ wie folgt geschrieben werden:

$$u_{tt} = c^2\Big(\frac{1}{r}\,u_r + u_{rr} + \frac{1}{r^2}\,u_{\varphi\varphi}\Big). \tag{6.9}$$

Lösungen können häufig in **Produktform** $u(r,\varphi,t) = W(r,\varphi)\cdot T(t)$ gefunden werden. Mit einem solchen **Separationsansatz** resultiert aus (6.9):

$$\frac{\ddot{T}}{c^2 T} = \frac{1}{W}\left(\frac{1}{r}\,W_r + W_{rr} + \frac{1}{r^2}\,W_{\varphi\varphi}\right) \stackrel{!}{=} -\lambda^2 = const. \tag{6.10}$$

An dieser Stelle wird eine **typische Schlußfolgerung** gezogen: Die linke Gleichungsseite ist eine Funktion der Variablen t allein, die rechte hingegen eine Funktion der Variablen r und φ allein. Beide Gleichungsseiten können sich nur in einer **Konstanten** treffen, die hier aus physikalischen Gründen mit $-\lambda^2$ angesetzt wurde. (Man erhält nur mit diesem Wert zeitperiodische Lösungen, die den erwarteten Schwingungen der Membran entsprechen.) Die **partielle DGl** (6.9) zerfällt jetzt in die zwei Teil–DGln

$$\ddot{T} + \lambda^2 c^2 T = 0, \qquad \frac{1}{r}\,W_r + W_{rr} + \frac{1}{r^2}\,W_{\varphi\varphi} + \lambda^2 W = 0.$$

Die erste Gleichung ist die bekannte Schwingungs–DGl. Die zweite Gleichung zerlegen wir nochmals mit einem Separationsansatz der Form $W(r,\varphi) = R(r)\cdot\Phi(\varphi)$ und erhalten:

$$\frac{\Phi_{\varphi\varphi}}{\Phi} = -\frac{r^2}{R}\Big(R_{rr} + \frac{1}{r}\,R_r + \lambda^2 R\Big) \stackrel{!}{=} -p^2 = const. \tag{6.11}$$

Wir argumentieren wie vorher: Die linke Gleichungsseite ist eine Funktion von φ allein, die rechte Seite eine Funktion von r allein. Beide Seiten können sich wieder nur in einer Konstanten treffen, die hier mit $-p^2$ angesetzt wurde. Somit erhält man die zwei gewöhnlichen DGln

$$\frac{d^2\Phi}{d\varphi^2} + p^2\Phi = 0,$$

und

$$r^2\frac{d^2R}{dr^2} + r\frac{dR}{dr} + (\lambda^2 r^2 - p^2)\,R = 0. \tag{6.12}$$

Die DGl (6.12) ist eine lineare DGl 2.Ordnung mit nichtkonstanten Koeffizienten; sie heißt BESSELsche Differentialgleichung. Wir werden in Abschnitt 16.9 näher auf die für Physik und Technik sehr wichtige DGl eingehen. Führt man neue abhängige Veränderliche gemäß $y_1(r) := R(r)$, $y_2(r) := R'(r)$ ein, so lässt sich die DGl (6.12) in der folgenden Form als homogenes System 1.Ordnung schreiben:

$$\begin{bmatrix} y_1' \\ y_2' \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \frac{p^2}{r^2} - \lambda^2 & -\frac{1}{r} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \qquad r > 0. \tag{6.13}$$

Wir befassen uns jetzt mit der Lösungsmenge des inhomogenen linearen Systems 1.Ordnung

$$\boxed{\vec{y}\,'(x) = A(x)\vec{y}(x) + \vec{g}(x), \qquad A(\cdot) := (a_{jk}(\cdot)) \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K}^{(n,n)}).} \tag{6.14}$$

Für solche Systeme gilt wie im Fall der linearen DGl 1.Ordnung das **Superpositionsprinzip**, man vergleiche Satz 16.3:

Satz 16.15 *Gegeben seien ein Intervall $I \subset \mathbf{R}$ und Funktionen $A \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K}^{(n,n)})$, $\vec{g} \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K}^n)$ mit $A \in C(I; \mathbf{K}^{(n,n)})$ und $\vec{g} \in C(I; \mathbf{K}^n)$.*

(a) *Die Lösungen $\vec{y}_h \in C^1(I; \mathbf{K}^n)$ des homogenen Systems $\vec{y}\,'(x) = A(x)\vec{y}(x)$ bilden einen Unterraum $U_h \subset C^1(I; \mathbf{K}^n)$; das heißt*

$$\boxed{\vec{y}\,'_1 - A(x)\vec{y}_1 = \vec{0} = \vec{y}\,'_2 - A(x)\vec{y}_2 \quad \Rightarrow \quad (\lambda\vec{y}_1 + \mu\vec{y}_2)' - A(x)(\lambda\vec{y}_1 + \mu\vec{y}_2) = \vec{0} \ \ \forall\, \lambda,\, \mu \in \mathbf{K}.}$$

(b) *Ist $\vec{y}_p \in C^1(I; \mathbf{K}^n)$ eine partikuläre Lösung des inhomogenen Systems $\vec{y}\,'(x) = A(x)\vec{y}(x) + \vec{g}(x)$, so ist die allgemeine Lösung des Systems* (6.14) *der affine Unterraum*

$$\boxed{\mathcal{L}(DGl) = \vec{y}_p + U_h = \{\vec{y} \in C^1(I; \mathbf{K}^n) \,:\, \vec{y}(x) = \vec{y}_p(x) + \vec{y}_h(x), \quad \vec{y}_h \in U_h\}.}$$

Auch hier zerfällt also die Lösungskonstruktion für das System (6.14) in die zwei folgenden Teilaufgaben:

(H) Bestimme den Unterraum $U_h \subset C^1(I; \mathbf{K}^n)$, das heißt, die Lösungsgesamtheit des homogenen Systems

$$\boxed{\vec{y}\,'(x) = A(x)\vec{y}(x), \qquad x \in I.} \tag{6.15}$$

(P) Bestimme eine partikuläre Lösung $\vec{y}_p \in C^1(I; \mathbf{K}^n)$ des inhomogenen Systems (6.14).

Beide Aufgaben (H) und (P) stellen für sich i.a. unlösbare Probleme im konstruktiven Sinne dar: Es sind keine allgemeinen **analytischen** Lösungsverfahren bekannt, mit deren Hilfe (H) und (P) für sich gelöst werden können. Immerhin bietet das Picard–Lindelöf–Verfahren einen konstruktiven Aspekt zur näherungsweisen Bestimmung einer Lösung. Aus dem Existenzsatz 16.14 von Picard folgern wir sogar:

Satz 16.16 *Auf dem Intervall $I := [x_0, x_0 + a]$ seien* **stetige** *Koeffizientenfunktionen $a_{jk} \in C(I)$ der Matrixfunktion $A(x) = (a_{jk}(x))$ gegeben. Dann existiert zu jeder Anfangsvorgabe*

$$\vec{y}(x_0) = \vec{y}_0 \in \mathbf{K}^n$$

genau eine Lösung $\vec{y} \in C^1(I; \mathbf{K}^n)$ des homogenen Systems (6.15).

Begründung: Wir setzen

$$L := \max_{x \in I} \Big(\sum_{j,k=1}^{n} |a_{jk}(x)|^2 \Big)^{1/2} = \max_{x \in I} \|A(x)\|_F$$

sowie $\vec{f}(x, \vec{y}) := A(x)\vec{y}$, $\vec{y} \in \mathbf{K}^n$. Dann erfüllt die Funktion $\vec{f}$ die Voraussetzungen des Picardschen Satzes 16.14, insbesondere auch die Lipschitz–Bedingung (5.8) mit der oben definierten Konstanten L. Die obige Existenzaussage folgt jetzt aus diesem Satz. □

Werden nun die n linear unabhängigen Vektoren $\vec{e}_j$ der Standardbasis als Anfangsvektoren $\vec{y}_0 = \vec{e}_j$ eingesetzt, so resultieren gemäß Satz 16.16 n Lösungen $\vec{y}_j \in C^1(I;\mathbf{K}^n)$, $1 \le j \le n$, des homogenen Systems (6.15) mit den Anfangswerten $\vec{y}_j(x_0) = \vec{e}_j$. Diese Lösungen sind **linear unabhängig**, da sie im Punkt $x = x_0$ linear unabhängig sind. Eine beliebige Lösung $\vec{y}_h \in C^1(I;\mathbf{K}^n)$ nimmt sicher einen Anfangswert $\vec{y}_h(x_0) = (C_1, C_2, \ldots, C_n)^T \in \mathbf{K}^n$ stetig an. Wegen der Eindeutigkeitsaussage in Satz 16.16 muss dann $\vec{y}_h(x) = \sum_{j=1}^{n} C_j \vec{y}_j(x)$ gelten. Also kann das homogene System (6.15) keine weiteren linear unabhängigen Lösungen haben:

Satz 16.17 *Es seien die Voraussetzungen des Satzes* 16.16 *erfüllt. Dann wird der Unterraum*

$$\boxed{U_h := \{\vec{y}_h \in C^1(I;\mathbf{K}^n) \,:\, \vec{y}_h'(x) = A(x)\vec{y}_h(x), \quad x \in I\}}$$

von genau n linear unabhängigen Vektorfunktionen $\vec{y}_1(x), \vec{y}_2(x), \ldots, \vec{y}_n(x)$ aufgespannt.

Definition 16.13 (a) *Ein System von n linear unabhängigen Lösungen $\vec{y}_1(x), \vec{y}_2(x), \ldots, \vec{y}_n(x)$ des homogenen DGl–Systems* (6.15) *heiße ein* **Fundamentalsystem**.

(b) *Bilden die Lösungen $\vec{y}_1(x), \vec{y}_2(x), \ldots, \vec{y}_n(x)$ des homogenen DGl–Systems* (6.15) *ein Fundamentalsystem, so heiße die Matrixfunktion*

$$\boxed{Y(x) := (\vec{y}_1(x), \vec{y}_2(x), \ldots, \vec{y}_n(x))} \tag{6.16}$$

eine **Fundamentalmatrix** *oder* WRONSKI**–Matrix** *des DGl–Systems* (6.15).

BSP. (16.6.3) Zu bestimmen sind ein Fundamentalsystem und eine Fundamentalmatrix des DGl–Systems

$$\vec{y}\,'(x) = \begin{bmatrix} 2x & 0 \\ 2x & 0 \end{bmatrix} \vec{y}(x).$$

Lösung: In den Komponentenfunktionen $\vec{y}(x) = (y_1(x), y_2(x))^T$ liegen die beiden skalaren DGln (i) $y_1' = 2xy_1$ und (ii) $y_2' = 2xy_1$ vor. Die DGl (i) integrieren wir mit dem Verfahren der TdV und erhalten $y_1(x) = C_1 e^{x^2}$. Dies führt in (ii) auf die DGl $y_2'(x) = C_1 2xe^{x^2} = C_1(e^{x^2})'$, die somit direkt integrierbar ist und die Lösung $y_2(x) = C_1 e^{x^2} + C_2$ liefert. Die allgemeine Lösung hat nun die Form

$$\vec{y}(x) = C_1 \underbrace{e^{x^2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{=:\vec{y}_1(x)} + C_2 \underbrace{\begin{bmatrix} 0 \\ 1 \end{bmatrix}}_{=:\vec{y}_2(x)} =: C_1\vec{y}_1(x) + C_2\vec{y}_2(x),$$

worin die beiden Vektorfunktionen $\vec{y}_1(x), \vec{y}_2(x)$ ein Fundamentalsystem bilden, aus welchem die folgende Fundamentalmatrix resultiert:

$$Y(x) = \begin{bmatrix} e^{x^2} & 0 \\ e^{x^2} & 1 \end{bmatrix}.$$

Man kann aus der allgemeinen Lösung ein Fundamentalsystem $\vec{y}_1(x)$, $\vec{y}_2(x)$ aber auch so bestimmen, dass die Anfangswerte $\vec{y}_1(0) = (1,0)^T$, $\vec{y}_2(0) = (0,1)^T$ angenommen werden. Dazu wähle man $(C_1, C_2) = (1, -1)$ bzw. $(C_1, C_2) = (0, 1)$. Es resultieren

$$\vec{y}_1(x) = \begin{bmatrix} e^{x^2} \\ e^{x^2} - 1 \end{bmatrix}, \quad \vec{y}_2(x) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad Y(x) = \begin{bmatrix} e^{x^2} & 0 \\ e^{x^2} - 1 & 1 \end{bmatrix}, \quad Y(0) = Id.$$

Aus der Definition einer Fundamentalmatrix resultieren einige einfache Folgerungen, die in dem folgenden Satz zusammengefasst sind:

Satz 16.18 (a) *Es sei $Y(x)$ eine Fundamentalmatrix des homogenen DGl–Systems* (6.15). *Dann ist auch $\tilde{Y}(x) := Y(x)B$ für jede Matrix $B \in \mathrm{Inv}\,(\mathbf{K}^n)$ eine Fundamentalmatrix desselben Systems.*

(b) *Für die* Wronski*–Determinante $W(x) := \det Y(x)$ gilt $W(x) \neq 0 \ \forall\ x \in I$, und folglich $Y(x) \in \mathrm{Inv}\,(\mathbf{K}^n)\ \forall\ x \in I$.*

(c) *Die Fundamentalmatrizen $Y(x)$ sind genau die invertierbaren Lösungen der Matrix–Differentialgleichung*

$$\boxed{Y'(x) = A(x)Y(x), \qquad x \in I.}$$

(d) *Die allgemeine Lösung des homogenen DGl–Systems* (6.15) *hat die Darstellung*

$$\boxed{\vec{y}(x) = Y(x)\vec{c}, \qquad \vec{c} \in \mathbf{K}^n.}$$

Insbesondere ist $\vec{y}(x) := Y(x)Y^{-1}(x_0)\vec{y}_0$ die eindeutig bestimmte Lösung der Anfangswertaufgabe

$$\boxed{\vec{y}\,'(x) = A(x)\vec{y}(x), \qquad \vec{y}(x_0) = \vec{y}_0 \in \mathbf{K}^n.} \tag{6.17}$$

Begründung zu (b): Wäre $W(x_1) = 0$ für ein $x_1 \in I$, so hätte das lineare Gleichungssystem $Y(x_1)\vec{c} = \vec{0}$ eine nichttriviale Lösung $\vec{c}_0 \neq \vec{0}$. Es wäre $\vec{y}(x) := Y(x)\vec{c}_0$ eine Lösung der AWA $\vec{y}\,'(x) = A(x)\vec{y}(x),\ \vec{y}(x_1) = \vec{0}$. Satz 16.16 lässt aber nur eine Lösung zu, nämlich $\vec{y}(x) = \vec{0}$. Wir haben einen Widerspruch konstruiert. Also muss $W(x) \neq 0$ für alle $x \in I$ gelten. □

Bemerkung 16.9 (a) Ist das DGl–System $\vec{y}\,'(x) = A(x)\vec{y}(x)$ aus der homogenen linearen DGl (6.3) entstanden, und ist $Y(x) := (\vec{y}_1(x), \vec{y}_2(x), \ldots, \vec{y}_n(x))$ eine Fundamentalmatrix dieses Systems, so bildet die **erste Zeile** $y_{11}(x), y_{12}(x), \ldots, y_{1n}(x)$ eine **Basis** für den Lösungsraum $\mathrm{Kern}\,L_n$ der homogenen DGl

$$L_n y := a_n(x)y^{(n)} + a_{n-1}(x)y^{(n-1)} + \cdots + a_1(x)y' + a_0(x)y = 0.$$

(b) Die bisher aufgelisteten Existenz– und Eindeutigkeitssätze enthalten keine brauchbaren Verfahren zur Auffindung der Lösungen. Wir hatten bereits erwähnt, dass solche Verfahren im allgemeinen auch nicht existieren. Ausgenommen sind die homogenen DGl–Systeme 1.Ordnung mit **konstanter Koeffizientenmatrix** $A(x) \equiv A \in \mathbf{K}^{(n,n)}$, die wir bereits in Abschnitt 11.7 vollständig behandelt haben.

(c) Die Wronski–Determinante $W(x) := \det Y(x)$ kann bis auf einen konstanten Faktor auch ohne Kenntnis eines Fundamentalsystems bestimmt werden. Dies zeigen wir im folgenden Satz. □

Satz 16.19 *Gegeben seien ein Intervall $I \subset \mathbf{R}$ und eine Matrixfunktion $A \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K}^{(n,n)})$, $A(x) = (a_{jk}(x))$, mit $a_{jk} \in C(I)$. Es bezeichne*

$$\boxed{\mathrm{Sp}\,(A(x)) := \sum_{j=1}^{n} a_{jj}(x)}$$

die **Spur** *der Matrixfunktion, und es sei $Y(x)$ eine Fundamentalmatrix des homogenen DGl–Systems* (6.15). *Dann gilt für die* Wronski*–Determinante $W(x) := \det Y(x)$:*

$$\boxed{\frac{d}{dx}\,W(x) = W(x) \cdot \mathrm{Sp}\,(A(x)) \ \ \forall\ x \in I,} \tag{6.18}$$

also nach Integration

$$\boxed{W(x) = W(x_0) \cdot \exp\Big(\int_{x_0}^{x} \mathrm{Sp}\,(A(t))\,dt \Big) \;\; \forall\; x,\, x_0 \in I.} \tag{6.19}$$

Bemerkung 16.10 Liegt die DGl (6.3) vor, so gilt

$$\mathrm{Sp}\,(A(x)) = b_{n-1}(x) = -\frac{a_{n-1}(x)}{a_n(x)},$$

und somit steht die DGl (6.18) im Einklang mit dem Resultat aus Satz 10.3. □

BSP. (16.6.4) Es sei $A(x)$ die Matrix aus BSP. (16.6.3). Wir haben $\mathrm{Sp}\,(A(x)) = 2x$, und somit gemäß (6.19) $W(x) = W(x_0)\,e^{x^2-x_0^2}$. Andererseits gilt für die dort berechnete Fundamentalmatrix $Y(x)$ die Beziehung $W(x) = \det Y(x) = e^{x^2}$ und folglich $W(x_0) = e^{x_0^2}$. Also haben wir in Übereinstimmung mit dem Resultat aus (6.19)

$$\frac{W(x)}{W(x_0)} = e^{x^2-x_0^2}.$$

BSP. (16.6.5) Das homogene DGl–System

$$\vec{y}\,'(x) = \begin{bmatrix} 0 & 1 \\ -\frac{6}{x^2} & \frac{4}{x} \end{bmatrix} \vec{y}(x), \qquad x > 0,$$

hat die beiden Lösungsvektoren

$$\vec{y}_1(x) := (x^2, 2x)^T, \quad \vec{y}_2(x) := (x^3, 3x^2)^T,$$

wie durch Einsetzen leicht verifiziert werden kann. Setzen wir

$$Y(x) := (\vec{y}_1(x), \vec{y}_2(x)) = \begin{bmatrix} x^2 & x^3 \\ 2x & 3x^2 \end{bmatrix}, \qquad x > 0,$$

so folgt $\det Y(x) = x^4 \neq 0 \;\forall\; x > 0$. Also ist $Y(x)$ eine Fundamentalmatrix, im Einklang mit der aus $\mathrm{Sp}\,(A(x)) = \frac{4}{x}$ gewonnenen WRONSKI–Determinante

$$W(x) = W(x_0)\,\exp\Big(\int_{x_0}^{x} \frac{4\,dt}{t} \Big) = W(x_0) \cdot \frac{x^4}{x_0^4}, \quad x,\, x_0 > 0.$$

Das obige DGl–System kann offenbar auf die EULERsche DGl

$$y'' - \frac{4}{x}\,y' + \frac{6}{x^2}\,y = 0, \quad x > 0,$$

zurückgeführt werden. Die erste Zeile der Fundamentalmatrix $Y(x)$ liefert nun für diese DGl die zwei linear unabhängigen Lösungen

$$y_1(x) := x^2, \quad y_2(x) := x^3,$$

die somit die allgemeine Lösung der EULERschen DGl aufspannen.

Nachfolgend geben wir zwei Methoden an, mit denen es in speziellen Fällen gelingt, ein Fundamentalsystem für das homogene DGl–System $\vec{y}\,'(x) = A(x)\vec{y}(x)$ zu bestimmen.

Satz 16.20 *Gegeben seien ein Intervall $I \subset \mathbf{R}$ und eine Matrixfunktion $A \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K}^{(n,n)})$, $A(x) = (a_{jk}(x))$, mit $a_{jk} \in C(I)$. Es sei ferner $\vec{0} \neq \vec{u} \in C^1(I; \mathbf{K}^n)$ eine* **partikuläre Lösung** *des homogenen DGl–System $\vec{y}\,'(x) = A(x)\vec{y}(x)$, $x \in I$. Für die j–te Komponente der Vektorfunktion $\vec{u}(x)$ gelte $u_j(x) \neq 0 \ \forall\, x \in I$. Dann erhält man mit dem Ansatz*

$$\boxed{\vec{y}(x) = v(x)\,\vec{u}(x) + \vec{z}(x), \qquad z_j(x) := 0,} \tag{6.20}$$

das **reduzierte System**

$$\boxed{z_i'(x) = \sum_{\substack{k=1\\k\neq j}}^{n} \Big(a_{ik}(x) - \frac{u_i(x)}{u_j(x)}\,a_{jk}(x)\Big)\, z_k(x), \quad i \neq j,} \tag{6.21}$$

von $n-1$ Gleichungen für die $n-1$ Unbekannten $z_1(x), \ldots, z_{j-1}(x), z_{j+1}(x), \ldots, z_n(x)$. Die Funktion $v(x)$ bestimmt man schließlich durch Integration der DGl

$$\boxed{v'(x) = \frac{1}{u_j(x)} \sum_{\substack{k=1\\k\neq j}}^{n} a_{jk}(x) z_k(x).} \tag{6.22}$$

Begründung: Aus dem Ansatz (6.20) erhält man

$$\vec{y}\,'(x) = v'(x)\cdot\vec{u}(x) + v(x)\cdot\vec{u}\,'(x) + \vec{z}\,'(x) \stackrel{!}{=} v(x)A(x)\vec{u}(x) + A(x)\vec{z}(x),$$

und somit wegen $\vec{u}\,'(x) = A(x)\vec{u}(x)$ das DGl–System

$$\vec{z}\,'(x) = A(x)\vec{z}(x) - v'(x)\vec{u}(x). \tag{6.23}$$

Ansatzgemäß gilt nun $z_j(x) := 0$, so dass durch Differentiation die Gleichung

$$0 = z_j'(x) = \sum_{\substack{k=1\\k\neq j}}^{n} a_{jk}(x) z_k(x) - v'(x)u_j(x)$$

folgt, aus der bereits (6.22) resultiert. Wird schließlich (6.22) in die Gleichung (6.23) eingesetzt, so ergibt sich (6.21). □

BSP. (16.6.6) Hier betrachten wir nochmals das DGl–System

$$\vec{y}\,'(x) = \begin{bmatrix} 2x & 0 \\ 2x & 0 \end{bmatrix} \vec{y}(x)$$

aus BSP. (16.6.3). Wir "erraten" eine Lösung $\vec{u}(x) = (0,1)^T \neq \vec{0}$, die die nichverschwindende Komponente $u_2(x) = 1$ besitzt. Mit dem Ansatz (6.20)

$$\vec{y}(x) = v(x) \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} z_1(x) \\ 0 \end{bmatrix}$$

erhält man gemäß (6.21):

$$z_1'(x) = \Big(a_{11}(x) - \frac{u_1(x)}{u_2(x)}\,a_{21}(x)\Big)\, z_1(x) = 2xz_1(x) \quad \Rightarrow \quad z_1(x) = e^{x^2}.$$

Ferner folgt aus (6.22)

$$v'(x) = \frac{a_{21}(x)}{u_2(x)}\, z_1(x) = 2xe^{x^2} \quad \Rightarrow \quad v(x) = e^{x^2}.$$

Wir haben wieder das bekannte Fundamentalsystem vorliegen:

$$\vec{y}(x) = e^{x^2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \vec{u}(x) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Das in Satz 16.20 vorgestellte Verfahren heißt D'ALEMBERTsches Verfahren der **Reduktion der Ordnung**. Im **Sonderfall** der homogenen linearen DGl (6.3) ergibt sich die folgende Variante:

Satz 16.21 *Gegeben seien ein Intervall $I \subset \mathbf{R}$ und Koeffizientenfunktionen $a_k \in C(I)$, $0 \le k \le n$. Es sei ferner $0 \neq u \in C^n(I)$ eine partikuläre Lösung der homogenen DGl*

$$\boxed{L_n y := \sum_{k=0}^{n} a_k(x) y^{(k)} = 0.}$$

Aus dem Ansatz

$$\boxed{y(x) = v(x) \cdot u(x)} \tag{6.24}$$

zur Bestimmung einer weiteren Lösung $y \in C^n(I)$ erhält man die DGl

$$b_n(x)v^{(n)} + b_{n-1}(x)v^{(n-1)} + \cdots + b_1(x)v' = 0,$$

die vermöge der Substitution $z(x) := v'(x)$ in eine DGl der Ordnung $n-1$ für $z(x)$ übergeht. Darin sind die Koeffizientenfunktionen $b_k(x)$ in der folgenden Weise definiert:

$$\boxed{b_k(x) := \sum_{j=k}^{n} \binom{j}{k} a_j(x) u^{(j-k)}(x), \quad k = 1, 2, \ldots, n.}$$

Den *Beweis* führt man durch Einsetzen von (6.24) in die homogene DGl $L_n y = 0$. □

BSP. (16.6.7) Die folgende DGl 2.Ordnung hat eine partikuläre Lösung $u(x) := e^{-x}$:

$$(x^2 + x)y'' + (x^2 - x - 1)y' - (1 + 2x)y = 0.$$

Dies bestätigt man durch Einsetzen. Wir reduzieren die Ordnung der DGl mit dem Produktansatz

$$\left.\begin{array}{lcl|l} y(x) & = & v(x)e^{-x} & \cdot(-1)(1+2x) \\ y'(x) & = & v'e^{-x} - ve^{-x} & \cdot(x^2 - x - 1) \\ y''(x) & = & v''e^{-x} - 2v'e^{-x} + ve^{-x} & \cdot(x^2 + x) \end{array}\right\} (+)$$

Einsetzen in die DGl liefert

$$e^{-x}((x^2 + x)v'' - (x^2 + 3x + 1)v' + \underbrace{(x^2 + x - x^2 + x + 1 - 1 - 2x)}_{=0} v) \stackrel{!}{=} 0.$$

Hieraus folgt wegen $e^{-x} \neq 0$ die folgende lineare DGl 1.Ordnung für die Unbekannte $z(x) := v'(x)$:

$$(x^2 + x)z' - (x^2 + 3x + 1)z = 0.$$

Man integriert wieder mit dem Verfahren der TdV unter Verwendung einer Partialbruchzerlegung und erhält die allgemeine Lösung $z(x) = C_1 x(x+1)e^x = v'(x)$. Hier führt direkte Integration auf die Ansatzfunktion $v(x) = C_1(x^2 - x + 1)e^x + C_2$, so dass die gegebene DGl die folgende allgemeine Lösung besitzt:

$$y(x) = C_1(x^2 - x + 1) + C_2 e^{-x}, \quad x \in \mathbf{R}.$$

Ein weiteres Verfahren zur Berechnung einer Fundamentalmatrix lehnt sich an die Verwendung der Matrix–Exponentialfunktion e^{xA} bei konstanten Matrizen $A \in \mathbf{K}^{(n,n)}$ an. Es gilt die folgende Aussage:

Satz 16.22 *Gegeben seien ein Intervall $I \subset \mathbf{R}$ und eine Matrixfunktion $A \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K}^{(n,n)})$, $A(x) = (a_{jk}(x))$, mit $a_{jk} \in C(I)$. Es gelte*

$$\boxed{A(t)A(s) = A(s)A(t) \ \ \forall\ s,t \in I.} \tag{6.25}$$

Dann ist

$$\boxed{Y(x) := e^{B(x)}, \qquad B(x) := \int_{x_0}^{x} A(t)\,dt,} \tag{6.26}$$

eine Fundamentalmatrix für das homogene DGl–System $\vec{y}\,'(x) = A(x)\vec{y}(x)$.

Begründung: Wegen (6.25) haben wir die Kommutator–Eigenschaft

$$B(x)B'(x) = B(x)A(x) = \int_{x_0}^{x} A(t)\,dt\,A(x) = \int_{x_0}^{x} A(x)A(t)\,dt = A(x)B(x) = B'(x)B(x).$$

Es resultiert $(B^2(x))' = B'(x)B(x) + B(x)B'(x) = 2B'(x)B(x)$, und mit vollständiger Induktion $(B^j(x))' = jB'(x)B^{j-1}(x)$ für alle $j \in \mathbf{N}$. Somit gilt

$$(e^{B(x)})' = \frac{d}{dx}\Big(\sum_{j=0}^{\infty} \frac{1}{j!}\,B^j(x)\Big) = B'(x)\sum_{k=0}^{\infty}\frac{1}{k!}B^k(x) = A(x)e^{B(x)}.$$

BSP. (16.6.8) Es sei $A(x) := \begin{bmatrix} 2x & 0 \\ 2x & 0 \end{bmatrix}$ wieder die Matrix aus BSP. (16.6.3). Nun gilt wegen $A(t) = 2t\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$ ganz offensichtlich die geforderte Kommutator–Eigenschaft (6.25). Weiterhin folgt mit der konstanten Matrix $C := \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$:

$$B(x) = \int_{x_0}^{x} 2tC\,dt = (x^2 - x_0^2)C, \quad B^2(x) = (x^2 - x_0^2)^2 C, \quad \cdots \quad , B^j(x) = (x^2 - x_0^2)^j C.$$

Somit erhalten wir die Fundamentalmatrix

$$Y(x) = e^{B(x)} = Id + \sum_{j=1}^{\infty}\frac{1}{j!}(x^2 - x_0^2)^j C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + (e^{x^2 - x_0^2} - 1)\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} e^{x^2-x_0^2} & 0 \\ e^{x^2-x_0^2} - 1 & 1 \end{bmatrix},$$

in Übereinstimmung mit unserem früheren Resultat in BSP. (16.6.3).

Bemerkung 16.11 Die Anwendungsmöglichkeiten des Satzes 16.22 sind eher begrenzt, da die Kommutator–Eigenschaft (6.25) nur in wenigen Ausnahmefällen – zum Beispiel für Matrizen $A(t) = h(t) \cdot \tilde{A}$ mit skalarer Funktion $h \in C(I)$ und konstanter Matrix $\tilde{A} \in \mathbf{K}^{(n,n)}$ – erfüllt ist. Insbesondere führt der Fall $h := 1$ wieder auf die DGl–Systeme 1.Ordnung mit konstanten Koeffizienten, mit denen wir uns in Abschnitt 11.7 erschöpfend befasst haben. □

Das Wesentliche über die Lösungsstruktur und über Lösungsverfahren des homogenen DGl–Systems (6.15) ist nun gesagt worden, und wir können die Teilaufgabe (H) als befriedigend gelöst ansehen. Zur Lösung der Teilaufgabe (P), nämlich der Bestimmung einer partikulären Lösung $\vec{y}_p(x)$ des inhomogenen DGl–Systems (6.14)

$$\vec{y}\,'(x) = A(x)\vec{y}(x) + \vec{g}(x),$$

hat man bis auf einige Glücksfälle geschickten Erratens oder intuitiver Direktansätze bisher nur das PICARD–LINDELÖFsche Iterationsverfahren zur näherungsweisen Berechnung von $\vec{y}_p(x)$ in der Hand. Die Sachlage vereinfacht sich ganz erheblich, wenn eine Fundamentalmatrix $Y(x)$ des homogenen DGl–Systems bekannt ist. In diesem Fall führt das Verfahren der **Variation der Konstanten von** LAGRANGE (VdK) erfolgreich zur Berechnung einer partikulären Lösung $\vec{y}_p$.

Ist $Y(x)$ eine Fundamentalmatrix des homogenen DGl–Systems $\vec{y}\,'(x) = A(x)\vec{y}(x)$, so setzt man den Konstantenvektor $\vec{c}$ in der allgemeinen Lösung $\vec{y}(x) = Y(x)\vec{c}$ als Vektorfunktion an: $\vec{c} = \vec{v}(x)$. Nun gilt für $\vec{y}_p(x) := Y(x)\vec{v}(x)$:

$$\vec{y}\,'_p(x) = Y'(x)\vec{v}(x) + Y(x)\vec{v}\,'(x) = A(x)\underbrace{Y(x)\vec{v}(x)}_{=\vec{y}_p(x)} + Y(x)\vec{v}\,'(x) \stackrel{!}{=} A(x)\vec{y}_p(x) + \vec{g}(x).$$

Somit folgt offenbar

$$Y(x)\vec{v}\,'(x) = \vec{g}(x) \qquad \Rightarrow \qquad \vec{v}(x) = \int_{x_0}^{x} Y^{-1}(t)\vec{g}(t)\,dt.$$

Wir haben zusammenfassend:

Satz 16.23 *Gegeben seien ein Intervall $I \subset \mathbf{R}$, eine Matrixfunktion $A \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K}^{(n,n)})$, $A(x) = (a_{jk}(x))$, mit $a_{jk} \in C(I)$ und eine Vektorfunktion $\vec{g} \in C(I;\mathbf{K}^n)$. Ist $Y(x)$ eine Fundamentalmatrix des homogenen DGl–Systems $\vec{y}\,'(x) = A(x)\vec{y}(x)$, so ist*

$$\boxed{\vec{y}(x) := Y(x)\Big(\vec{c} + \int_{x_0}^{x} Y^{-1}(t)\vec{g}(t)\,dt\Big), \quad \vec{c} = const,} \tag{6.27}$$

die allgemeine Lösung des inhomogenen DGl–Systems (6.14)*. Die Anfangswertaufgabe*

$$\vec{y}\,'(x) = A(x)\vec{y}(x) + \vec{g}(x), \quad x \in I, \qquad \vec{y}(x_0) = \vec{y}_0 \in \mathbf{K}^n,$$

hat für jedes $x_0 \in I$ die eindeutig bestimmte Lösung

$$\boxed{\vec{y}(x) = Y(x)\Big(Y^{-1}(x_0)\vec{y}_0 + \int_{x_0}^{x} Y^{-1}(t)\vec{g}(t)\,dt\Big), \qquad x \in I.} \tag{6.28}$$

BSP. (16.6.9) Zu bestimmen ist die allgemeine Lösung des inhomogenen DGl–Systems

$$\begin{aligned} \dot{x} &= x + 2y + 3z - 2t, \\ \dot{y} &= -3x - 2y - z + 4t, \\ \dot{z} &= x + y + z + 6, \end{aligned}$$

sowie die Lösung der Anfangswertaufgabe zum Anfangswert $(x(0), y(0), z(0)) = (5, 5, 5)$.

Lösung: Mit den folgenden Vektorfunktionen $\vec{x}(t), \vec{g}(t)$ und der konstanten Matrix $A \in \mathbf{R}^{(3,3)}$ haben wir das inhomogene DGl–System $\dot{\vec{x}}(t) = A\vec{x}(t) + \vec{g}(t)$ vorliegen:

$$\vec{x}(t) := \begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix}, \quad \vec{g}(t) := \begin{bmatrix} -2t \\ 4t \\ 6 \end{bmatrix}, \quad A := \begin{bmatrix} 1 & 2 & 3 \\ -3 & -2 & -1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Eine Fundamentalmatrix $Y(t)$ berechnen wir wie in Abschnitt 11.7 durch Ermittlung der Eigenwerte und –vektoren der Matrix A. Das charakteristische Polynom

$$P_3(\lambda) = \det(A - \lambda\, Id) = \begin{vmatrix} 1-\lambda & 2 & 3 \\ -3 & -2-\lambda & -1 \\ 1 & 1 & 1-\lambda \end{vmatrix} = -\lambda(\lambda^2 + 1)$$

hat offensichtlich die drei Nullstellen $\lambda_1 = 0$, $\lambda_2 = i$, $\lambda_3 = -i$. Wir berechnen die zugeordneten Eigenvektoren aus den Lösungen der homogenen linearen Gleichungssysteme

$$\vec{0} \;=\; (A - 0\; Id)\vec{v}_1 = \begin{bmatrix} 1 & 2 & 3 \\ -3 & -2 & -1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}, \quad \text{also} \quad \vec{v}_1 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix},$$

$$\vec{0} \;=\; (A - i\; Id)\vec{v}_2 = \begin{bmatrix} 1-i & 2 & 3 \\ -3 & -2-i & -1 \\ 1 & 1 & 1-i \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}, \quad \text{also} \quad \vec{v}_2 = \begin{bmatrix} 5 \\ 4i-7 \\ 3-i \end{bmatrix} = \overline{\vec{v}_3}.$$

Da eine vollständige Basis des $\mathbf{C}^3$ aus Eigenvektoren vorliegt, resultiert nun ein komplexes Fundamentalsystem des homogenen DGl–Systems, nämlich:

$$\vec{x}_1(t) = e^{\lambda_1 t}\vec{v}_1 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}, \quad \vec{x}_2(t) = e^{\lambda_2 t}\vec{v}_2 = e^{it} \begin{bmatrix} 5 \\ 4i-7 \\ 3-i \end{bmatrix}, \quad \vec{x}_3(t) = e^{\lambda_3 t}\vec{v}_3 = e^{-it} \begin{bmatrix} 5 \\ -4i-7 \\ 3+i \end{bmatrix}.$$

Wie wir bereits in Abschnitt 11.7 festgestellt haben, gibt es zu der reellen Systemmatrix $A \in \mathbf{R}^{(n,n)}$ stets auch ein **reelles Fundamentalsystem**, da komplexe Eigenwerte nur als konjugiert komplexe Paare auftreten. Ist $\lambda \in \mathbf{C}$ ein solcher Eigenwert und $\vec{v} \in \mathbf{C}^n$ der zugeordnete komplexe Eigenvektor, so überprüft man sehr einfach, dass die komplexe Fundamentallösung $\vec{x}(t) := e^{\lambda t}\vec{v}$ in zwei reelle Fundamentallösungen zerlegbar ist, nämlich

$$\boxed{\vec{x}_1(t) := \operatorname{Re}\vec{x}(t), \qquad \vec{x}_2(t) = \operatorname{Im}\vec{x}(t).}$$

Im vorliegenden Beispiel resultiert deshalb anstelle des komplexen Fundamentalsystems das folgende **reelle Fundamentalsystem**

$$\vec{x}_1(t) = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}, \quad \vec{x}_2(t) = \cos t \begin{bmatrix} 5 \\ -7 \\ 3 \end{bmatrix} + \sin t \begin{bmatrix} 0 \\ -4 \\ 1 \end{bmatrix}, \quad \vec{x}_3(t) = \sin t \begin{bmatrix} 5 \\ -7 \\ 3 \end{bmatrix} + \cos t \begin{bmatrix} 0 \\ 4 \\ -1 \end{bmatrix}.$$

Hiermit liegt die **reelle** Fundamentalmatrix

$$Y(t) := (\vec{x}_1(t), \vec{x}_2(t), \vec{x}_3(t))$$

vor. Zur Ermittlung einer partikulären Lösung des inhomogenen DGl–Systems benötigen wir nun gemäß (6.27) die Inverse $Y^{-1}(t)$. Ihre Berechnung kann wegen der Funktionsabhängigkeit von der Variablen t sehr unangenehm werden. Für konstante Systemmatrizen $A \in \mathbf{K}^{(n,n)}$ ist jedoch die spezielle Fundamentalmatrix $Y_0(x) := e^{xA}$ wegen der folgenden Eigenschaft sehr vorteilhaft:

$$Y_0^{-1}(x) = e^{-xA} = Y_0(-x), \qquad Y_0(x)Y_0^{-1}(t) = e^{(x-t)A} = Y_0(x-t).$$

Bevor wir das vorliegende Beispiel beenden, zeigen wir die Gültigkeit einer ähnlichen Beziehung für beliebige Fundamentalmatrizen $Y(x)$, solange nur die Systemmatrix A konstant ist:

Satz 16.24 *Gegeben sei die konstante Matrix $A \in \mathbf{K}^{(n,n)}$, und es sei $Y(x)$ eine Fundamentalmatrix des homogenen DGl–Systems $\vec{y}\,'(x) = A\vec{y}(x)$, $x \in \mathbf{R}$. Dann gilt stets*

$$\boxed{Y(x) = e^{(x-x_0)A}Y(x_0) \;\; \forall\, x_0,\, x \in \mathbf{R}, \quad \textit{und somit} \quad Y(x) = e^{xA}Y(0).} \tag{6.29}$$

Ferner ist

$$\boxed{\vec{y}(x) := Y(x)\vec{c} + \int_{x_0}^{x} Y(x + x_0 - t)Y^{-1}(x_0)\vec{g}(t)\,dt, \qquad x_0,\, x \in I,} \tag{6.30}$$

die allgemeine Lösung des inhomogenen DGl–Systems $\vec{y}\,'(x) = A\vec{y}(x) + \vec{g}(x)$ mit $\vec{g} \in C(I; \mathbf{K}^n)$. Die Anfangswertaufgabe

$$\vec{y}\,'(x) = A\vec{y}(x) + \vec{g}(x), \;\; x \in I, \qquad \vec{y}(x_0) = \vec{y}_0 \in \mathbf{K}^n,$$

hat für jedes $x_0 \in I$ die eindeutig bestimmte Lösung

$$\boxed{\vec{y}(x) := Y(x)Y^{-1}(x_0)\vec{y}_0 + \int_{x_0}^{x} Y(x + x_0 - t)Y^{-1}(x_0)\vec{g}(t)\,dt, \qquad x \in I.} \tag{6.31}$$

Begründungen: (a) Es gilt für alle $x \in \mathbf{R}$ die Gleichung

$$\frac{d}{dx}\, e^{-(x-x_0)A}Y(x) = e^{-(x-x_0)A}Y'(x) - e^{-(x-x_0)A}\underbrace{AY(x)}_{=Y'(x)} = 0,$$

das heißt, wir haben

$$e^{-(x-x_0)A}Y(x) = const = e^{-(x_0-x_0)A}Y(x_0) = Y(x_0).$$

Dies führt bereits auf die behauptete Gleichung (6.29).

(b) Aus (6.29) folgern wir

$$Y(x)Y^{-1}(t) = e^{(x-x_0)A}Y(x_0)Y^{-1}(x_0)e^{-(t-x_0)A} = e^{(x-t)A} = Y(x + x_0 - t)Y^{-1}(x_0).$$

Der Rest folgt nun aus dem vorangegangenen Satz 16.23. □

BSP. (16.6.9) (Fortsetzung.) Für die reelle Fundamentalmatrix $Y(t) = (\vec{x}_1(t), \vec{x}_2(t), \vec{x}_3(t))$ gelten nun

$$Y(0) = \begin{bmatrix} 1 & 5 & 0 \\ -2 & -7 & 4 \\ 1 & 3 & -1 \end{bmatrix}, \quad Y^{-1}(0) = \frac{1}{5}\begin{bmatrix} -5 & 5 & 20 \\ 2 & -1 & -4 \\ 1 & 2 & 3 \end{bmatrix}.$$

Wir setzen zur Abkürzung

$$\vec{h}(t) := Y^{-1}(0)\vec{g}(t) = \frac{1}{5}\begin{bmatrix} -5 & 5 & 20 \\ 2 & -1 & -4 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} -2t \\ 4t \\ 6 \end{bmatrix} = \frac{2}{5}\begin{bmatrix} 15t + 60 \\ -4t - 12 \\ 3t + 9 \end{bmatrix}.$$

Unter Beachtung der Relationen

$$\int_0^t s\sin(t-s)\,ds = t - \sin t, \quad \int_0^t s\cos(t-s)\,ds = 1 - \cos t = \int_0^t \sin(t-s)\,ds, \quad \int_0^t \cos(t-s)\,ds = \sin t,$$

resultiert nun aus (6.30) die folgende partikuläre Lösung des inhomogenen DGl–Systems:

$$\vec{x}_p(t) = \int_0^t Y(t-s)\vec{h}(s)\,ds = \begin{bmatrix} 3t^2+30t-10\cos t-30\sin t+10 \\ -6t^2-50t-10\cos t+50\sin t+10 \\ 3t^2+26t-20\sin t \end{bmatrix}.$$

Die allgemeine Lösung des inhomogenen DGl–Systems ergibt sich durch Superposition:

$$\vec{x}(t) = C_1\vec{x}_1(t) + C_2\vec{x}_2(t) + C_3\vec{x}_3(t) + \vec{x}_p(t),$$

und wegen

$$Y^{-1}(0)\begin{bmatrix} x(0) \\ y(0) \\ z(0) \end{bmatrix} = \frac{1}{5}\begin{bmatrix} -5 & 5 & 20 \\ 2 & -1 & -4 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix} = \begin{bmatrix} 20 \\ -3 \\ 6 \end{bmatrix}$$

hat die Anfangswertaufgabe die eindeutig bestimmte Lösung

$$\vec{x}(t) = 20\vec{x}_1(t) - 3\vec{x}_2(t) + 6\vec{x}_3(t) + \vec{x}_p(t), \qquad t \in \mathbf{R}.$$

Es ist sicher **nicht** opportun, die Lösung des Problems (P), das heißt, die Bestimmung einer partikulären Lösung $y_p(x)$ der inhomogenen linearen DGl n–ter Ordnung (6.3)

$$L_n y := \sum_{k=0}^{n} a_k(x)y^{(k)} = f(x), \qquad x \in I,$$

in der oben dargestellten Weise durch Rückführung auf ein DGl–System 1.Ordnung zu bewerkstelligen. Man wendet hier vielmehr das **Verfahren der Variation der Konstanten** (VdK) direkt auf die DGl (6.3) an, vorausgesetzt, es ist ein Fundamentalsystem $y_1(x), y_2(x), \ldots, y_n(x)$ der homogenen DGl $L_n y = 0$ bekannt. Auch hier definieren wir eine **Fundamentalmatrix** oder Wronski–Matrix vermöge

$$Y(x) := \begin{bmatrix} y_1(x) & y_2(x) & \cdots & y_n(x) \\ y_1'(x) & y_2'(x) & \cdots & y_n'(x) \\ \vdots & \vdots & \ddots & \vdots \\ y_1^{(n-1)}(x) & y_2^{(n-1)}(x) & \cdots & y_n^{(n-1)}(x) \end{bmatrix}, \qquad W(x) := \det Y(x).$$

Wir zeigen, dass mit dem folgenden Ansatz der VdK

$$\boxed{y_p(x) := v_1(x)y_1(x) + v_2(x)y_2(x) + \cdots + v_n(x)y_n(x)} \tag{6.32}$$

und durch Lösen der Integrationsaufgabe

$$Y(x)\vec{v}\,'(x) =$$

$$\begin{bmatrix} v_1'(x)y_1(x) + v_2'(x)y_2(x) + \cdots + v_n'(x)y_n(x) \\ v_1'(x)y_1'(x) + v_2'(x)y_2'(x) + \cdots + v_n'(x)y_n'(x) \\ \vdots \\ v_1'(x)y_1^{(n-1)}(x) + v_2'(x)y_2^{(n-1)}(x) + \cdots + v_n'(x)y_n^{(n-1)}(x) \end{bmatrix} \stackrel{!}{=} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \frac{1}{a_n(x)}\,f(x) \end{bmatrix} =: \vec{g}(x) \tag{6.33}$$

tatsächlich eine partikuläre Lösung $y_p(x)$ der inhomogenen DGl $L_n y = f(x)$ bestimmt ist. Aus dem Ansatz (6.32) folgt nämlich nach n–maliger Differentiation, jeweils unter Beachtung der Nebenbedingung (6.33):

$$\left.\begin{array}{rcl|l} y_p(x) & = & v_1y_1 + v_2y_2 + \cdots + v_ny_n & \cdot a_0(x) \\ y_p'(x) & = & v_1y_1' + v_2y_2' + \cdots + v_ny_n' & \cdot a_1(x) \\ y_p''(x) & = & v_1y_1'' + v_2y_2'' + \cdots + v_ny_n'' & \cdot a_2(x) \\ & \vdots & & \cdot \\ y_p^{(n)}(x) & = & v_1y_1^{(n)} + v_2y_2^{(n)} + \cdots + v_ny_n^{(n)} + \frac{1}{a_n(x)}\,f(x) & \cdot a_n(x) \end{array}\right\} \ (+)$$

Man erhält nun

$$L_n y_p = v_1(x)\underbrace{L_n y_1}_{=0} + v_2(x)\underbrace{L_n y_2}_{=0} + \cdots + v_n(x)\underbrace{L_n y_n}_{=0} + f(x) = f(x),$$

also wie gewünscht eine partikuläre Lösung der inhomogenen DGl. Nun muss noch die Integrationsaufgabe (6.33) gelöst werden. Wegen $W(x) \neq 0$ kann die Fundamentalmatrix $Y(x)$ invertiert werden, so dass das DGl–System $\vec{v}\,'(x) = Y^{-1}(x)\vec{g}(x)$ resultiert. Setzen wir $Y^{-1}(x) := (\eta_{jk}(x))$, so ergibt sich aus der speziellen Form der Funktion $\vec{g}(x)$ komponentenweise:

$$v_j'(x) = \eta_{jn}(x) \cdot \frac{f(x)}{a_n(x)}, \quad j = 1, 2, \ldots, n.$$

Wir hatten in Satz 5.37 mit Hilfe der Cramerschen Regel die Koeffizienten einer inversen Matrix berechnet. Das Verfahren liefert hier

$$\eta_{jn}(x) = \frac{W_{nj}(x)}{W(x)}, \quad j = 1, 2, \ldots, n,$$

worin $W_{nj}(x)$ der Kofaktor der Wronski–Determinante $W(x)$ zum Platz (n, j) ist:

$$W_{nj}(x) = \begin{vmatrix} y_1(x) & \cdots & y_j(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_j'(x) & \cdots & y_n'(x) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_1^{(n-2)}(x) & \cdots & y_j^{(n-2)}(x) & \cdots & y_n^{(n-2)}(x) \\ 0 & \cdots & 1 & \cdots & 0 \end{vmatrix}.$$

In der Zusammenfassung haben wir gezeigt:

Satz 16.25 *Gegeben seien ein Intervall $I \subset \mathbf{R}$ und Koeffizientenfunktionen $a_j \in C(I)$, $1 \leq j \leq n$, mit $a_n(x) \neq 0 \ \forall\, x \in I$. Ist $y_1, y_2, \ldots, y_n \in C^n(I)$ ein Fundamentalsystem der homogenen linearen DGl*

$$L_n y := \sum_{j=0}^{n} a_j(x) y^{(j)} = 0,$$

und bezeichnet $W(x)$ die zugeordnete Wronski*–Determinante, so ist die allgemeine Lösung der inhomogenen DGl $L_n y = f \in C(I)$ in der folgenden Form gegeben:*

$$\boxed{y(x) = \sum_{j=1}^{n} y_j(x)\Big(C_j + \int_{x_0}^{x} \frac{W_{nj}(t)}{a_n(t)W(t)}\, f(t)\, dt\Big), \qquad x \in I, \quad C_j = const.} \tag{6.34}$$

Die Anfangswertaufgabe

$$L_n y = f(x), \;\; x \in I, \quad y(x_0) = y_0,\, y'(x_0) = y_1,\, \ldots,\, y^{(n-1)}(x_0) = y_{n-1}, \quad x_0 \in I,$$

hat eine eindeutig bestimmte Lösung $y \in C^n(I)$ in der Form (6.34), worin die Konstanten C_j gemäß folgender Vorschrift zu berechnen sind:

$$\begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_n \end{bmatrix} = Y^{-1}(x_0) \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix}.$$

BSP. (16.6.10) Zu bestimmen ist die allgemeine Lösung der inhomogenen DGl

$$L_2 y := \frac{1}{4}\, y'' + y = \frac{1}{\cos 2x}, \quad \cos 2x \neq 0.$$

Lösung: Die homogene DGl $L_2 y = 0$ ist die bekannte **Schwingungs–DGl** mit der allgemeinen Lösung

$$y_h(x) = C_1 \cos 2x + C_2 \sin 2x, \quad x \in \mathbf{R}.$$

Eine partikuläre Lösung der inhomogenen DGl $L_2 y = \frac{1}{\cos 2x}$ bestimmen wir mit dem Ansatz der VdK:

$$\left.\begin{array}{rcll} y_p(x) &=& v_1(x)\cos 2x + v_2(x)\sin 2x & \cdot 1 \\ y_p'(x) &=& \underbrace{v_1'\cos 2x + v_2'\sin 2x}_{=:0} - 2v_1\sin 2x + 2v_2\cos 2x & \cdot 0 \\ y_p''(x) &=& -2v_1'\sin 2x + 2v_2'\cos 2x - 4v_1\cos 2x - 4v_2\sin 2x & \cdot \frac{1}{4} \end{array}\right\} (+)$$

Nach Einsetzen in die inhomogene DGl resultiert das folgende DGl–System

$$\begin{array}{rcrcl} v_1'(x)\cos 2x &+& v_2'(x)\sin 2x &=& 0, \\ -2v_1'(x)\sin 2x &+& 2v_2'(x)\cos 2x &=& \frac{4}{\cos 2x}, \end{array}$$

welches wir mit der CRAMERschen Regel nach der Vorschrift (6.34) auflösen und anschließend integrieren:

$$W(x) = \begin{vmatrix} \cos 2x & \sin 2x \\ -2\sin 2x & 2\cos 2x \end{vmatrix} = 2,$$

$$W_{21}(x) = \begin{vmatrix} 0 & \sin 2x \\ 1 & 2\cos 2x \end{vmatrix} = -\sin 2x, \quad W_{22}(x) = \begin{vmatrix} \cos 2x & 0 \\ -2\sin 2x & 1 \end{vmatrix} = \cos 2x.$$

Hiermit resultieren:

$$v_1(x) = -2\int_{x_0}^{x} \frac{\sin 2t}{\cos 2t}\, dt = \ln\left|\frac{\cos 2x}{\cos 2x_0}\right|, \quad v_2(x) = 2\int_{x_0}^{x} dt = 2(x - x_0),$$

und daraus die allgemeine Lösung

$$y(x) = C_1\cos 2x + C_2\sin 2x + 2(x - x_0)\sin 2x + \cos 2x\, \ln\left|\frac{\cos 2x}{\cos 2x_0}\right|, \quad \cos 2x \neq 0.$$

Die Anfangswertaufgabe zum Anfangswert $y(0) = y_0$, $y'(0) = y_1$ wird gelöst, indem wir $x_0 = 0$ einsetzen und die Konstanten C_1, C_2 aus dem linearen Gleichungssystem

$$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \end{bmatrix}$$

bestimmen. Es folgt $C_1 = y_0$, $C_2 = \frac{1}{2}\, y_1$, und somit schließlich

$$y(x) = y_0\cos 2x + \frac{y_1}{2}\, \sin 2x + 2x\sin 2x + \cos 2x\, \ln|\cos 2x|, \quad x \in I := (-\frac{\pi}{4}, \frac{\pi}{4}).$$

16.7 Ergänzungen

Die Klasse derjenigen gewöhnlichen Differentialgleichungen der Ordnung $n \geq 2$

$$F(x, y, y', \ldots, y^{(n)}) = 0, \tag{7.1}$$

für die elementare Integrationsverfahren angegeben werden können, ist weitaus kleiner als die entsprechende Klasse der gewöhnlichen Differentialgleichungen 1.Ordnung. Man versucht deshalb häufig, das Verfahren der **Reduktion der Ordnung** durchzuführen mit dem Ziel, eine vorgelegte DGl in eine solche niederer Ordnung zurückzuführen. Für diese sollte dann ein Integrationsverfahren bekannt sein. Ein solches Vorgehen kann besonders dann erfolgversprechend angewendet werden, wenn einige der Veränderlichen in der DGl (7.1) nicht explizit auftreten. Wir geben nachfolgend die zwei Hauptfälle für diese Sachlage an.

Fall (A): Die Variablen y, y', $\ldots$, $y^{(k)}$, $k < n$, treten nicht explizit auf:

$$F(x, y^{(k+1)}, \ldots, y^{(n)}) = 0.$$

In diesem Fall führt der folgende Ansatz auf eine DGl $F(x, u, u', \ldots, u^{(n-k-1)}) = 0$ der Ordnung $n-k-1$ für eine neue Variable $u(x)$:

$$\boxed{u(x) = y^{(k+1)}(x).}$$

BSP. (16.7.1) Die Lösung der folgenden DGl 2.Ordnung ist zu bestimmen, in der die Variable $y(x)$ nicht explizit auftritt:

$$y'' - \frac{1}{x+1}\, y' = (x+1)^2, \qquad x \neq -1.$$

Lösung: Setzen wir $u(x) := y'(x)$, so genügt $u(x)$ der inhomogenen linearen DGl 1.Ordnung

$$u' - \frac{1}{x+1}\, u = (x+1)^2, \qquad x \neq -1,$$

die nach dem bekannten Verfahren aus Abschnitt 16.2 gelöst wird. Man erhält die allgemeine Lösung

$$u(x) = (x+1)\Big(C + \frac{1}{2}\,(x+1)^2\Big) = y'(x),$$

und eine weitere unbestimmte Integration führt auf die gesuchte Lösung

$$y(x) = C_1(x+1)^2 + C_2 + \frac{1}{8}\,(x+1)^4, \qquad x \neq -1, \quad C_1 := \frac{C}{2}.$$

BSP. (16.7.2) Zu bestimmen ist die Lösung der folgenden DGl 3.Ordnung, in der weder die Variable x noch die Variable y explizit auftreten:

$$y''' = \frac{2y'y''^2}{1+y'^2}.$$

Lösung: Auch hier setzen wir zunächst wieder $u(x) := y'(x)$, womit sich die folgende DGl 2.Ordnung ergibt:

$$u'' = \frac{2uu'^2}{1+u^2}. \tag{7.2}$$

Diesen Typ von DGln behandeln wir unter

Fall (B): Die Variable x tritt nicht explizit auf:

$$F(y, y', \ldots, y^{(n)}) = 0.$$

In diesem Fall bestimmen wir y' als Funktion der Variablen y; das heißt, wir machen den Ansatz

$$\boxed{y' = z(y) \quad \Rightarrow \quad y'' = \frac{dz}{dy}\,\frac{dy}{dx} = z \cdot z'(y), \quad y''' = z \cdot z'^2(y) + z^2 z''(y) \text{ usw.}}$$

Es resultiert eine DGl $G(y, z, z', \ldots, z^{(n-1)}) = 0$ der Ordnung $n-1$ für die Funktion $z(y)$.

BSP. (16.7.2) (Fortsetzung.) Wir kehren zurück zur DGl (7.2) und setzen dort $z(u) := u'$. Die resultierende DGl können wir mit dem Verfahren der TdV integrieren:

$$\frac{dz}{du} = u\,\frac{2z}{1+u^2} \quad \Rightarrow \quad \frac{dz}{z} = \frac{2u}{1+u^2}\,du \quad \Rightarrow \quad \ln|z| = \ln|C_1(1+u^2)|,$$

und hieraus erschließen wir die Hilfsfunktion $z(u) = C_1(1+u^2) = u'(x)$. Eine weitere Integration mit Hilfe des Verfahrens der TdV führt auf $\operatorname{arc\,tan}_H u = C_1 x + C_2$, und somit

$$u(x) = \tan(C_1 x + C_2) = y'(x).$$

Hieraus erhalten wir schließlich durch eine weitere unbestimmte Integration die gesuchte Lösung

$$y(x) = \int \tan(C_1 x + C_2)\,dx = -\frac{1}{C_1}\,\ln|\cos(C_1 x + C_2)| + C_3.$$

BSP. (16.7.3) Der obige Fall (B) enthält insbesondere explizite DGln 2.Ordnung vom Typ

$$\boxed{y'' = f(y),}$$

in dem die beiden Variablen x und y' nicht explizit auftreten. Wegen

$$\boxed{y'' = \frac{dy'}{dy}\,\frac{dy}{dx} = y' \cdot \frac{dy'}{dy} = \frac{1}{2}\,\frac{d}{dy}\,(y'^2)}$$

erhält man sofort ein sogenanntes **Energie–Integral**

$$\frac{1}{2}\,y'^2 - \int f(y)\,dy = E = const.$$

Dieser Begriff ist der Dynamik entlehnt: Für die eindimensionale Bewegung $x(t)$ einer Masse m in einem ortsabhängigen Kraftfeld $K(x)$ gilt die NEWTONsche Bewegungsgleichung $m\ddot{x} = K(x)$ und somit $\frac{m}{2}\,\dot{x}^2 - \int K(x)\,dx = const.$ Ist $K(x)$ ein **Potentialfeld** mit Potential $\varphi(x)$, das heißt, gilt $K(x) = -\frac{d\varphi}{dx}\,(x)$, so resultiert der **Energiesatz**

$$\frac{1}{2}\,m\dot{x}^2 + \varphi(x) = E = const,$$

oder in Worten

$$\boxed{\textbf{kinetische Energie + potentielle Energie = const.}}$$

Es folgen

$$\dot{x}(t) = \frac{dx}{dt} = \sqrt{\frac{2}{m}\,(E - \varphi(x))}, \qquad t - t_0 = \int_{x_0}^{x} \frac{ds}{\sqrt{\frac{2}{m}\,(E - \varphi(s))}}.$$

Ein konkretes Beispiel ist die DGl des **Fadenpendels** $m\ddot{x} = -\frac{mg}{\ell}\sin x$, die unter den Anfangsbedingungen $x(0) = x_0 > 0$ und $\dot{x}(0) = 0$ zu integrieren ist. Es gilt hier $\varphi(x) := -\frac{mg}{\ell}\cos x$, und somit $E = \frac{1}{2}m\dot{x}^2(0) + \varphi(x(0)) = -\frac{mg}{\ell}\cos x_0$. Das heißt, wir haben die Lösung $x(t)$ in impliziter Form durch folgendes Integral vorliegen:

$$\boxed{t = \sqrt{\frac{\ell}{2g}}\int_{x_0}^{x}\frac{ds}{\sqrt{\cos s - \cos x_0}}.}$$

Der Operatorenkalkül von HEAVISIDE für DGl–Systeme mit konstanten Koeffizienten.

Wir zeigen die Vorgehensweise exemplarisch auf, indem wir das folgende Beispiel eines inhomogenen DGl–Systems 2.Ordnung analysieren:

$$\begin{array}{rcrcrcl} \ddot{x} & + & 2\dot{y} & - & 3x & = & e^t, \\ \ddot{y} & + & 4\dot{x} & + & 3y & = & 0. \end{array} \tag{7.3}$$

Wird der Differentialoperator $D := \frac{d}{dt}$ eingeführt, so resultiert als formale Struktur ein inhomogenes lineares Gleichungssystem für die Bestimmung der Unbekannten x und y, dessen Koeffizienten Polynome in der Veränderlichen D sind:

$$(D^2 - 3)x + 2Dy = e^t, \tag{7.4}$$

$$4Dx + (D^2 + 3)y = 0. \tag{7.5}$$

Die algebraischen Verknüpfungen "+" und "$\cdot$" des Operators D sind denselben Rechengesetzen untergeordnet wie die Zahlen des Körpers $\mathbf{R}$ mit zwei Ausnahmen: Eine **Division** durch D ist verboten, und die **Multiplikation** mit D darf nur von links erfolgen. Insbesondere kann das "lineare Gleichungssystem" (7.4), (7.5) mit dem GAUSS–Algorithmus wie ein gewöhnliches Gleichungssystem behandelt werden, sofern man den GAUSS–Algorithmus in divisionsfreier Version verwendet: In den elementaren Zeilenoperationen darf nur **polynomiale Multiplikation** von links mit D zugelassen werden. Wir behandeln unter diesen Einschränkungen das obige System:

$$\left.\begin{array}{rrcrcrcl} -4D\cdot(7.4)\ : & -4D(D^2-3)x & - & & 8D^2y & = & -4De^t = -4e^t \\ (D^2-3)\cdot(7.5)\ : & 4D(D^2-3)x & + & (D^2-3)(D^2+3)y & = & 0 & \end{array}\right\}\ (+)$$

Wird die obige Summe gebildet, so erhält man eine einzige DGl für eine Unbekannte, nämlich

$$((D^2-3)(D^2+3) - 8D^2)y = -4e^t \quad \Leftrightarrow \quad y^{(4)} - 8\ddot{y} - 9y = -4e^t.$$

Dieser Eliminationsprozess wird häufig das **Auskoppeln** einer DGl genannt. Die hier ausgekoppelte lineare DGl 4.Ordnung mit konstanten Koeffizienten lässt sich mit den Standardverfahren des Abschnitts 10.5 integrieren. Man erhält die Lösung

$$y(t) = C_1e^{3t} + C_2e^{-3t} + C_3\cos t + C_4\sin t + \frac{1}{4}e^t,$$

die wir in die Gleichung (7.5) einsetzen und diese dann nach $\dot{x}(t)$ auflösen:

$$\dot{x}(t) = -3C_1e^{3t} - 3C_2e^{-3t} - \frac{1}{2}C_3\cos t - \frac{1}{2}C_4\sin t - \frac{1}{4}e^t.$$

Eine unbestimmte Integration führt zur Lösung

$$x(t) = -C_1e^{3t} + C_2e^{-3t} - \frac{1}{2}C_3\sin t + \frac{1}{2}C_4\cos t - \frac{1}{4}e^t + C_5,$$

worin die zusätzlich auftretende Integrationskonstante C_5 durch Einsetzen der Lösungen $x(t)$ und $y(t)$ in die Gleichung (7.4) zu $C_5 = 0$ determiniert ist.

16.8 Numerische Lösungsverfahren

Im Rahmen dieser Vorlesung kann nur auf die einfachsten Verfahren zur numerischen Integration der Anfangswertaufgabe

$$y' = f(x,y), \qquad y(x_0) = y_0,$$

eingegangen werden. Verfeinerte Verfahren müssen in speziellen Lehrveranstaltungen über die Numerik von Differentialgleichungen erlernt werden. Die Einschränkung der Grundproblemstellung auf eine gewöhnliche DGl 1.Ordnung ist vom algorithmischen Aspekt her allerdings bedeutungslos; alle vorgestellten Algorithmen gestatten ohne Zusatz eine vektorielle Formulierung; das heißt, sie sind in gleicher Weise auf DGl–Systeme 1.Ordnung anwendbar.

Beachte: (a) Vor dem Ansetzen eines numerischen Lösungsverfahrens sollte man zunächst auf analytischem Wege nach Lösungen suchen. Der numerische Apparat ist in der Regel um ein Vielfaches aufwendiger als der analytische Lösungsversuch.
(b) Hat die analytische Methode versagt, so vergewissere man sich vor Beginn numerischer Rechnungen, ob **Existenz** und **Eindeutigkeit** der Lösung gewährleistet sind. Näherungsverfahren finden nämlich häufig auch dort "Lösungen", wo überhaupt keine existieren können.

16.8.1 Einschrittverfahren zur Lösung von Anfangswertaufgaben

Zu einer ersten numerischen Methode zur Lösung der einfachsten Anfangswertaufgabe für eine gewöhnliche Differentialgleichung 1.Ordnung, nämlich

$$\boxed{y' = f(x,y), \qquad y(x_0) = y_0,} \tag{8.1}$$

gelangt man durch folgende Überlegung. Ist etwa die Lösung $y(x)$ auf dem Intervall $I := [a,b]$ mit $a := x_0$ zu bestimmen, so wählt man $n+1$ Stützstellen $x_j \in I$ in der Anordnung $a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$. Die Differenzen $h_j := x_{j+1} - x_j$ heißen **Schrittweiten** der obigen Zerlegung von I, und n heißt die **Schrittzahl**. Im Sonderfall einer äquidistanten Zerlegung des Intervalls I mit der Schrittzahl n hat man eine konstante Schrittweite $h = \frac{b-a}{n}$ und somit äquidistante Stützstellen

$$x_j = a + jh, \quad j = 0,1,\ldots,n.$$

Da der Funktionswert $f(x,y(x))$ gemäß (8.1) die Steigung $y'(x)$ der gesuchten exakten Lösung $y(x)$ im Punkt $x \in I$ angibt, kommt man zu einer Näherung, wenn die Tangentensteigung $y'(x)$ durch die Sekantensteigung $\frac{1}{h}\,(y(x+h) - y(x))$ ersetzt wird:

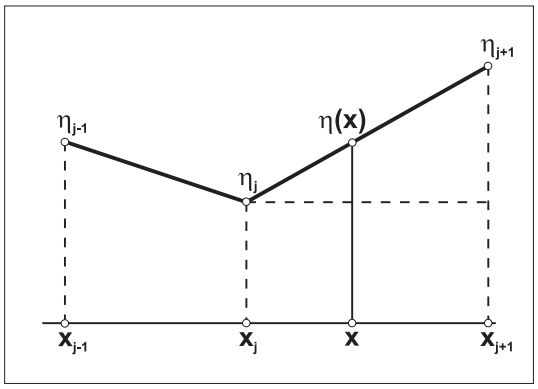
$$y(x+h) \approx y(x) + h\, f(x,y(x)).$$

Ausgehend von den Anfangswerten x_0, $y_0 = y(x_0)$, gelangt man auf der endlichen Zerlegung des Intervalls I zu Näherungswerten η_j für die Funktionswerte $y_j := y(x_j)$ der exakten Lösung:

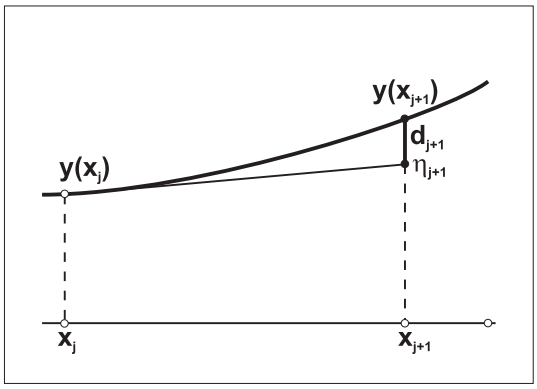
$$\boxed{\eta_0 := y_0 \quad \text{und} \quad \eta_{j+1} := \eta_j + h_j\, f(x_j,\eta_j), \quad x_{j+1} := x_j + h \;\; \forall\, j = 0,1,\ldots,n-1.} \tag{8.2}$$

Verbindet man die Knotenpunkte (x_j,η_j) und (x_{j+1},η_{j+1}) (siehe Skizze) durch eine Gerade, so erhält man als Näherungslösung zur Anfangswertaufgabe (8.1) einen **Polygonzug**, nämlich

$$\boxed{\begin{aligned} \eta(x) &:= \eta_j + \frac{1}{h_j}\,(\eta_{j+1} - \eta_j)(x - x_j) \\ &= \eta_j + f(x_j,\eta_j)(x - x_j) \;\; \forall\, x \in [x_j, x_{j+1}], \quad 0 \le j \le n-1. \end{aligned}} \tag{8.3}$$



Das Eulersche Polygonzugverfahren



Zur geometrischen Bedeutung des lokalen Diskretisationsfehlers

Definition 16.14 *Das Verfahren* (8.2) *heißt* Euler–Cauchy**sches Polygonzugverfahren**.

Bemerkung 16.12 Die Näherungswerte η_j hängen sicher von der Wahl der Schrittweiten $h_0, h_1, \ldots, h_j$ ab. Gelangt man mit anderen Schrittweiten $\tilde{h}_0, \tilde{h}_1, \ldots, \tilde{h}_j$ ebenfalls zur Stützstelle x_j, so wird der jetzt berechnete Näherungswert $\tilde{\eta}_j$ im allgemeinen von η_j verschieden sein. □

Das Euler–Cauchysche Polygonzugverfahren – kurz Euler–Verfahren – zählt zu den sogenannten **Einschrittverfahren**.

Definition 16.15 *Ein* **Einschrittverfahren** *zur näherungsweisen Lösung der Aufgabe* (8.1) *besteht in der Vorgabe einer geeigneten Funktion* $\Phi = \Phi(x, y; h; f)$ *und der Vorgabe der Berechnungsvorschrift*

$$\boxed{\begin{array}{llll} \eta_0 := y_0, & \eta_{j+1} & := & \eta_j + h_j\,\Phi(x_j, \eta_j; h_j; f) \;\; \forall\; j = 0, 1, \ldots, n-1, \\ & x_{j+1} & := & x_j + h_j, \end{array}} \tag{8.4}$$

mit deren Hilfe Näherungen η_j *für die Werte* $y_j := y(x_j)$ *der exakten Lösung bestimmt werden.*

BSP. (16.8.1) Beim Euler–Verfahren liegt gerade der Fall $\Phi(x, y; h; f) := f(x, y)$ vor. Hier ist die Funktionsvorschrift Φ von der Schrittweite h unabhängig.

BSP. (16.8.2) Eine ganze Klasse von Einschrittverfahren gewinnt man aus der Taylor–Entwicklung der exakten Lösung $y(x)$ in einer Umgebung des Startpunktes (x_0, y_0) :

$$y(x) = \sum_{k=0}^{p} \frac{1}{k!}\, y^{(k)}(x_0)\,(x - x_0)^k + R_{p+1}.$$

Vernachlässigt man das Restglied R_{p+1}, so erhält man mit der Schrittweite $h_j := x_{j+1} - x_j$ den Näherungswert η_{j+1} nach der Rechenvorschrift

$$\boxed{\eta_{j+1} = \eta_j + h_j \sum_{k=1}^{p} \frac{1}{k!}\, h_j^{k-1} y_j^{(k)} \equiv \eta_j + h_j\,\Phi(x_j, \eta_j; h_j; f).} \tag{8.5}$$

Hier beschreibt $y_j^{(k)}$ den Wert der k–ten Ableitung im Punkt (x_j, η_j). Offenbar gilt $y_j' = f(x_j, \eta_j)$ gemäß (8.1). Die zweiten und höheren Ableitungen lassen sich im Prinzip durch wiederholte Differentiation nach x und Substitution von $y'(x)$ aus (8.1) gewinnen:

$$\left.\begin{array}{rcl} y'(x) &=& f(x,y), \\ y''(x) &=& f_x(x,y) + f_y(x,y)\cdot f(x,y), \\ y'''(x) &=& f_{xx}(x,y) + 2f_{xy}(x,y)\cdot f(x,y) + f_{yy}(x,y)\cdot f^2(x,y) + f_y(x,y)\cdot y''(x), \\ & \vdots & \end{array}\right\} \tag{8.6}$$

Setzt man hier in den rechten Seiten $x = x_j$ und $y = \eta_j$ ein, so hat man formelmäßige Ausdrücke für die Größen $y_j^{(k)}$ vorliegen. Man beachte, dass die Berechnung der höheren Ableitungen $y^{(k)}(x)$ nach dem Schema (8.6) allerdings sehr rasch kompliziert wird, so dass man sich bei den Verfahren (8.5) nur auf kleine Zahlen p beschränkt. Mit dem Fall $p = 1$ liegt das EULER–Verfahren vor.

Bemerkung 16.13 Im Gegensatz zum Einschrittverfahren kann die Funktion Φ bei **Mehrschrittverfahren** auch noch von vorhergehenden Knotenstellen (x_{j-l}, η_{j-l}) abhängen. □

Konsistenz, Diskretisationsfehler, Fehlerordnung

Die Rechenvorschrift (8.4) kann sicher nur dann ein brauchbares Näherungsverfahren zur Lösung der AWA (8.1) darstellen, wenn die Funktion $\Phi(x, y; h; f)$ mit $f(x, y)$ in einer bestimmten Relation steht. Diese Relation resultiert aus (8.4) im Limes $h_j \to 0$:

$$\lim_{h_j \to 0} \frac{1}{h_j} (\eta_{j+1} - \eta_j) = y'(x_j) = \Phi(x_j, \eta_j; 0; f).$$

Definition 16.16 *Ein Einschrittverfahren* (8.4) *heiße mit der Differentialgleichung in* (8.1) **konsistent**, *falls gilt:*

$$\boxed{\Phi(x, y; 0; f) = f(x, y) \ \ \forall\, x, y.} \tag{8.7}$$

BSP. (16.8.3) Das EULER–Verfahren ist wegen $\Phi(x, y; h; f) = f(x, y)$ ganz offensichtlich konsistent.

Wegen des Zusammenhangs (8.7) verzichten wir nun bei konsistenten Einschrittverfahren auf das Argument f in der Funktion Φ. Die Wahl der Funktion Φ nimmt ganz entscheidend Einfluss auf die Güte der Approximation der exakten Lösung $y(x)$ durch das Einschrittverfahren (8.4). Gilt nämlich $\Phi(x, y; h) \neq f(x, y)$, so wird in jedem Schritt des Verfahrens (8.4) mit einer falschen Steigung gerechnet. Die dadurch verursachte Abweichung $y(x_{j+1}) - \eta_{j+1}$ wird durch den **lokalen Diskretisationsfehler** gemessen.

Definition 16.17 *Sei $y(x)$ die exakte Lösung der Anfangswertaufgabe* (8.1). *Dann heiße die Differenz*

$$\boxed{d_{j+1} := y(x_{j+1}) - y(x_j) - h\, \Phi(x_j, y(x_j); h)} \tag{8.8}$$

der **lokale Diskretisationsfehler** *an der Stelle $x_{j+1} = x_j + h$.*

BSP. (16.8.4) Lokaler Diskretisationsfehler des EULER–Verfahrens: Hat die Funktion $f(x, y)$ stetige partielle Ableitungen nach x und y bis zur Ordnung $p - 1$, so ist die Lösung $y(x)$ der AWA (8.1) sicher p–mal stetig differenzierbar, und es gilt die TAYLOR–Entwicklung

$$y(x + h) = \sum_{k=0}^{p-1} \frac{1}{k!}\, y^{(k)}(x) h^h + \frac{1}{p!}\, h^p y^{(p)}(x + \theta h), \quad 0 < \theta < 1. \tag{8.9}$$

Unter Verwendung der Formeln (8.6) resultiert

$$y(x + h) - y(x) = h\, f(x, y) + \frac{h^2}{2} \left(f_x(x, y) + f_y(x, y) \cdot f(x, y)\right) + \mathcal{O}(h^3) \ \text{ für } \ h \to 0.$$

Setzt man hier $x := x_j$ und $x_{j+1} := x_j + h$ sowie $y_j := y(x_j)$, so folgt unter Beachtung von $\Phi(x, y; h) = f(x, y)$ für das EULER–Verfahren der lokale Diskretisationsfehler an der Stelle x_{j+1}, nämlich:

$$\boxed{d_{j+1} = \frac{h^2}{2}\,(f_x(x_j, y_j) + f_y(x_j, y_j) \cdot f(x_j, y_j)) + \mathcal{O}(h^3) = \mathcal{O}(h^2) \quad \text{für} \quad h \to 0.} \tag{8.10}$$

Die Relation (8.10) des EULER–Verfahren–Beispiels veranlasst zu folgender

Definition 16.18 *Ein Einschrittverfahren* (1.4) *besitze die* **Fehlerordnung** $p \geq 0$, *falls für seinen lokalen Diskretisationsfehler* d_j *die Abschätzung*

$$\boxed{\max_{1 \leq j \leq n} |d_j| \leq const \cdot h^{p+1} = \mathcal{O}(h^{p+1}) \quad \textit{für} \quad h \to 0+} \tag{8.11}$$

gilt, wobei $f \in C^p([a, b] \times \mathbf{R})$ *vorausgesetzt wird.*

Gemäß dieser Definition hat das EULER–Verfahren die Fehlerordnung $p = 1$. Das Einschrittverfahren (8.5) hat den lokalen Diskretisationsfehler

$$d_{j+1} = y(x_{j+1}) - y(x_j) - h \sum_{k=1}^{p} \frac{1}{k!}\, h^{k-1} y_j^{(k)} = \frac{1}{(p+1)!}\, h^{p+1} y^{(p+1)}(x_j + \theta h),\ 0 < \theta < 1.$$

Für $f \in C^p([a, b] \times \mathbf{R})$ resultiert daraus eine Fehlerordnung genau p.

Die Bedeutung der Fehlerordnung p wird augenscheinlicher durch Einführung der folgenden

Definition 16.19 *Sei* $y(x)$ *die exakte Lösung der AWA* (8.1). *Dann heiße die Differenz*

$$\boxed{g_j := y(x_j) - \eta_j} \tag{8.12}$$

der **globale Diskretisationsfehler** *an der Stelle* x_j.

Der globale Diskretisationsfehler gibt den totalen Fehler an, den die mit dem Einschrittverfahren (8.4) berechnete Näherung $\eta(x)$ gegenüber der exakten Lösung $y(x)$ aufweist. Eine Abschätzung des globalen Diskretisationsfehlers lässt sich vornehmen, falls die Funktion Φ in einem geeignet gewählten Bereich

$$G := \{(x, y; h) \,:\, a \leq x \leq b,\ \ |y - y(x)| \leq \gamma,\ \ 0 \leq |h| \leq h_0\}$$

bezüglich der Variablen y LIPSCHITZ–stetig ist:

$$\boxed{|\Phi(x, y_1; h) - \Phi(x, y_2; h)| \leq L\,|y_1 - y_2| \ \ \forall\, (x, y_i; h) \in G,\ \ i = 1, 2.} \tag{8.13}$$

Setzt man jetzt $g_j = y(x_j) - \eta_j$ und $g_{j+1} = y(x_{j+1}) - \eta_{j+1}$ in (8.8) ein und verwendet die Vorschrift (8.4), so resultiert

$$g_{j+1} = g_j + h[\Phi(x_j, y(x_j); h) - \Phi(x_j, \eta_j; h)] + d_{j+1}.$$

Mit (8.13) folgt hieraus die Abschätzung

$$|g_{j+1}| \leq |g_j| + |h| L\,|y(x_j) - \eta_j| + |d_{j+1}| = (1 + |h|L)\,|g_j| + |d_{j+1}|.$$

Hat das Einschrittverfahren (8.4) eine Fehlerordnung $p \geq 0$, so gibt es nach (8.11) sicher eine Konstante D mit $\max\limits_{1 \leq j \leq n} |d_j| \leq D$ für $0 \leq |h| \leq h_0$. Deshalb erfüllt die Folge $(|g_j|)_{j \geq 0}$ die Differenzenungleichung

$$|g_{j+1}| \leq (1 + |h|L)\,|g_j| + D, \quad j = 0, 1, 2, \ldots. \tag{8.14}$$

Zur Lösung von (8.14) benötigen wir:

Satz 16.26 *Gegeben sei eine Folge $\xi_j \in \mathbf{C}$, welche einer Abschätzung der Form*

$$|\xi_{j+1}| \le (1+\delta)\,|\xi_j| + D, \quad \delta > 0, \quad D \ge 0, \quad j = 0, 1, 2, \ldots,$$

genügt. Dann gilt

$$|\xi_n| \le \frac{D}{\delta}\,\left[e^{n\delta} - 1\right] + e^{n\delta}|\xi_0| \;\; \forall\, n \in \mathbf{N}.$$

Begründung: Durch vollständige Induktion nach n zeigt man die Ungleichung

$$|\xi_n| \le (1+\delta)^n\,|\xi_0| + \frac{D}{\delta}\,((1+\delta)^n - 1).$$

Diese gilt offenbar für $n = 1$. Den Schluss von n auf $(n+1)$ vollzieht man unter Verwendung der angegebenen Differenzenungleichung. Nun beachtet man weiterhin, dass für $f(t) := 1 + t - e^t$ stets gilt: $f'(t) = 1 - e^t \le 0 \;\forall\, t \ge 0$. Deshalb folgt $f(t) \le f(0) = 0$, also $1 + t \le e^t \;\forall\, t \ge 0$. Daraus folgen die Ungleichungen $(1+\delta)^n \le e^{n\delta} \;\forall\, n \in \mathbf{N}$. □

Als unmittelbare Folgerung aus Satz 16.26 erhalten wir eine Abschätzung des globalen Diskretisationsfehlers.

Satz 16.27 *Gegeben sei ein Einschrittverfahren* (8.4) *der Fehlerordnung $p \ge 0$, welches der* Lipschitz*–Bedingung* (8.13) *genüge. Dann gilt für den globalen Diskretisationsfehler g_n an der festen Stelle $x_n := x_0 + nh$, $|h| \le h_0$, die Abschätzung*

$$\boxed{|g_n| \le |h|^p\,\frac{N}{L}\left(e^{L\,|x_n - x_0|} - 1\right),} \tag{8.15}$$

mit einer von h unabhängigen Konstanten N.

Begründung: Bei konstanter Schrittweite $h \ne 0$ kann (8.11) in der Form

$$\max_{1 \le j \le n} |d_j| \le N|h|^{p+1} \equiv D, \quad h \to 0,$$

abgeschätzt werden, mit einer nur von der Funktion $f(x, y)$ und ihren partiellen Ableitungen bis zur Ordnung p abhängigen Konstanten N. Mit dieser Zahl D in (8.14) und mit $\delta := |h| \cdot L$ erhalten wir aus Satz 16.26:

$$|g_n| \le \frac{N|h|^{p+1}}{|h|L}\left(e^{nL|h|} - 1\right) + e^{nL|h|}|g_0|.$$

Beachten wir $g_0 = y(x_0) - y_0 = 0$ sowie $|h| = \frac{1}{n}\,|x_n - x_0|$, so resultiert daraus schon die Behauptung (8.15). □

Bemerkung 16.14 (a) Man erkennt an der Abschätzung (8.15), dass die Fehlerordnung $p \ge 0$ mit der Konvergenzordnung bezüglich der Schrittweite h eines Einschrittverfahrens übereinstimmt.

(b) Insbesondere folgt auch aus (8.15), dass alle Einschrittverfahren (8.4), die der Lipschitz–Bedingung (8.13) genügen und eine Fehlerordnung $p > 0$ haben, sowohl konsistent als auch konvergent sind. Denn wegen (8.11) folgt für $p > 0$:

$$\lim_{h \to 0}\left|\frac{d_{j+1}}{h}\right| = \lim_{h \to 0}\left|\frac{1}{h}\,(y(x+h) - y(x)) - \Phi(x, y(x); h)\right| \le const \cdot \lim_{h \to 0}|h|^p = 0,$$

also

$$y'(x) = f(x, y) = \Phi(x, y; 0).$$

Das ist die Konsistenzbedingung (8.7). Wegen (8.15) gilt ferner

$$\lim_{h\to 0} |y(x_n) - \eta_n| = \lim_{h\to 0} |g_n| = 0,$$

und dies bedeutet die Konvergenz der Folge η_n bei festgehaltenem x_n gegen den exakten Wert $y(x_n)$.

(c) Die LIPSCHITZ–Bedingung (8.13) gilt zum Beispiel für Funktionen Φ mit stetiger partieller Ableitung $(\partial/\partial y)\Phi(x, y; h)$ auf dem kompakten Bereich G. Für das EULER–Verfahren mit $\Phi(x, y; h) = f(x, y)$ wird durch die Bedingung (8.13) gerade die Existenz und Eindeutigkeit einer Lösung $y(x)$ zur Anfangswertaufgabe (8.1) gemäß Satz 16.14 garantiert.

(d) Die Ungleichung (8.15) könnte dazu benutzt werden, zu gegebenem x und $\epsilon > 0$ die Schrittweite h zu ermitteln, mit der $y(x)$ bis auf einen Fehler ϵ genau berechnet werden kann. Leider lassen sich die Konstanten N und L in der Praxis nur schwer explizit bestimmen. □

16.8.2 RUNGE–KUTTA–Verfahren

Das EULER–Verfahren (8.2) mit der Fehlerordnung $p = 1$ wird bei groben Schrittweiten h_j nur eine unzureichende Näherungslösung der Anfangswertaufgabe (8.4) liefern. Um zu Einschrittverfahren höherer Fehlerordnung $p \geq 2$ zu gelangen, geht man im Prinzip von einer **Quadraturformel** für das Integral in der Identität

$$y(x_{j+1}) - y(x_j) = \int_{x_j}^{x_{j+1}} f(x, y(x))\, dx, \qquad x_{j+1} := x_j + h_j, \tag{8.16}$$

aus. Eine allgemeine Quadraturformel mit Stützstellen $\xi_1, \xi_2, \ldots, \xi_m \in [x_j, x_{j+1}]$ und zugehörigen Integrationsgewichten $a_1, a_2, \ldots, a_m$ führt zu folgendem Ansatz für die Näherung η_{j+1} des Funktionswertes $y(x_{j+1})$:

$$\eta_{j+1} = \eta_j + h_j \sum_{l=1}^{m} a_l k_l \quad \text{mit} \quad k_l := f(\xi_l, y(\xi_l)). \tag{8.17}$$

Die spezielle Wahl der **Trapezregel** als Quadraturformel mit $m = 2, a_1 = a_2 = \frac{1}{2}$ und $\xi_1 = x_j, \xi_2 = x_{j+1}$ in (8.17) führt bei Wahl von $\eta_j \approx y(x_j)$ und $\eta_{j+1} \approx y(x_{j+1})$ auf das **implizite Integrationsverfahren**:

$$\boxed{\eta_{j+1} = \eta_j + \frac{h_j}{2}\,(f(x_j, \eta_j) + f(x_{j+1}, \eta_{j+1})),} \tag{8.18}$$

in welchem der Näherungswert η_{j+1} **implizit** definiert ist. Jeder Integrationsschritt erfordert im allgemeinen die Lösung einer nichtlinearen Gleichung.

Spezialfall: Bei einer **linearen** Differentialgleichung

$$y' = a(x)y + b(x) =: f(x, y)$$

kann das Verfahren (8.18) in ein **explizites** Einschrittverfahren überführt werden. Durch Auflösen nach η_{j+1} ergibt sich

$$\eta_{j+1} = \frac{[2 + h_j\, a(x_j)]\, \eta_j + h_j\, [b(x_j) + b(x_{j+1})]}{2 - h_j\, a(x_{j+1})}, \qquad j = 0, 1, \ldots.$$

Hängt die Funktion $f(x,y)$ nichtlinear von y ab, so kann die Fixpunktgleichung (8.18) zum Beispiel durch sukzessive Approximation nach η_{j+1} aufgelöst werden. Für einen geeigneten Startwert $\eta_{j+1}^{(0)}$ konvergiert die Fixpunktiteration

$$\boxed{\eta_{j+1}^{(k+1)} = \eta_j + \frac{h_j}{2}\Big(f(x_j,\eta_j) + f(x_{j+1},\eta_{j+1}^{(k)})\Big), \qquad k = 0,1,\ldots,} \tag{8.19}$$

sofern $f(x,y)$ die übliche LIPSCHITZ–Bedingung $|f(x,y_1) - f(x,y_2)| \le L\,|y_1 - y_2|$ erfüllt und zudem $h_j L/2 < 1$ gilt. Da in (8.19) ohnehin der Funktionswert $f(x_j,\eta_j)$ zu berechnen ist, wählt man als geeigneten Startwert einen Schritt des EULER–Verfahrens (8.2):

$$\eta_{j+1}^{(0)} := \eta_j + h_j\, f(x_j,\eta_j). \tag{8.20}$$

Wird in der Fixpunktiteration (8.19) nur ein einziger Iterationsschritt ausgeführt, so resultiert mit dem Startwert (8.20) das folgende Einschritt–Verfahren:

$$\boxed{\begin{aligned} \eta_{j+1}^{(P)} &= \eta_j + h_j\, f(x_j,\eta_j), \\ \eta_{j+1} &= \eta_j + \frac{h_j}{2}\Big(f(x_j,\eta_j) + f(x_{j+1},\eta_{j+1}^{(P)})\Big). \end{aligned}} \tag{8.21}$$

Definition 16.20 *Das Verfahren* (8.21) *heißt* HEUN**sches Verfahren.** *Es gehört zu den expliziten* **Prädiktor–Korrektor–Verfahren,** *da zunächst ein* **Prädiktorwert** $\eta_{j+1}^{(P)}$ *mit dem* EULER*-Verfahren (1. Ordnung) bestimmt wird, der dann mit dem impliziten Trapezverfahren korrigiert wird.*

Das HEUNsche Verfahren (ca. 1900) lässt sich folgendermaßen algorithmisch formulieren:

$$\boxed{\begin{aligned} k_1 &:= f(x_j,\eta_j), \\ k_2 &:= f(x_j + h_j, \eta_j + h_j\, k_1); \\ \eta_{j+1} &:= \eta_j + \tfrac{1}{2}\, h_j\{k_1 + k_2\}. \end{aligned}} \tag{8.22}$$

Die Formulierung unterstreicht deutlich, dass das HEUNsche Verfahren zur Klasse der Einschrittverfahren gehört mit

$$\boxed{\Phi(x,y;h) := a_1 k_1 + a_2 k_2 = a_1 f(x,y) + a_2 f(x + q_1 h, y + q_2 h\, f(x,y)),} \tag{8.23}$$

worin speziell

$$a_1 = a_2 = \frac{1}{2}, \quad q_1 = q_2 = 1 \tag{8.24}$$

zu setzen sind. Betreffs der Fehlerordnung haben wir die folgende Aussage:

Satz 16.28 *Das mit der Funktion* Φ *aus* (8.23) *gebildete Einschrittverfahren hat mindestens die Fehlerordnung* $p = 2$, *falls* a_1, a_2, q_1 *und* q_2 *wie folgt gewählt werden:*

$$\boxed{a_1 + a_2 = 1, \qquad a_2 q_1 = a_2 q_2 = \frac{1}{2}.} \tag{8.25}$$

Begründung: Entwickelt man (8.23) an der Stelle $h = 0$ in eine TAYLOR–Reihe, so erhält man zusammen mit (8.9) für den lokalen Diskretisationsfehler:

$$\begin{aligned} d_{j+1} &= (1 - a_1 - a_2)h_j\, f(x_j, y_j) \\ &\quad + \frac{1}{2}\, h_j^2((1 - 2a_2q_1)f_x(x_j, y_j) + (1 - 2a_2q_2)f_y(x_j, y_j) \cdot f(x_j, y_j)) + \mathcal{O}(h_j^3) \ \text{ für } \ h_j \to 0. \end{aligned}$$

Ein Verfahren der Fehlerordnung $p \geq 2$ resultiert also bei Wahl von

$$0 = (1 - a_1 - a_2) = (1 - 2a_2q_1) = (1 - 2a_2q_2),$$

und diese Bedingungen sind äquivalent mit (8.25). □

Folgerung 16.3 (a) *Das* HEUN*sche Verfahren erfüllt mit dem Parametersatz* (8.24) *die Bedingungen* (8.25). *Es ist also ein Einschrittverfahren der Fehlerordnung* $p = 2$.

(b) *Mit der Parameterwahl* $a_1 := 0,\ a_2 := 1;\ q_1 = q_2 = \frac{1}{2}$ *erhält man aus dem Ansatz* (8.23) *ein weiteres Einschrittverfahren der Fehlerordnung* $p = 2$, *nämlich*

$$\boxed{\Phi(x, y; h) := f\Big(x + \frac{h}{2}, y + \frac{h}{2}\, f(x, y)\Big).} \tag{8.26}$$

Das durch (8.26) *definierte Einschrittverfahren heißt* **modifiziertes** EULER**-Verfahren** (L.COLLATZ, *1960*). *Es lässt sich wie folgt algorithmisch formulieren:*

$$\boxed{\begin{aligned} k_1 &:= f(x_j, \eta_j), \\ k_2 &:= f(x_j + \tfrac{1}{2}\, h_j, \eta_j + \tfrac{1}{2}\, h_j k_1); \\ \eta_{j+1} &:= \eta_j + h_j k_2. \end{aligned}} \tag{8.27}$$

Bemerkung 16.15 Einschrittverfahren, bei denen die Funktion $f(x, y)$ pro Integrationsschritt k–mal ausgewertet werden muss, heißen **k–stufige Verfahren.** Das HEUN–Verfahren (HV) und das modifizierte EULER–Verfahren (MEV) sind also **zweistufige Einschrittverfahren.** Sie gehören außerdem zur Klasse der RUNGE–KUTTA–Verfahren. □

Allgemeine RUNGE–KUTTA–Verfahren resultieren aus dem Ansatz (8.17), wenn die Integrationsstützstellen ξ_l gemäß

$$\xi_1 := x_j, \quad \xi_l := x_j + q_l h_j \ \ \forall\ 2 \leq l \leq m \ \text{ mit } \ 0 < q_l \leq 1 \tag{8.28}$$

fixiert werden und die unbekannten Funktionswerte $y(\xi_l)$ nach der Idee des Prädiktorverfahrens festgelegt werden. Zunächst setzt man $y(\xi_1) \approx \eta_j$, ferner:

$$y(\xi_l) \approx y_l^* := \eta_j + h_j b_{l1}\, f(x_j, \eta_j) + h_j \sum_{r=2}^{l-1} b_{lr} f(x_j + q_r h_j, y_r^*), \qquad 2 \leq l \leq m. \tag{8.29}$$

Für die spezielle Differentialgleichung $y' = 1$ mit der Lösung $y(x) = C + x$ hat man $y(x_j) = C + x_j = \eta_j$ sowie $y(\xi_l) = \eta_j + q_l h_j$. Wegen $y_l^* = \eta_j + h_j \sum\limits_{r=1}^{l-1} b_{lr}$ werden in diesem Fall die Prädiktorwerte y_l^* exakt, wenn man fordert

$$\boxed{q_l = \sum_{r=1}^{l-1} b_{lr}, \qquad 2 \leq l \leq m.} \tag{8.30}$$

Diese Forderung ist mehr oder minder künstlich; sie schränkt aber die Parametermannigfaltigkeit ein.

Bemerkung 16.16 Aus (8.17) und (8.29) ergibt sich eine rekursive Definition der Funktionswerte $k_l = f(\xi_l, y_l^*)$ gemäß

$$\left.\begin{aligned} k_1 &:= f(x_j, \eta_j), \\ k_l &:= f\Big(x_j + q_l h_j, \eta_j + h_j \sum_{r=1}^{l-1} b_{lr} k_r\Big) \quad \text{für} \quad l = 2, 3, \ldots, m. \end{aligned}\right\} \tag{8.31}$$

Die Berechnung von (8.17) erfordert also im allgemeinen in jedem Schritt m Funktionsauswertungen, so dass der Algorithmus (8.17) ein m–stufiges RUNGE–KUTTA–Verfahren (RK–Verfahren) darstellt. Das EULER–Verfahren ist ein einstufiges RK–Verfahren. □

Dreistufige RK–Verfahren. Gemäß den Formeln (8.17) und (8.31) können dreistufige RK–Verfahren in der folgenden algorithmischen Form definiert werden:

$$\boxed{\begin{aligned} k_1 &:= f(x_j, \eta_j), \\ k_2 &:= f(x_j + q_2 h_j, \eta_j + h_j b_{21} k_1), \\ k_3 &:= f(x_j + q_3 h_j, \eta_j + h_j\,(b_{31} k_1 + b_{32} k_2)); \\ \eta_{j+1} &:= \eta_j + h_j\{a_1 k_1 + a_2 k_2 + a_3 k_3\}. \end{aligned}} \tag{8.32}$$

Für die acht Parameter $a_1, a_2, a_3,\ b_{21}, b_{31}, b_{32},\ q_2, q_3$, gelten momentan nur die zwei Bedingungen (8.30), das sind

$$q_2 = b_{21}, \quad q_3 = b_{31} + b_{32}. \tag{8.33}$$

Wir verschaffen uns weitere Bedingungen durch die Forderung, dass das Verfahren (8.32) mindestens eine Fehlerordnung $p = 3$ aufweisen soll. Dazu ist der lokale Diskretisationsfehler

$$d_{j+1} = y(x_{j+1}) - y(x_j) - h_j\,\{a_1 \overline{k}_1 + a_2 \overline{k}_2 + a_3 \overline{k}_3\}, \qquad \overline{k}_l := f(\xi_l, y(\xi_l)), \tag{8.34}$$

zu berechnen.

Satz 16.29 *Das gemäß* (8.32) *definierte dreistufige* RUNGE–KUTTA*–Verfahren hat mindestens die Fehlerordnung $p = 3$, falls neben den Gleichungen* (8.33) *noch die folgenden Bedingungen erfüllt sind:*

$$\boxed{a_1 + a_2 + a_3 = 1; \quad a_2 q_2 + a_3 q_3 = \frac{1}{2}; \quad a_3 q_2 b_{32} = \frac{1}{6}; \quad a_2 q_2^2 + a_3 q_3^2 = \frac{1}{3}.} \tag{8.35}$$

Begründung: Wir verwenden in den folgenden Ausdrücken TAYLOR–Entwicklungen an der Stelle $h := h_j = 0$ und schreiben f anstelle von $f(x_j, y(x_j))$. Es gilt mit (8.33):

$$\begin{aligned} \overline{k}_1 &= f(x_j, y(x_j)) = f; \\ \overline{k}_2 &= f[x_j + q_2 h, y(x_j) + q_2 h\, f(x_j, y(x_j))] \\ &= f + q_2 h\,(f_x + f \cdot f_y) + \tfrac{1}{2}\,(q_2 h)^2\,(f_{xx} + 2f \cdot f_{xy} + f^2 \cdot f_{yy}) + \mathcal{O}(h^3) \\ &=: f + q_2 h \cdot F + \tfrac{1}{2}\,(q_2 h)^2 \cdot G + \mathcal{O}(h^3); \end{aligned}$$

$$\begin{aligned} \overline{k}_3 &= f(x_j + q_3 h, y(x_j) + h\,(b_{31}\overline{k}_1 + b_{32}\overline{k}_2)) \\ &= f + q_3 h\, f_x + h\,(b_{31}\overline{k}_1 + b_{32}\overline{k}_2)\, f_y \\ &\quad + \tfrac{1}{2}\,(q_3 h)^2\, f_{xx} + q_3\,(b_{31}\overline{k}_1 + b_{32}\overline{k}_2)\, h^2\, f_{xy} + \tfrac{1}{2}\,(b_{31}\overline{k}_1 + b_{32}\overline{k}_2)^2 h^2\, f_{yy} + \mathcal{O}(h^3) \\ &= f + h\,(q_3 f_x + (b_{31} + b_{32})\, f \cdot f_y) \\ &\quad + h^2\,(q_2 b_{32} F \cdot f_y + \tfrac{1}{2}\, q_3^2\, f_{xx} + q_3\,(b_{31} + b_{32})\, f \cdot f_{xy} + \tfrac{1}{2}\,(b_{31} + b_{32})^2 f^2 \cdot f_{yy}) + \mathcal{O}(h^3) \\ &\overset{(8.33)}{=} f + q_3 h\, F + h^2\,(q_2 b_{32} F \cdot f_y + \tfrac{1}{2}\, q_3^2 G) + \mathcal{O}(h^3). \end{aligned}$$

Berücksichtigt man die TAYLOR–Formel

$$y(x_j+h)-y(x_j)=hf+\frac{1}{2}h^2F+\frac{1}{6}h^3\,(f_y\cdot F+G)+\mathcal{O}(h^4),$$

so erhält man aus (8.34) die folgende Darstellung des lokalen Diskretisationsfehlers:

$$\left.\begin{aligned} d_{j+1} \;=\;& (1-a_1-a_2-a_3)h\cdot f+(\tfrac{1}{2}-a_2q_2-a_3q_3)h^2\cdot F \\ &+\Big((\tfrac{1}{6}-a_3q_2b_{32})\,F\cdot f_y+(\tfrac{1}{6}-\tfrac{1}{2}\,a_2q_2^2-\tfrac{1}{2}\,a_3q_3^2)\,G\Big)+\mathcal{O}(h^4). \end{aligned}\right\} \tag{8.36}$$

Sind die vier nichtlinearen Gleichungen (8.35) erfüllt, so ergibt sich gerade $d_{j+1}=\mathcal{O}(h^4)$ für $h\to 0$, also eine Fehlerordnung von mindestens $p=3$. □

Bemerkung 16.17 Im dreistufigen RK–Verfahren (8.32) kann die Fehlerordnung $p=4$ **nicht** erreicht werden. Die explizite Berechnung der Terme $\mathcal{O}(h^4)$ in der TAYLOR–Entwicklung (8.36) zeigt das Auftreten eines Gliedes, das von den Parametern a_r, b_{rs}, q_s unabhängig ist und das im allgemeinen nicht verschwindet. □

Die Gleichungen (8.35) definieren eine zweiparametrige Mannigfaltigkeit. Unter den Einschränkungen $q_2\neq q_3$ und $q_2\neq\frac{2}{3}$ kann die (q_2,q_3)-Ebene als Parameterraum verwendet werden. Man erhält dann mit den verfahrensorientierten Restriktionen $0<q_l\leq 1$:

$$\boxed{\begin{array}{lll} a_1=\dfrac{2+6q_2q_3-3(q_2+q_3)}{6q_2q_3}, & a_2=\dfrac{3q_3-2}{6q_2(q_3-q_2)}, & a_3=\dfrac{2-3q_2}{6q_3(q_3-q_2)}; \\[2ex] b_{21}=q_2, & b_{31}=q_3-b_{32}, & b_{32}=\dfrac{q_3(q_3-q_2)}{q_2(2-3q_2)}. \end{array}} \tag{8.37}$$

Im Rahmen dieser Bildungsgesetze liefert (8.32) für jede Parameterwahl $0<q_2,q_3\leq 1$ mit $\frac{2}{3}\neq q_2\neq q_3$ ein dreistufiges RK–Verfahren der Fehlerordnung $p=3$. Aus historischen Gründen zitieren wir hier *zwei Sonderfälle.*

HEUNsches Verfahren dritter Ordnung. Dieses resultiert aus (8.32), wenn in (8.37) $q_2:=\frac{1}{3}$ und $q_3:=\frac{2}{3}$ gesetzt werden. Man erhält

$$a_1=\frac{1}{4},\quad a_2=0,\quad a_3=\frac{3}{4},\quad b_{21}=\frac{1}{3},\quad b_{31}=0,\quad b_{32}=\frac{2}{3},$$

und mit diesen Parametern folgt der Algorithmus:

$$\boxed{\begin{aligned} k_1 &:= f(x_j,\eta_j),\\ k_2 &:= f(x_j+\tfrac{1}{3}\,h_j,\eta_j+\tfrac{1}{3}\,h_jk_1),\\ k_3 &:= f(x_j+\tfrac{2}{3}\,h_j,\eta_j+\tfrac{2}{3}\,h_jk_2);\\ \eta_{j+1} &:= \eta_j+\tfrac{1}{4}\,h_j\{k_1+3k_3\}. \end{aligned}} \tag{8.38}$$

KUTTA–Verfahren dritter Ordnung. Wir setzen in (8.37) $q_2=\frac{1}{2}$ und $q_3=1$. Dann ergeben sich die Parameterwerte

$$a_1=\frac{1}{6},\quad a_2=\frac{2}{3},\quad a_3=\frac{1}{6},\quad b_{21}=\frac{1}{2},\quad b_{31}=-1,\quad b_{32}=2.$$

Mit diesen Parametern resultiert aus (8.32) der Algorithmus

$$\boxed{\begin{array}{lll} k_1 & := & f(x_j, \eta_j), \\ k_2 & := & f(x_j + \frac{1}{2}\,h_j, \eta_j + \frac{1}{2}\,h_j k_1), \\ k_3 & := & f(x_j + h_j, \eta_j - h_j k_1 + 2h_j k_2); \\ \eta_{j+1} & := & \eta_j + \frac{1}{6}\,h_j\{k_1 + 4k_2 + k_3\}. \end{array}} \tag{8.39}$$

Offenbar dient hier die SIMPSON–Regel als Korrektor.

Prinzip der Schrittweitensteuerung. Die absolute Größe des lokalen Diskretisationsfehlers d_{j+1} kann dazu verwendet werden, die Schrittweite h_{j+1} für den nächstfolgenden Integrationsschritt festzulegen. Dazu setzt man *a priori* eine Toleranzgrenze $\epsilon > 0$ fest. Wird im j–ten Integrationsschritt mit der Schrittweite h_j gerechnet, so setzt man $h_{j+1} := 2h_j$, falls $|d_{j+1}|$ die Toleranzgrenze um eine Zehnerpotenz unterschreitet, das heißt, falls $|d_{j+1}| < 0.1\epsilon$ gilt. Gilt hingegen $|d_{j+1}| > \epsilon$, so führe man den j–ten Integrationsschritt erneut aus, jedoch mit halber Schrittweite $\tilde{h}_j := h_j/2$.

Die praktische Realisierung der Schrittweitensteuerung erfordert eine einfache Methode zur näherungsweisen Berechnung des lokalen Diskretisationsfehlers d_{j+1}. Die folgende Strategie führt am schnellsten zum Ziel. Zu einem RK–Verfahren mit der Fehlerordnung p_0 und den Funktionskoeffizienten k_l sucht man ein RK–Verfahren mit einer **höheren** Fehlerordnung $p > p_0$, welches **dieselben** Funktionskoeffizienten k_l verwendet. Der lokale Diskretisationsfehler d_{j+1} lässt sich dann durch das Verfahren höherer Ordnung schätzen, wobei allerdings noch zusätzliche Funktionskoeffizienten k_l zu berechnen sind. Nun wäre es völlig widersinnig, wenn der Mehraufwand zur Berechnung der zusätzlichen Koeffizienten nicht nutzbringend zur Genauigkeitssteigerung eingebracht würde. Deshalb wird man den neuen Näherungswert η_{j+1} nach dem genaueren RK–Verfahren der Fehlerordnung p berechnen und **nicht** nach dem schlechteren Verfahren der Fehlerordnung p_0.

BSP. (16.8.5) Das modifizierte EULER–Verfahren (8.27) hat den lokalen Diskretisationsfehler

$$d_{j+1}^{(MEV)} = y(x_{j+1}) - y(x_j) - h_j\overline{k}_2;$$

hingegen gilt für das KUTTA–Verfahren (8.39)

$$d_{j+1}^{(K)} = y(x_{j+1}) - y(x_j) - \frac{1}{6}\,h_j\left\{\overline{k}_1 + 4\overline{k}_2 + \overline{k}_3\right\}, \qquad \overline{k}_l := f(\xi_l, y(\xi_l)).$$

Die unbekannten Werte $\overline{k}_l$ ersetzt man durch die bekannten Näherungen k_l. Da für beide Verfahren der Wert $k_2 = f(x_j + \frac{1}{2}\,h_j, \eta_j + \frac{1}{2}\,h_j k_1)$ übereinstimmt, resultiert durch Elimination von $y(x_{j+1}) - y(x_j)$:

$$d_{j+1}^{(MEV)} \approx \frac{1}{6}\,h_j\,\{k_1 - 2k_2 + k_3\} + d_{j+1}^{(K)} = \frac{1}{6}\,h_j\,\{k_1 - 2k_2 + k_3\} + \mathcal{O}(h_j^4). \tag{8.40}$$

Der Ausdruck $\frac{1}{6}\,h_j\{k_1 - 2k_2 + k_3\}$, in welchem lediglich k_3 zusätzlich zu berechnen ist, liefert somit einen Schätzwert für $d_{j+1}^{(MEV)}$.

BSP. (16.8.6) Man bestimme zum HEUNschen Verfahren (8.22) einen Schätzwert für den lokalen Diskretisationsfehler

$$d_{j+1}^{(H2)} = y(x_{j+1}) - y(x_j) - \frac{1}{2}h_j\,\{\overline{k}_1 + \overline{k}_2\}.$$

Man ersetzt wieder $\overline{k}_l$ durch k_l und sucht wegen (8.22) ein RK–Verfahren 3.Ordnung mit dem Funktionsterm $k_2 = f(x_j + h_j, \eta_j + h_j k_1)$. Wegen (8.32) und (8.33) muss die Wahl $q_2 = b_{21} = 1$ getroffen

werden. Im Rahmen der Gleichungen (8.37) kann jetzt $0 < q_3 < 1$ beliebig fixiert werden. Zum Beispiel resultiert mit $q_3 = \frac{1}{2}$ aus (8.37) der Parametersatz

$$a_1 = \frac{1}{6}, \quad a_2 = \frac{1}{6}, \quad a_3 = \frac{2}{3}, \quad b_{21} = 1, \quad b_{31} = \frac{1}{4}, \quad b_{32} = \frac{1}{4}.$$

Wir erhalten das folgende, mit dem HEUNschen Verfahren (8.22) kompatible RK–Verfahren dritter Ordnung:

$$\boxed{\begin{aligned} k_1 &:= f(x_j, \eta_j), \\ k_2 &:= f(x_j + h_j, \eta_j + h_j k_1), \\ k_3 &:= f(x_j + \tfrac{1}{2} h_j, \eta_j + \tfrac{1}{4} h_j (k_1 + k_2)); \\ \eta_{j+1} &:= \eta_j + \tfrac{1}{6} h_j \{k_1 + k_2 + 4k_3\}. \end{aligned}} \tag{8.41}$$

Analog zu (8.40) erhalten wir

$$d_{j+1}^{(H2)} \approx \frac{1}{6} h_j \{k_1 + k_2 + 4k_3\} - \frac{1}{2} h_j \{k_1 + k_2\} + \mathcal{O}(h_j^4) = \frac{1}{3} h_j \{-k_1 - k_2 + 2k_3\} + \mathcal{O}(h_j^4).$$

Somit liefert der Term $\frac{1}{3} h_j \{-k_1 - k_2 + 2k_3\}$ einen brauchbaren Schätzwert für den lokalen Diskretisationsfehler $d_{j+1}^{(H2)}$.

m–stufige RK–Verfahren mit $m \geq 4$. Gemäß (8.17) und (8.31) hat das allgemeine vierstufige RUNGE–KUTTA–Verfahren die algorithmische Form

$$\left.\begin{aligned} k_1 &:= f(x_j, \eta_j), \\ k_2 &:= f(x_j + q_2 h_j, \eta_j + h_j b_{21} k_1), \\ k_3 &:= f(x_j + q_3 h_j, \eta_j + h_j (b_{31} k_1 + b_{32} k_2)), \\ k_4 &:= f(x_j + q_4 h_j, \eta_j + h_j (b_{41} k_1 + b_{42} k_2 + b_{43} k_3)); \\ \eta_{j+1} &:= \eta_j + h_j \{a_1 k_1 + a_2 k_2 + a_3 k_3 + a_4 k_4\}. \end{aligned}\right\} \tag{8.42}$$

Die dreizehn freien Parameter sind zunächst nur durch die drei Nebenbedingungen (8.30), nämlich

$$q_l = \sum_{r=1}^{l-1} b_{lr} \quad \text{für} \quad l = 2, 3, 4, \tag{8.43}$$

restringiert. Aus der Forderung, dass das Verfahren (8.42) mindestens die Fehlerordnung $p = 4$ hat, ergeben sich weitere acht Bestimmungsgleichungen, so dass als Lösungsmenge eine zweiparametrige Mannigfaltigkeit resultiert. Analog zu (8.37) können auch hier q_2 und q_3 als frei wählbare Parameter verwendet werden. Wir verweisen auf die Spezialliteratur (zum Beispiel: R.D. GRIGORIEFF: Numerik gewöhnlicher Differentialgleichungen, Band 1. Stuttgart: Teubner Verlag 1972).

Wir zitieren aus der Literatur die drei folgenden vierstufigen RK–Verfahren mit der Fehlerordnung $p = 4$:

Klassisches RK–Verfahren 4.Ordnung. Dieses hat die algorithmische Form:

$$\boxed{\begin{aligned} k_1 &:= f(x_j, \eta_j), \\ k_2 &:= f(x_j + \tfrac{1}{2} h_j, \eta_j + \tfrac{1}{2} h_j k_1), \\ k_3 &:= f(x_j + \tfrac{1}{2} h_j, \eta_j + \tfrac{1}{2} h_j k_2), \\ k_4 &:= f(x_j + h_j, \eta_j + h_j k_3); \\ \eta_{j+1} &:= \eta_j + \tfrac{1}{6} h_j \{k_1 + 2k_2 + 2k_3 + k_4\}. \end{aligned}} \tag{8.44}$$

$\frac{3}{8}$–RK–Verfahren 4.Ordnung. Dieses Verfahren verwendet die $\frac{3}{8}$–Integrationsregel der NEWTON–CÔTES–Formeln (vgl. Tabelle in Abschnitt 8.5). Es hat die algorithmische Form:

$$\begin{aligned} k_1 &:= f(x_j, \eta_j), \\ k_2 &:= f(x_j + \tfrac{1}{3} h_j, \eta_j + \tfrac{1}{3} h_j k_1), \\ k_3 &:= f(x_j + \tfrac{2}{3} h_j, \eta_j - \tfrac{1}{3} h_j k_1 + h_j k_2), \\ k_4 &:= f(x_j + h_j, \eta_j + h_j(k_1 - k_2 + k_3)); \\ \eta_{j+1} &:= \eta_j + \tfrac{1}{8} h_j \{k_1 + 3k_2 + 3k_3 + k_4\}. \end{aligned} \tag{8.45}$$

RK–Verfahren 4.Ordnung nach ENGLAND. Der Algorithmus lautet:

$$\begin{aligned} k_1 &:= f(x_j, \eta_j), \\ k_2 &:= f(x_j + \tfrac{1}{2} h_j, \eta_j + \tfrac{1}{2} h_j k_1), \\ k_3 &:= f(x_j + \tfrac{1}{2} h_j, \eta_j + \tfrac{1}{4} h_j (k_1 + k_2)), \\ k_4 &:= f(x_j + h_j, \eta_j + h_j(-k_2 + 2k_3)); \\ \eta_{j+1} &:= \eta_j + \tfrac{1}{6} h_j \{k_1 + 4k_3 + k_4\}. \end{aligned} \tag{8.46}$$

Bemerkung 16.18 (a) Das Prinzip der Schrittweitensteuerung bei RK–Verfahren 3.Ordnung durch ein RK–Verfahren 4.Ordnung ist unmöglich. Die entsprechenden Parametersätze lassen keine Lösungen zu, für die k_1, k_2 und k_3 in beiden Verfahren identisch sind. Man findet aber geschickte Varianten der Schrittweitensteuerung in der Literatur.

(b) In den bisherigen Fällen war die erreichbare Fehlerordnung p eines expliziten m–stufigen RK–Verfahrens stets durch $p = m,\ m = 1, 2, 3, 4$, charakterisiert. Für $m \geq 5$ gilt diese Relation nicht mehr. J.C. BUTCHER ("On the attainable order of Runge–Kutta methods." Math.Comp. **19** (1965), 408–417) hat die folgende Tabelle der maximalen Fehlerordnungen berechnet:

$m =$	1	2	3	4	5	6	7	8	9
$p =$	1	2	3	4	4	5	6	6	7

(c) Zu einem vierstufigen RK–Verfahren 4.Ordnung kann ein Schätzwert für den lokalen Diskretisationsfehler d_{j+1} wieder mit einem Verfahren 5.Ordnung gewonnen werden. Allerdings ist dafür nach obiger Tabelle ein sechsstufiges RK–Verfahren erforderlich. Ein kompatibles Paar wurde von R. ENGLAND ("Error estimates for Runge–Kutta type solutions to systems of ordinary differential equations." Comp. J. **12** (1969), 166–170) angegeben. Das vierstufige Verfahren (8.46) kann durch Hinzunahme der folgenden Formeln zu einem sechsstufigen RK–Verfahren 5.Ordnung erweitert werden:

$$\begin{aligned} k_5 &:= f(x_j + \tfrac{2}{3} h_j, \eta_j + \tfrac{1}{27} h_j (7k_1 + 10k_2 + k_4)), \\ k_6 &:= f(x_j + \tfrac{1}{5} h_j, \eta_j + \tfrac{1}{625} h_j (28k_1 - 125k_2 + 546k_3 + 54k_4 - 378k_5)); \\ \eta_{j+1} &:= \eta_j + \tfrac{1}{336} h_j \{14k_1 + 35k_4 + 162k_5 + 125k_6\}. \end{aligned} \tag{8.47}$$

Für den lokalen Diskretisationsfehler $d_{j+1}^{(ENGL)}$ des Verfahrens (8.46) resultiert aus diesen Formeln der Schätzwert □

$$\boxed{d_{j+1}^{(ENGL)} \approx \frac{1}{336}\, h_j\{-42k_1 - 224k_3 - 21k_4 + 162k_5 + 125k_6\} + \mathcal{O}(h_j^6).} \tag{8.48}$$

Tabelle der numerischen Näherungslösungen:

j	Stützstelle x_j	Näherung $\eta_j \approx y(x_j)$	Schrittweite h_j	lokaler Fehler
1	$0.000\,000\mathrm{E}^{+00}$	$5.310\,100\mathrm{E}^{-01}$	$1.000\,000\mathrm{E}^{-02}$	$-1.039\,422\mathrm{E}^{-14}$
2	$1.000\,000\mathrm{E}^{-02}$	$5.256\,999\mathrm{E}^{-01}$	$2.000\,000\mathrm{E}^{-02}$	$-1.385\,896\mathrm{E}^{-14}$
3	$3.000\,000\mathrm{E}^{-02}$	$5.150\,797\mathrm{E}^{-01}$	$4.000\,000\mathrm{E}^{-02}$	$-4.157\,690\mathrm{E}^{-14}$
4	$7.000\,000\mathrm{E}^{-02}$	$4.938\,393\mathrm{E}^{-01}$	$8.000\,000\mathrm{E}^{-02}$	$8.315\,380\mathrm{E}^{-14}$
5	$1.500\,000\mathrm{E}^{-01}$	$4.513\,585\mathrm{E}^{-01}$	$1.600\,000\mathrm{E}^{-01}$	$-1.108\,717\mathrm{E}^{-13}$
6	$3.100\,000\mathrm{E}^{-01}$	$3.663\,969\mathrm{E}^{-01}$	$3.200\,000\mathrm{E}^{-01}$	$-1.552\,204\mathrm{E}^{-12}$
7	$6.300\,000\mathrm{E}^{-01}$	$1.964\,737\mathrm{E}^{-01}$	$3.200\,000\mathrm{E}^{-01}$	$-5.469\,302\mathrm{E}^{-10}$
8	$9.500\,000\mathrm{E}^{-01}$	$2.655\,053\mathrm{E}^{-02}$	$4.000\,000\mathrm{E}^{-02}$	$-8.854\,216\mathrm{E}^{-10}$
9	$9.900\,000\mathrm{E}^{-01}$	$5.310\,191\mathrm{E}^{-03}$	$5.000\,000\mathrm{E}^{-03}$	$-5.537\,003\mathrm{E}^{-11}$
10	$9.950\,000\mathrm{E}^{-01}$	$2.655\,196\mathrm{E}^{-03}$	$2.500\,000\mathrm{E}^{-03}$	$-8.920\,324\mathrm{E}^{-11}$
11	$9.975\,000\mathrm{E}^{-01}$	$1.327\,761\mathrm{E}^{-03}$	$1.250\,000\mathrm{E}^{-03}$	$-1.440\,544\mathrm{E}^{-10}$
12	$9.987\,500\mathrm{E}^{-01}$	$6.641\,437\mathrm{E}^{-04}$	$6.250\,000\mathrm{E}^{-04}$	$-2.317\,342\mathrm{E}^{-10}$
13	$9.993\,750\mathrm{E}^{-01}$	$3.324\,959\mathrm{E}^{-04}$	$3.125\,000\mathrm{E}^{-04}$	$-3.714\,869\mathrm{E}^{-10}$
14	$9.996\,875\mathrm{E}^{-01}$	$1.669\,315\mathrm{E}^{-04}$	$1.562\,500\mathrm{E}^{-04}$	$-5.859\,218\mathrm{E}^{-10}$
15	$9.998\,437\mathrm{E}^{-01}$	$8.456\,570\mathrm{E}^{-05}$	$7.812\,500\mathrm{E}^{-05}$	$-8.781\,929\mathrm{E}^{-10}$
16	$9.999\,218\mathrm{E}^{-01}$	$4.404\,360\mathrm{E}^{-05}$	$3.906\,250\mathrm{E}^{-05}$	$-1.081\,854\mathrm{E}^{-09}$
17	$9.999\,609\mathrm{E}^{-01}$	$2.479\,428\mathrm{E}^{-05}$	$1.953\,125\mathrm{E}^{-05}$	$-4.221\,097\mathrm{E}^{-10}$
18	$9.999\,804\mathrm{E}^{-01}$	$1.655\,371\mathrm{E}^{-05}$	$9.765\,625\mathrm{E}^{-06}$	$8.289\,263\mathrm{E}^{-10}$
19	$9.999\,902\mathrm{E}^{-01}$	$1.382\,501\mathrm{E}^{-05}$	$9.765\,632\mathrm{E}^{-06}$	$8.542\,698\mathrm{E}^{-10}$
20	$1.000\,000\mathrm{E}^{+00}$	$1.312\,356\mathrm{E}^{-05}$		

Tabelle der numerischen Näherungslösungen:

j	Stützstelle x_j	Näherung $\eta_j \approx y(x_j)$	Schrittweite h_j	lokaler Fehler
1	$0.000\,000\mathrm{E}^{+00}$	$5.310\,100\mathrm{E}^{-01}$	$1.000\,000\mathrm{E}^{-02}$	$0.000\,000\mathrm{E}^{+00}$
2	$1.000\,000\mathrm{E}^{-02}$	$5.256\,999\mathrm{E}^{-01}$	$2.000\,000\mathrm{E}^{-02}$	$0.000\,000\mathrm{E}^{+00}$
3	$3.000\,000\mathrm{E}^{-02}$	$5.150\,797\mathrm{E}^{-01}$	$4.000\,000\mathrm{E}^{-02}$	$9.701\,276\mathrm{E}^{-14}$
4	$7.000\,000\mathrm{E}^{-02}$	$4.938\,393\mathrm{E}^{-01}$	$8.000\,000\mathrm{E}^{-02}$	$6.548\,361\mathrm{E}^{-13}$
5	$1.500\,000\mathrm{E}^{-01}$	$4.513\,585\mathrm{E}^{-01}$	$1.600\,000\mathrm{E}^{-01}$	$8.197\,578\mathrm{E}^{-12}$
6	$3.100\,000\mathrm{E}^{-01}$	$3.663\,969\mathrm{E}^{-01}$	$3.200\,000\mathrm{E}^{-01}$	$2.480\,616\mathrm{E}^{-10}$
7	$6.300\,000\mathrm{E}^{-01}$	$1.964\,737\mathrm{E}^{-01}$	$1.600\,000\mathrm{E}^{-01}$	$2.767\,289\mathrm{E}^{-10}$
9	$9.500\,000\mathrm{E}^{-01}$	$2.655\,052\mathrm{E}^{-02}$	$2.000\,000\mathrm{E}^{-02}$	$6.048\,564\mathrm{E}^{-10}$
11	$9.900\,000\mathrm{E}^{-01}$	$5.310\,152\mathrm{E}^{-03}$	$5.000\,000\mathrm{E}^{-03}$	$3.857\,991\mathrm{E}^{-09}$
12	$9.950\,000\mathrm{E}^{-01}$	$2.655\,129\mathrm{E}^{-03}$	$2.500\,000\mathrm{E}^{-03}$	$5.934\,724\mathrm{E}^{-09}$
13	$9.975\,000\mathrm{E}^{-01}$	$1.327\,647\mathrm{E}^{-03}$	$1.250\,000\mathrm{E}^{-03}$	$9.128\,510\mathrm{E}^{-09}$
14	$9.987\,500\mathrm{E}^{-01}$	$6.639\,509\mathrm{E}^{-04}$	$3.125\,000\mathrm{E}^{-04}$	$1.020\,555\mathrm{E}^{-09}$
16	$9.993\,749\mathrm{E}^{-01}$	$3.321\,802\mathrm{E}^{-04}$	$1.562\,500\mathrm{E}^{-04}$	$1.617\,424\mathrm{E}^{-09}$
18	$9.996\,874\mathrm{E}^{-01}$	$1.664\,148\mathrm{E}^{-04}$	$7.812\,500\mathrm{E}^{-05}$	$2.556\,941\mathrm{E}^{-09}$
20	$9.998\,437\mathrm{E}^{-01}$	$8.372\,220\mathrm{E}^{-05}$	$3.906\,250\mathrm{E}^{-05}$	$4.009\,808\mathrm{E}^{-09}$
21	$9.998\,828\mathrm{E}^{-01}$	$6.313\,973\mathrm{E}^{-05}$	$1.953\,125\mathrm{E}^{-05}$	$1.234\,137\mathrm{E}^{-09}$
24	$9.999\,414\mathrm{E}^{-01}$	$3.256\,670\mathrm{E}^{-05}$	$9.765\,625\mathrm{E}^{-06}$	$1.864\,404\mathrm{E}^{-09}$
27	$9.999\,707\mathrm{E}^{-01}$	$1.785\,111\mathrm{E}^{-05}$	$4.882\,812\mathrm{E}^{-06}$	$2.474\,586\mathrm{E}^{-09}$
30	$9.999\,853\mathrm{E}^{-01}$	$1.130\,585\mathrm{E}^{-05}$	$2.441\,406\mathrm{E}^{-06}$	$2.140\,462\mathrm{E}^{-09}$
34	$9.999\,951\mathrm{E}^{-01}$	$8.501\,539\mathrm{E}^{-06}$	$2.441\,406\mathrm{E}^{-06}$	$-1.070\,443\mathrm{E}^{-09}$
35	$9.999\,975\mathrm{E}^{-01}$	$8.295\,499\mathrm{E}^{-06}$	$2.441\,420\mathrm{E}^{-06}$	$-7.704\,466\mathrm{E}^{-09}$
36	$1.000\,000\mathrm{E}^{+00}$	$8.269\,178\mathrm{E}^{-06}$		

BSP. (16.8.7) Wir zeigen die Wirkungsweise der Schrittweitensteuerung am Beispiel der Anfangswertaufgabe

$$y'(x) = \frac{y}{\sqrt{(1-x)^2+y^2}} - 1, \quad x \in [0,1], \qquad y(0) = y_0.$$

Die rechte Seite der Differentialgleichung wird für diejenige Lösung $y(x)$ indefinit, für die $y(1) = 0$ erfüllt ist. Diese Lösung hat den Anfangswert $y(0) = y_0 \doteq 0.531\,010\,06$. Wir integrieren zunächst die Anfangswertaufgabe numerisch mit dem Verfahren (8.46) von ENGLAND unter Verwendung des Schätzers (8.48) für den lokalen Diskretisationsfehler. Wir starten das Verfahren mit dem Anfangswert $y_0 := 0.531\,010\,06$ bei einer Toleranz $\epsilon := 10^{-8}$ und einer Anfangsschrittweite $h_0 := 0.01$. In der ersten der beiden obigen Tabellen wird sichtbar, wie die Schrittweite h_j nachgesteuert werden muss, damit $|d_{j+1}| < \epsilon$ gewährleistet ist.

Wir integrieren nun die Anfangswertaufgabe mit dem modifizierten EULER–Verfahren (8.27). Zur Schätzung des lokalen Diskretisationsfehlers verwenden wir das ENGLAND–Verfahren (8.46). Es gilt hier:

$$\boxed{d_{j+1}^{(MEV)} \approx \frac{1}{6} h_j \{k_1 - 6k_2 + 4k_3 + k_4\} + \mathcal{O}(h_j^4).}$$

Die Anfangsdaten und die Toleranz geben wir wie vorher vor. Wegen des erheblich geringeren Genauigkeitsgrades gegenüber dem ENGLAND–Verfahren arbeitet das modifizierte EULER–Verfahren mit kleineren Schrittweiten, um den lokalen Diskretisationsfehler unterhalb der Toleranz zu halten. Insgesamt werden auch mehr Integrationsschritte benötigt, wie die zweite der obigen Tabellen zeigt.

Um zu einer qualitativen Bewertung der hier vorgestellten 10 RK–Verfahren zu kommen, sind nachfolgend alle Verfahren an zwei Testbeispielen mit 6 verschiedenen Schrittzahlen getestet worden. In den Tabellen ist jeweils der maximale Fehler $\max\limits_{1 \le j \le n} |y(x_j) - \eta_j|$ zwischen der exakten Lösung $y(x)$ und der RK–Lösung η_j in den äquidistanten Stützstellen $x_j = j/n$ berechnet worden.

Vergleichstabelle der RUNGE–KUTTA–Verfahren.

(Die Referenznummer bezeichnet den im Skript verwendeten Algorithmus.)

BSP. (16.8.8) Die Anfangswertaufgabe

$$y' = 2x\, e^{-x^2} \cdot y^2, \quad x \in [0,1], \qquad y(0) = 1,$$

mit der exakten Lösung $y(x) = \exp(x^2)$ wird numerisch integriert:

	1.Ordnung	2.Ordnung		3.Ordnung	
Schritt–zahl	EULER (8.2)	HEUN (8.22)	**MEV** (8.27)	HEUN (8.38)	KUTTA (8.39)
$n = 10$	$5.539\,809\text{E}^{-01}$	$5.068\,502\text{E}^{-02}$	$4.612\,707\text{E}^{-02}$	$3.246\,723\text{E}^{-03}$	$7.223\,686\text{E}^{-04}$
$n = 15$	$4.132\,285\text{E}^{-01}$	$2.344\,836\text{E}^{-02}$	$2.219\,985\text{E}^{-02}$	$1.039\,163\text{E}^{-03}$	$2.406\,253\text{E}^{-04}$
$n = 20$	$3.301\,576\text{E}^{-01}$	$1.342\,453\text{E}^{-02}$	$1.297\,932\text{E}^{-02}$	$4.553\,958\text{E}^{-04}$	$1.081\,749\text{E}^{-04}$
$n = 30$	$2.358\,704\text{E}^{-01}$	$6.060\,872\text{E}^{-03}$	$5.988\,834\text{E}^{-03}$	$1.401\,163\text{E}^{-04}$	$3.427\,442\text{E}^{-05}$
$n = 60$	$1.272\,964\text{E}^{-01}$	$1.535\,763\text{E}^{-03}$	$1.552\,031\text{E}^{-03}$	$1.818\,217\text{E}^{-05}$	$4.597\,997\text{E}^{-06}$
$n = 80$	$9.744\,139\text{E}^{-02}$	$8.664\,479\text{E}^{-04}$	$8.806\,566\text{E}^{-04}$	$7.742\,561\text{E}^{-06}$	$1.975\,666\text{E}^{-06}$

	3.Ordnung	4.Ordnung			5.Ordnung
Schritt–		**klass. RK**	3/8–**RK**	ENGLAND	ENGLAND
zahl	(8.41)	(8.44)	(8.45)	(8.46)	(8.47)
$n = 10$	$5.600\,013\mathrm{E}^{-03}$	$7.048\,056\mathrm{E}^{-05}$	$2.076\,513\mathrm{E}^{-05}$	$1.241\,347\mathrm{E}^{-04}$	$2.483\,311\mathrm{E}^{-05}$
$n = 15$	$1.763\,064\mathrm{E}^{-03}$	$1.431\,217\mathrm{E}^{-05}$	$3.134\,158\mathrm{E}^{-06}$	$2.599\,724\mathrm{E}^{-05}$	$3.578\,097\mathrm{E}^{-06}$
$n = 20$	$7.659\,794\mathrm{E}^{-04}$	$4.579\,684\mathrm{E}^{-06}$	$9.857\,867\mathrm{E}^{-07}$	$8.461\,174\mathrm{E}^{-06}$	$8.880\,269\mathrm{E}^{-07}$
$n = 30$	$2.335\,854\mathrm{E}^{-04}$	$9.131\,690\mathrm{E}^{-07}$	$2.140\,714\mathrm{E}^{-07}$	$1.717\,919\mathrm{E}^{-06}$	$1.223\,961\mathrm{E}^{-07}$
$n = 60$	$3.003\,422\mathrm{E}^{-05}$	$5.769\,106\mathrm{E}^{-08}$	$1.473\,381\mathrm{E}^{-08}$	$1.105\,472\mathrm{E}^{-07}$	$4.267\,349\mathrm{E}^{-09}$
$n = 80$	$1.275\,961\mathrm{E}^{-05}$	$1.854\,641\mathrm{E}^{-08}$	$4.676\,621\mathrm{E}^{-09}$	$3.544\,118\mathrm{E}^{-08}$	$1.280\,568\mathrm{E}^{-09}$

BSP. (16.8.9) Die Anfangswertaufgabe

$$y' = -\frac{\sin 2x}{2y}, \quad x \in [0,1], \qquad y(0) = 1,$$

mit der exakten Lösung $y(x) = \cos x$ wird numerisch integriert:

	1.Ordnung	2.Ordnung		3.Ordnung	
Schritt–	EULER	HEUN	**MEV**	HEUN	KUTTA
zahl	(8.2)	(8.22)	(8.27)	(8.38)	(8.39)
$n = 10$	$6.071\,371\mathrm{E}^{-02}$	$2.339\,002\mathrm{E}^{-03}$	$5.416\,909\mathrm{E}^{-04}$	$3.190\,141\mathrm{E}^{-06}$	$5.952\,776\mathrm{E}^{-05}$
$n = 15$	$4.170\,394\mathrm{E}^{-02}$	$1.018\,092\mathrm{E}^{-03}$	$2.417\,934\mathrm{E}^{-04}$	$9.419\,327\mathrm{E}^{-07}$	$1.727\,422\mathrm{E}^{-05}$
$n = 20$	$3.179\,605\mathrm{E}^{-02}$	$5.663\,087\mathrm{E}^{-04}$	$1.362\,235\mathrm{E}^{-04}$	$3.960\,249\mathrm{E}^{-07}$	$7.214\,817\mathrm{E}^{-06}$
$n = 30$	$2.157\,317\mathrm{E}^{-02}$	$2.487\,721\mathrm{E}^{-04}$	$6.061\,396\mathrm{E}^{-05}$	$1.168\,209\mathrm{E}^{-07}$	$2.116\,898\mathrm{E}^{-06}$
$n = 60$	$1.099\,201\mathrm{E}^{-02}$	$6.144\,205\mathrm{E}^{-05}$	$1.516\,433\mathrm{E}^{-05}$	$1.455\,919\mathrm{E}^{-08}$	$2.620\,563\mathrm{E}^{-07}$
$n = 80$	$8.284\,827\mathrm{E}^{-03}$	$3.445\,386\mathrm{E}^{-05}$	$8.530\,873\mathrm{E}^{-06}$	$6.168\,193\mathrm{E}^{-09}$	$1.102\,489\mathrm{E}^{-07}$

	3.Ordnung	4.Ordnung			5.Ordnung
Schritt–		**klass. RK**	3/8–**RK**	ENGLAND	ENGLAND
zahl	(8.41)	(8.44)	(8.45)	(8.46)	(8.47)
$n = 10$	$1.390\,401\mathrm{E}^{-04}$	$1.715\,017\mathrm{E}^{-06}$	$1.108\,122\mathrm{E}^{-06}$	$1.031\,897\mathrm{E}^{-06}$	$4.301\,364\mathrm{E}^{-08}$
$n = 15$	$4.136\,776\mathrm{E}^{-05}$	$3.395\,334\mathrm{E}^{-07}$	$2.165\,525\mathrm{E}^{-07}$	$2.046\,590\mathrm{E}^{-07}$	$5.629\,772\mathrm{E}^{-09}$
$n = 20$	$1.747\,878\mathrm{E}^{-05}$	$1.075\,641\mathrm{E}^{-07}$	$6.813\,297\mathrm{E}^{-08}$	$6.489\,244\mathrm{E}^{-08}$	$1.321\,495\mathrm{E}^{-09}$
$n = 30$	$5.185\,116\mathrm{E}^{-06}$	$2.125\,307\mathrm{E}^{-08}$	$1.336\,320\mathrm{E}^{-08}$	$1.283\,115\mathrm{E}^{-08}$	$1.537\,046\mathrm{E}^{-10}$
$n = 60$	$6.487\,507\mathrm{E}^{-07}$	$1.282\,387\mathrm{E}^{-09}$	$7.867\,129\mathrm{E}^{-10}$	$7.630\,660\mathrm{E}^{-10}$	$3.456\,079\mathrm{E}^{-11}$
$n = 80$	$2.737\,779\mathrm{E}^{-07}$	$3.628\,883\mathrm{E}^{-10}$	$2.055\,458\mathrm{E}^{-10}$	$1.973\,603\mathrm{E}^{-10}$	$5.911\,715\mathrm{E}^{-11}$

Man erkennt, dass bei Verfahren derselben Ordnung problemspezifische Qualitätsunterschiede auftreten (in BSP. (16.8.8) ist die Lösungsfunktion $y(x)$ konvex, im BSP(16.8.9) ist sie konkav). Deshalb kann generell bei Verfahren derselben Ordnung kein qualitativer Unterschied festgestellt werden. Geht man jedoch davon aus, dass die **Rechenzeit** für ein Verfahren etwa proportional zur Anzahl der Funktionsauswertungen steigt, so ist bei dem sechsstufigen Verfahren (8.47) der Aufwand bei einer Schrittzahl $n = 10$ vergleichbar mit dem Aufwand für das einstufige EULER–Verfahren bei einer Schrittzahl $n = 60$. Die folgende Tabelle stellt bei einem Aufwand von 60 Funktionsauswertungen pro Verfahren die Fehlerbilanz verschiedener Verfahren zusammen:

BSP. (16.8.8)	EULER (8.2)	HEUN (8.22)	**MEV** (8.27)	HEUN (8.38)	KUTTA (8.39)	**RK** (8.44)	3/8–**RK** (8.45)	ENGLAND (8.47)
n	60	30	30	20	20	15	15	10
Fehler	$1.27\mathrm{E}^{-1}$	$6.06\mathrm{E}^{-3}$	$5.99\mathrm{E}^{-3}$	$4.55\mathrm{E}^{-4}$	$1.08\mathrm{E}^{-4}$	$1.43\mathrm{E}^{-5}$	$3.13\mathrm{E}^{-6}$	$2.48\mathrm{E}^{-5}$

Man erkennt, dass bei gleichem Rechenaufwand die Verfahren höherer Ordnung eindeutig zu besseren Resultaten führen.

Aus der Abschätzung (8.14) des globalen Diskretisationsfehlers g_n geht hervor, dass der Fehler sich bei einer Schrittzahlverdoppelung in einem Verfahren p–ter Ordnung um den Faktor 2^{-p} verkleinern sollte. Man kann diesen Sachverhalt an den Tabellen auf Seite 246 nachvollziehen. In BSP. (16.8.8) erhält man für das klassische RK–Verfahren (8.44) die folgende Tabelle:

Schrittzahl	max $\lvert g_n\rvert$	Quotient bei Schrittzahlhalb.
$n = 10$	$7.048 \cdot 10^{-5}$	$\searrow$ 15.4
$n = 20$	$4.580 \cdot 10^{-6}$	$\nearrow$ $\searrow$ 15.8
$n = 40$	$2.902 \cdot 10^{-7}$	$\nearrow$ $\searrow$ 15.6
$n = 80$	$1.855 \cdot 10^{-8}$	$\nearrow$

$\sim 16 = 2^4$.

Das Ergebnis entspricht der Tatsache, dass das klassische RK–Verfahren die Ordnung 4 hat.

16.9 Potenzreihenansätze

Eine Methode, die im weitesten Sinn zu den numerischen Verfahren zur Lösung der Anfangswertaufgabe

$$\boxed{y' = f(x, y), \qquad y(x_0) = y_0,} \tag{9.1}$$

gezählt werden kann, ist die Methode des **Potenzreihenansatzes**. Ein solcher Ansatz führt häufig auf einfach zu handhabende Rekursionsformeln. Allerdings müssen sehr starke Voraussetzungen gestellt werden: Die Funktion $f(x, y)$ muss in der Umgebung $U(x_0, y_0)$ eines Punktes (x_0, y_0) **reell analytisch** sein. Das heißt, $f(x, y)$ gestattet eine konvergente Potenzreihenentwicklung

$$f(x, y) = \sum_{j,k=0}^{\infty} a_{jk}(x - x_0)^j (y - y_0)^k, \qquad (x, y) \in U(x_0, y_0). \tag{9.2}$$

Es sei hier ohne Beweis festgestellt, dass in diesem Fall die Lösung der AWA (9.1) in einer Umgebung $V(x_0)$ des Punktes x_0 ebenfalls in eine Potenzreihe entwickelbar ist:

$$y(x) = \sum_{i=0}^{\infty} c_i (x - x_0)^i, \qquad x \in V(x_0). \tag{9.3}$$

Es ist zumindest plausibel, dass die gemäß dem Taylorschen Satz definierten Koeffizienten $c_i = \frac{1}{i!}\, y^{(i)}(x_0)$ durch fortgesetzte Differentiation der Gleichung $y' = f(x, y)$ berechnet werden können. Die Anfangsbedingung $y(x_0) = y_0$ führt schon unmittelbar auf $c_0 = y_0$; ferner ergeben sich

$$\begin{aligned} y'(x_0) &= f(x_0, y_0) && \Rightarrow \quad c_1 = \frac{1}{1!}\, y'(x_0), \\ y''(x_0) &= (f_x + f_y \cdot f)(x_0, y_0) && \Rightarrow \quad c_2 = \frac{1}{2!}\, y''(x_0), \end{aligned}$$

usw. Eine andere Möglichkeit der Bestimmung von $c_1, c_2, \ldots$ ergibt sich durch Einsetzen von (9.3) in die Gleichungen (9.1) und (9.2) und anschließendem Koeffizientenvergleich. Man gewinnt allgemeine Rekursionsformeln, die hier aber wegen ihres geringen praktischen Nutzens nicht angegeben werden sollen. Es ist vielmehr sinnvoller, sich jeweils am speziellen Beispiel zu orientieren.

BSP. (16.9.1) Es soll die Lösung der Anfangswertaufgabe

$$y' = x^2 + y^2, \qquad y(0) = 1,$$

in der Umgebung des Anfangspunktes $x_0 = 0$ durch eine Potenzreihenentwicklung berechnet werden.

Lösung: (a) Wir wenden die Methode des sukzessiven Differenzierens an. Ausgehend vom Startwert $y_0 = c_0 = 1$ berechnet man:

$$\begin{array}{lcll} y'(0) & = & 0+1 & = \ 1, \\ y''(0) & = & (2x + 2y \cdot y')(0,1) & = \ 2, \\ y'''(0) & = & (2 + 2y'^2 + 2y \cdot y'')(0,1) & = \ 8, \\ y^{(4)}(0) & = & (6y'y'' + 2y \cdot y''')(0,1) & = \ 28, \end{array}$$

usw. Somit gilt die folgende Approximation der gesuchten Lösung durch ihr TAYLOR–Polynom vom Grade 4:

$$y(x) = 1 + x + \frac{2}{2!}\,x^2 + \frac{8}{3!}\,x^3 + \frac{28}{4!}\,x^4 + \cdots$$

(b) Wir setzen für die gesuchte Lösung $y(x)$ eine Potenzreihe mit unbestimmten Koeffizienten an:

$$y(x) = \sum_{i=0}^{\infty} c_i x^i, \qquad y'(x) = \sum_{i=0}^{\infty} i c_i x^{i-1} = \sum_{i=0}^{\infty} (i+1) c_{i+1} x^i.$$

Indem wir diese Ausdrücke in die gegebene DGl einsetzen, erhalten wir:

$$y'(x) = \sum_{i=0}^{\infty} (i+1) c_{i+1} x^i \stackrel{!}{=} x^2 + \Big(\sum_{i=0}^{\infty} c_i x^i\Big)^2 \stackrel{\text{Cauchy-Prod.}}{=} x^2 + \sum_{i=0}^{\infty} x^i \sum_{k=0}^{i} c_k c_{i-k}.$$

Ein Koeffizientenvergleich liefert nun, beginnend mit dem Koeffizienten $c_0 = 1$:

$$\begin{array}{lcll} [x^0] \ : \ 1 \cdot c_1 & = & c_0^2 & \Rightarrow \ c_1 = 1, \\ [x^1] \ : \ 2 \cdot c_2 & = & c_0 c_1 + c_1 c_0 & \Rightarrow \ c_2 = 1, \\ [x^2] \ : \ 3 \cdot c_3 & = & 1 + c_0 c_2 + c_1 c_1 + c_2 c_0 & \Rightarrow \ c_3 = \frac{4}{3}, \\ [x^3] \ : \ 4 \cdot c_4 & = & c_0 c_3 + c_1 c_2 + c_2 c_1 + c_3 c_0 & \Rightarrow \ c_4 = \frac{7}{6}, \end{array}$$

usw. Man erhält allgemein die **Rekursionsformel**

$$c_{i+1} = \frac{1}{i+1} \sum_{k=0}^{i} c_k c_{i-k}, \qquad i \geq 3,$$

die sich algorithmisch sehr einfach mit Hilfe von Computern bis zu jeder endlichen Ordnung i lösen lässt.

Die Methode des Potenzreihenansatzes kann auch auf Differentialgleichungen höherer Ordnung angewendet werden; man verwendet sie insbesondere bei linearen DGln der Ordnung n mit analytischen Koeffizienten, wenn sonst kein anderes Verfahren zur Bestimmung eines Fundamentalsystems greift. Wir beschränken uns hier ausschließlich auf lineare DGln 2.Ordnung.

Satz 16.30 *Sind in der DGl*

$$\boxed{y'' + p(x)y' + q(x)y = 0} \tag{9.4}$$

die Koeffizientenfunktionen $p, q \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R})$ an der Stelle $x_0 \in \mathbf{R}$ in konvergente Potenzreihen entwickelbar, so kann die allgemeine Lösung $y(x)$ der DGl (9.4) ebenfalls mittels einer konvergenten Potenzreihenentwicklung um den Punkt x_0 berechnet werden. Die Koeffizienten dieser Potenzreihenentwicklung bestimmt man nach der Methode des Koeffizientenvergleichs.

Begründung: (Skizze.) Wir können ohne Einschränkung $x_0 = 0$ annehmen. Es seien

$$p(x) = \sum_{n=0}^{\infty} p_n x^n, \quad q(x) = \sum_{n=0}^{\infty} q_n x^n, \qquad x \in B_\rho(0) \text{ für ein } \rho > 0,$$

die vorausgesetzten Potenzreihenentwicklungen. Wir setzen für $y(x)$ eine Potenzreihe an der Stelle $x_0 = 0$ mit unbestimmten Koeffizienten a_n an:

$$\left.\begin{array}{lcl|l} y(x) & = & \displaystyle\sum_{n=0}^{\infty} a_n x^n & \cdot \displaystyle\sum_{n=0}^{\infty} q_n x^n \\ y'(x) & = & \displaystyle\sum_{n=0}^{\infty} n a_n x^{n-1} = \sum_{n=0}^{\infty} (n+1) a_{n+1} x^n & \cdot \displaystyle\sum_{n=0}^{\infty} p_n x^n \\ y''(x) & = & \displaystyle\sum_{n=0}^{\infty} n(n-1) a_n x^{n-2} = \sum_{n=0}^{\infty} (n+2)(n+1) a_{n+2} x^n & \cdot 1 \end{array}\right\} (+)$$

Wird die obige Summe gebildet, so resultiert unter Verwendung des CAUCHY–Produkts nach geeigneter Zusammenfassung gleicher x–Potenzen:

$$\sum_{n=0}^{\infty} x^n \Big((n+2)(n+1) a_{n+2} + \sum_{j=0}^{n} (j+1) a_{j+1} p_{n-j} + \sum_{j=0}^{n} a_j q_{n-j} \Big) = 0.$$

Durch Koeffizientenvergleich ergibt sich daraus die folgende Rekursionsformel

$$a_{n+2} = -\frac{1}{(n+1)(n+2)} \sum_{j=0}^{n} \left((j+1) a_{j+1} p_{n-j} + a_j q_{n-j} \right), \quad n = 0, 1, 2, \ldots$$

Über die beiden Koeffizienten a_0 und a_1 kann frei verfügt werden. Will man zwei Lösungen

$$y_{1,2}(x) = \sum_{n=0}^{\infty} a_n^{(1,2)} x^n$$

so bestimmen, dass die Anfangsbedingungen $y_1(0) = 1$, $y_1'(0) = 0$ bzw. $y_2(0) = 0$, $y_2'(0) = 1$ angenommen werden, so sind a_0, a_1 in der folgenden Weise festzulegen:

$$a_0^{(1)} := 1,\ a_1^{(1)} := 0, \qquad a_0^{(2)} := 0,\ a_1^{(2)} := 1.$$

Wird es als bereits erwiesen angenommen, dass die zugeordneten Potenzreihen für $y_{1,2}(x)$ einen positiven Konvergenzradius haben – auf einen Konvergenzbeweis wird hier verzichtet –, so folgt aus Satz 10.3, dass die WRONSKI–Determinante $W(x) := \begin{vmatrix} y_1(x) & y_2(x) \\ y_1'(x) & y_2'(x) \end{vmatrix}$ die DGl $W'(x) = -p(x)W(x)$ löst, und somit wegen

$$W(x) = W(0) e^{-\int_0^x p(t)\,dt} = 1 \cdot e^{-\int_0^x p(t)\,dt} \neq 0$$

nicht verschwindet. Also bilden die beiden Lösungen $y_1(x)$ und $y_2(x)$ ein Fundamentalsystem für die DGl (9.4). □

BSP. (16.9.2) Zu bestimmen ist mit der Methode des Potenzreihenansatzes die Lösung der Anfangswertaufgabe

$$y'' + xy = x^3 + 2, \qquad y(0) = 0, \quad y'(0) = 1.$$

Hier sind die Voraussetzungen des Satzes 16.30 im Entwicklungspunkt $x_0 = 0$ erfüllt, und deshalb darf die gesuchte Lösung $y(x)$ als Potenzreihe im Entwicklungspunkt x_0 angesetzt werden:

$$\left.\begin{array}{lcll} y(x) & = & \displaystyle\sum_{k=0}^{\infty} c_k x^k & \cdot x \\ y'(x) & = & \displaystyle\sum_{k=0}^{\infty} k c_k x^{k-1} = \sum_{k=0}^{\infty} (k+1) c_{k+1} x^k & \cdot 0 \\ y''(x) & = & \displaystyle\sum_{k=0}^{\infty} k(k-1) c_k x^{k-2} = \sum_{k=0}^{\infty} (k+2)(k+1) c_{k+2} x^k & \cdot 1 \end{array}\right\} (+)$$

Die obige Summenbildung führt nach Einsetzen in die DGl auf die Gleichung

$$\sum_{k=0}^{\infty} c_k x^{k+1} + \sum_{k=0}^{\infty} (k+2)(k+1) c_{k+2} x^k = 2c_2 + \sum_{k=1}^{\infty} x^k (c_{k-1} + (k+2)(k+1) c_{k+2}) \stackrel{!}{=} x^3 + 2.$$

Die Anfangsbedingungen erfordern $c_0 = 0$ und $c_1 = 1$, und nach der Methode des Koeffizientenvergleichs berechnet man

$$\begin{array}{lllllllll} [x^0] & : & 2c_2 & = & 2 & \Rightarrow & c_2 & = & 1, \\ [x^1] & : & c_0 + 3 \cdot 2c_3 & = & 0 & \Rightarrow & c_3 & = & 0, \\ [x^2] & : & c_1 + 4 \cdot 3c_4 & = & 0 & \Rightarrow & c_4 & = & -\frac{1}{3 \cdot 4}, \\ [x^3] & : & c_2 + 5 \cdot 4c_5 & = & 1 & \Rightarrow & c_5 & = & 0, \end{array}$$

usw. Allgemein bestimmt man nun c_k, $k \geq 6$, aus der Rekursionsformel

$$\boxed{c_{k+3} = -\frac{c_k}{(k+3)(k+2)}, \quad k \geq 3,}$$

aus der wegen $c_3 = 0 = c_5$ sofort $c_{3n} = 0 = c_{3n+2}$ für alle $n \in \mathbf{N}$ folgen. Setzen wir noch $k = 3n - 2$ in die Rekursionsformel ein, so nimmt diese die Form

$$c_{3n+1} = -\frac{c_{3n-2}}{3n(3n+1)}, \quad n \in \mathbf{N},$$

an, und mit der Induktionsverankerung $c_1 = 1$ kann daraus zum Beispiel durch vollständige Induktion das folgende Bildungsgesetz bewiesen werden:

$$c_{3n+1} = \frac{(-1)^n}{(3 \cdot 4)(6 \cdot 7)(9 \cdot 10) \cdots 3n(3n+1)}, \quad n \in \mathbf{N}.$$

Somit liegt die gesuchte Lösung in der folgenden Form vor:

$$y(x) = x + x^2 + \sum_{n=1}^{\infty} c_{3n+1} x^{3n+1}.$$

Wir diskutieren den **Konvergenzbereich** der Potenzreihe mit Hilfe des Quotientenkriteriums und setzen dazu $a_n := c_{3n+1}$ sowie $z := x^3$. Dann gilt $y(x) = x + x^2 + x \sum\limits_{n=1}^{\infty} a_n z^n$, und die Potenzreihe ist für alle $z \in \mathbf{R}$ konvergent mit

$$|z| < \lim_{n \to \infty} \left| \frac{a_n}{a_{n+1}} \right| = \lim_{n \to \infty} (3n+3)(3n+4) = \infty.$$

Die Lösung $y(x)$ wird also durch eine beständig konvergente Potenzreihe dargestellt.

Die bisherige Diskussion der Potenzreihenansätze für die DGl (9.4) ging von der Hypothese aus, dass die Koeffizientenfunktionen $p(x)$ und $q(x)$ in einem Punkt $x_0 \in \mathbf{R}$ jeweils konvergente Potenzreihenentwicklungen zulassen. In diesen Fällen heiße der Punkt x_0 ein **regulärer Punkt** der DGl (9.4).

Definition 16.21 *Für eine lineare Differentialgleichung 2.Ordnung in der Form*

$$\boxed{y'' + \frac{P(x)}{x - x_0}\, y' + \frac{Q(x)}{(x - x_0)^2} y = 0} \tag{9.5}$$

mit

$$P(x) = \sum_{n=0}^{\infty} p_n (x - x_0)^n, \quad Q(x) = \sum_{n=0}^{\infty} q_n (x - x_0)^n \quad \text{und} \quad p_0^2 + q_0^2 + q_1^2 \neq 0$$

heiße der Punkt x_0 eine **Stelle der Bestimmtheit** *oder eine* **außerwesentliche Singularität**. *Die DGl (9.5) selbst heiße* **schwach singulär**.

Eine außerwesentliche Singularität x_0 tritt in der DGl (9.4) lax gesprochen dann auf, wenn die Funktion $p(x)$ in x_0 einen Pol höchstens 1.Ordnung und die Funktion $q(x)$ in x_0 einen Pol höchstens 2.Ordnung besitzt. Treten Pole höherer Ordnung auf, so heißt x_0 eine **wesentliche Singularität**.

Bei Differentialgleichungen mit einer außerwesentlichen Singularität führt im allgemeinen ein **verallgemeinerter Potenzreihenansatz** zur Bestimmung eines Fundamentalsystems:

Satz 16.31 *In der Umgebung einer außerwesentlichen Singularität x_0 der DGl (9.5) mit*

$$P(x) = \sum_{n=0}^{\infty} p_n (x - x_0)^n, \quad Q(x) = \sum_{n=0}^{\infty} q_n (x - x_0)^n \quad \text{und} \quad p_0^2 + q_0^2 + q_1^2 \neq 0,$$

kann die allgemeine Lösung $y(x)$ durch einen **verallgemeinerten Potenzreihenansatz**

$$\boxed{y(x) = (x - x_0)^{\rho} \sum_{n=0}^{\infty} c_n (x - x_0)^n} \tag{9.6}$$

gewonnen werden. Darin ist ρ als Wurzel der **determinierenden Gleichung** *oder* **Indexgleichung**

$$\boxed{\rho(\rho - 1) + p_0 \rho + q_0 = 0} \tag{9.7}$$

bestimmt, und der Koeffizient $c_0 \neq 0$ ist frei wählbar. Gilt für die Wurzeln ρ_1, ρ_2 der quadratischen Gleichung (9.7) die Beziehung $\rho_1 - \rho_2 \notin \mathbf{Z}$, so liefert die Methode des verallgemeinerten Potenzreihenansatzes für ρ_1 und ρ_2 zwei linear unabhängige Lösungen $y_1(x)$ und $y_2(x)$.

Für $\rho_1 - \rho_2 \in \mathbf{Z}$ muss mit dem folgenden Ansatz eine zweite linear unabhängige Lösung berechnet werden, wobei ohne Beschränkung der Allgemeinheit $\rho_1 \geq \rho_2$ angenommen wird:

$$\boxed{y_2(x) = (x - x_0)^{\rho_2} \sum_{n=0}^{\infty} b_n (x - x_0)^n + c y_1(x) \cdot \ln(x - x_0).} \tag{9.8}$$

Hier kann auch der Fall $c = 0$ eintreten.

Begründung: (Skizze) Wir lassen hier Konvergenzuntersuchungen wieder weg. Ferner nehmen wir ohne Einschränkung $x_0 = 0$ an. Aus dem Ansatz (9.6) resultiert zunächst

$$\left.\begin{array}{lcl|l} y(x) & = & \sum\limits_{n=0}^{\infty} c_n x^{n+\rho} & \cdot\, x^{-2} \sum\limits_{n=0}^{\infty} q_n x^n \\ y'(x) & = & \sum\limits_{n=0}^{\infty} (n+\rho) c_n x^{n+\rho-1} & \cdot\, x^{-1} \sum\limits_{n=0}^{\infty} p_n x^n \\ y''(x) & = & \sum\limits_{n=0}^{\infty} (n+\rho)(n+\rho-1) c_n x^{n+\rho-2} & \cdot\, 1 \end{array}\right\} \quad (+)$$

Bilden wir die obige Summe, so ergibt sich nach Einsetzen in die DGl (9.5):

$$\sum_{n=0}^{\infty} x^{n+\rho-2} \Big((n+\rho)(n+\rho-1) c_n + \sum_{k=0}^{n} (k+\rho) c_k p_{n-k} + \sum_{k=0}^{n} c_k q_{n-k} \Big) = 0.$$

Ein Koeffizientenvergleich führt nun für $n = 0, 1, 2, \ldots$ auf die Rekursionsformel

$$[(n+\rho)(n+\rho-1) + (n+\rho)p_0 + q_0] c_n + \sum_{k=0}^{n-1} c_k ((k+\rho) p_{n-k} + q_{n-k}) = 0.$$

Für $n = 0$ erhält man die Indexgleichung (9.7). Für $n > 0$ bestimmt die Rekursionsformel den Koeffizienten c_n aus den schon berechneten Koeffizienten $c_1, c_2, \ldots, c_{n-1}$ und dem frei wählbaren Koeffizienten c_0, vorausgesetzt, der Faktor in den Klammern $[\cdots]$ vor c_n verschwindet nicht. Es gilt aber unter Berücksichtigung der Indexgleichung (9.7):

$$[\cdots] = (n+\rho)(n+\rho-1) + (n+\rho)p_0 + q_0 = n(n+\rho-1) + n\rho + np_0 = n(n + 2\rho - \rho_1 - \rho_2),$$

wobei wegen der VIËTAschen Wurzelsätze die Identität $\rho_1 + \rho_2 = -(p_0 - 1)$ gilt. Wir haben nun für $\rho = \rho_1$ die Beziehung $2\rho - \rho_1 - \rho_2 = \rho_1 - \rho_2$ und für $\rho = \rho_2$ ganz analog $2\rho - \rho_1 - \rho_2 = \rho_2 - \rho_1$. Das heißt, im Falle von $\rho_1 - \rho_2 = \pm n$ gilt $n + 2\rho - \rho_1 - \rho_2 = 0$ entweder für ρ_1 oder für ρ_2. Der Ansatz (9.6) liefert nur eine Lösung, und zwar für denjenigen Exponenten ρ_j, für den $n + 2\rho_j - \rho_1 - \rho_2 \neq 0$ gilt. Das ist der in Satz 16.31 angegebene Fall. Der dann notwendige Ansatz (9.8) kann am speziellen Beispiel verifiziert werden. □

BSP. (16.9.3) Die BESSELsche **Differentialgleichung.** Das ist die folgende lineare DGl 2.Ordnung mit einer außerwesentlichen Singularität bei $x_0 = 0$:

$$\boxed{y'' + \frac{1}{x} y' + \frac{x^2 - p^2}{x^2} y = 0, \qquad p = const \in \mathbf{R}.} \tag{9.9}$$

Diese DGl tritt bei einigen kreissymmetrischen Problemen der Potentialtheorie und der Elastomechanik auf, zum Beispiel bei Schwingungen von ebenen elastischen Vollkreismembranen; man vgl. BSP. (16.6.2).

Wegen der außerwesentlichen Singularität im Punkt $x_0 = 0$ werden wir ein Fundamentalsystem für die BESSELsche DGl mit Hilfe eines verallgemeinerten Potenzreihenansatzes bestimmen:

$$\left.\begin{array}{lcl|l} y(x) & = & \sum\limits_{n=0}^{\infty} c_n x^{n+\rho} & \cdot\, x^{-2}(x^2 - p^2) \\ y'(x) & = & \sum\limits_{n=0}^{\infty} (n+\rho) c_n x^{n+\rho-1} & \cdot\, x^{-1} \\ y''(x) & = & \sum\limits_{n=0}^{\infty} (n+\rho)(n+\rho-1) c_n x^{n+\rho-2} & \cdot\, 1 \end{array}\right\} \quad (+)$$

Bilden wir die obige Summe, so ergibt sich nach Einsetzen in die DGl (9.9):

$$\sum_{n=0}^{\infty} c_n((n+\rho)(n+\rho-1+1)-p^2)x^{n+\rho-2} + \sum_{n=2}^{\infty} c_{n-2}x^{n+\rho-2} = 0.$$

Wir führen die Indexfunktion $f(\rho) := \rho^2 - p^2$ ein. Mit dieser Notation liefert ein Koeffizientenvergleich:

$$c_0 f(\rho) = 0, \qquad c_1 f(1+\rho) = 0, \qquad c_n f(n+\rho) + c_{n-2} = 0 \ \ \forall\, n \geq 2. \tag{9.10}$$

Unter der Annahme $c_0 \neq 0$ folgt die Indexgleichung $f(\rho) = 0$ mit den beiden Wurzeln

$$\rho_{1,2} = \pm p.$$

Wir treffen nun Fallunterscheidungen gemäß den Vorgaben des Satzes 16.31.

Fall (A): Es sei $\rho_1 - \rho_2 = 2p \notin \mathbf{Z}$. Das heißt, p darf kein ganzzahliges Vielfaches von $\frac{1}{2}$ sein. In diesem Fall gilt $f(n+\rho) = f(n \pm p) = n(n \pm 2p) \neq 0$. Die Koeffizienten c_n können daher eindeutig aus den Rekursionsformeln (9.10) berechnet werden:

$$\boxed{c_{2m} = \begin{cases} \dfrac{(-1)^m c_0}{2^{2m}(1+p)(2+p)\cdots(m+p)m!} & : \ \rho_1 = +p, \\ \dfrac{(-1)^m c_0}{2^{2m}(1-p)(2-p)\cdots(m-p)m!} & : \ \rho_2 = -p, \end{cases} \qquad c_{2m-1} = 0, \quad m \in \mathbf{N}.} \tag{9.11}$$

Diese Koeffizienten können mit Hilfe der Euler**schen Gamma–Funktion** Γ auf eine einfachere Form gebracht werden. Die Funktion Γ wird in der Regel über eine *definierende Funktionalgleichung* eingeführt, nämlich:

$$\boxed{\Gamma(1+x) = x\Gamma(x), \quad x > 0, \qquad \Gamma(1) = 1.}$$

Durch sukzessives Einsetzen von $x = 1, 2, \ldots, n$ bestätigt man leicht den folgenden Zusammenhang mit den Fakultäten: $\Gamma(1+n) = n! \ \forall\, n \in \mathbf{N}_0$. Für negative $x \in \mathbf{R}$ wird die Gamma–Funktion durch die funktionale Beziehung

$$\boxed{\Gamma(x)\Gamma(1-x) = \frac{\pi}{\sin \pi x}, \qquad x \notin \mathbf{Z},} \tag{9.12}$$

erklärt. In den Punkten $x = 0, -1, -2, \ldots$ ist $\Gamma(x)$ nicht erklärt; dort hat die Funktion Polstellen. Um den Zusammenhang mit den Koeffizienten (9.11) herzustellen, setzen wir sukzessive $x = (1 \pm p),\ (2 \pm p), \ldots, (m \pm p)$ in die Funktionalgleichung ein mit dem Ergebnis:

$$\Gamma(1+m \pm p) = (1 \pm p)(2 \pm p)\cdots(m \pm p)\Gamma(1 \pm p).$$

Daraus gewinnen wir die folgende Darstellung der nichtverschwindenden Koeffizienten:

$$c_{2m} = \frac{(-1)^m}{2^{2m}\Gamma(1+m \pm p)\, m!}\, c_0 \Gamma(1 \pm p), \quad m \in \mathbf{N}, \quad \rho_{1,2} = \pm p.$$

Da über den Koeffizienten c_0 noch frei verfügt werden kann, wird aus Normierungsgründen die folgende Wertzuweisung vorgenommen:

$$c_0 := \frac{1}{2^p \Gamma(1+p)} \text{ für } \rho_1 = +p, \quad c_0 := \frac{2^p}{\Gamma(1-p)} \text{ für } \rho_2 = -p.$$

Es resultiert nun ein Fundamentalsystem für die Besselsche DGl (9.9) in der Form

$$y_1(x) \ = \ J_p(x) := \Big(\frac{x}{2}\Big)^p \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\Gamma(1+m+p)} \Big(\frac{x}{2}\Big)^{2m},$$

$$y_2(x) \ = \ J_{-p}(x) := \Big(\frac{x}{2}\Big)^{-p} \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\Gamma(1+m-p)} \Big(\frac{x}{2}\Big)^{2m}.$$

Definition 16.22 *Die Funktion*

$$\boxed{J_p(x) := \Big(\frac{x}{2}\Big)^p \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\Gamma(1+m+p)} \Big(\frac{x}{2}\Big)^{2m}, \quad x \in \mathbf{R},} \tag{9.13}$$

heiße Bessel*–Funktion erster Art der Ordnung p. Die Funktion $J_p(x)$ ist für alle $p \in \mathbf{R}$ mit $-p \notin \mathbf{N}$ erklärt. Die Potenzreihe in* (9.13) *konvergiert beständig.*

Da die Bessel–DGl (9.9) ein Sonderfall der allgemeinen DGl (9.4) mit den Spezifikationen $p(x) := \frac{1}{x}$ und $q(x) := \frac{x^2-p^2}{x^2}$ ist, haben die beiden Funktionen $J_p(x)$ und $J_{-p}(x)$ die folgende Wronski–Determinante:

$$W(x) = W(x_0)e^{-\int_{x_0}^{x} \frac{dt}{t}} = \frac{x_0 W(x_0)}{x} \overset{x_0 \to 0+}{=} -\frac{2}{x\Gamma(p)\Gamma(1-p)} \overset{(9.12)}{=} -\frac{2}{\pi x} \sin \pi p.$$

Das heißt, die Wronski–Determinante verschwindet nur für $p \in \mathbf{Z}$, obwohl wir in unseren Herleitungen $2p \notin \mathbf{Z}$ annehmen mussten. Die Bessel–Funktionen $J_p(x)$ und $J_{-p}(x)$ liefern auch dann noch ein Fundamentalsystem, wenn p ein ungerades Vielfaches von $\frac{1}{2}$ ist:

Satz 16.32 *Die* Bessel*–Funktionen $J_p(x)$ und $J_{-p}(x)$ bilden für jeden Index $p \notin \mathbf{Z}$ ein Fundamentalsystem für die* Besselsche *DGl* (9.9).

Fall (B): In vielen Anwendungen wird man auf den Fall $p = 0$ der Besselschen DGl (9.9) geführt. Die Indexgleichung $f(\rho) = 0$ hat nun eine Doppelwurzel $\rho_0 = 0$, und eine Lösung der DGl (9.9) ist durch die Bessel–Funktion $J_0(x)$ gegeben:

$$y_1(x) = J_0(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{(m!)^2} \Big(\frac{x}{2}\Big)^{2m}, \quad x \in \mathbf{R}.$$

Eine weitere linear unabhängige Lösung erhält man gemäß Satz 16.31 aus dem Ansatz

$$\left.\begin{array}{lll}
y_2(x) = \displaystyle\sum_{n=0}^{\infty} b_n x^n + J_0(x)\,\ln x = \sum_{n=0}^{\infty} b_n x^n + \ln x \sum_{m=0}^{\infty} c_{2m} x^{2m} & \Big| & \cdot 1 \\
y_2'(x) = \displaystyle\sum_{n=0}^{\infty} n b_n x^{n-1} + \ln x \sum_{m=0}^{\infty} 2m c_{2m} x^{2m-1} + \sum_{m=0}^{\infty} c_{2m} x^{2m-1} & \Big| & \cdot \frac{1}{x} \\
y_2''(x) = \displaystyle\sum_{n=0}^{\infty} n(n-1) b_n x^{n-2} + \ln x \sum_{m=0}^{\infty} 2m(2m-1) c_{2m} x^{2m-2} + \sum_{m=0}^{\infty} (4m-1) c_{2m} x^{2m-2} & \Big| & \cdot 1
\end{array}\right\} (+)$$

In der Summe erhalten wir

$$b_0 + b_1 x^{-1} + 4b_2 + \sum_{n=3}^{\infty} (n^2 b_n + b_{n-2}) x^{n-2} + \sum_{m=0}^{\infty} 4m c_{2m} x^{2m-2} = 0,$$

wobei wir die Koeffizienten c_{2m} von $J_0(x)$ einsetzen müssen:

$$c_{2m} = \frac{(-1)^m}{2^{2m}(m!)^2} = -4(m+1)^2 c_{2m+2}.$$

Mit einem Koeffizientenvergleich ergibt sich:

$$b_0 = 0, \quad b_{2m+1} = 0 \ \forall\, m \geq 0, \quad b_{2m} = -\frac{1}{4m^2} b_{2m-2} - \frac{1}{m} c_{2m} \ \forall\, m \geq 1.$$

Aus der letzten Rekursion folgt – zum Beispiel mit vollständiger Induktion:

$$b_{2m} = -c_{2m}\Big(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{m}\Big), \qquad m \geq 1.$$

Somit haben wir eine zweite, linear unabhängige Lösung in der folgenden Form berechnet:

$$y_2(x) = J_0(x)\,\ln x - \sum_{m=1}^{\infty} \frac{(-1)^m}{(m!)^2}\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{m}\right)\left(\frac{x}{2}\right)^{2m}.$$

Es ist klar, dass mit $y_1(x), y_2(x)$ auch die Funktionen $y_1(x)$, $\alpha y_1(x)+\beta y_2(x)$, $\beta \neq 0$, ein Fundamentalsystem für die BESSEL–DGl (9.9) bilden. Mit der speziellen Wahl

$$\alpha := \frac{2}{\pi}\,(C-\ln 2), \quad \beta := \frac{2}{\pi}$$

erhält man schließlich

$$\alpha y_1(x)+\beta y_2(x) = \frac{2}{\pi}\left(\ln\frac{x}{2}+C\right)J_0(x) - \frac{2}{\pi}\sum_{m=1}^{\infty}\frac{(-1)^m}{(m!)^2}\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{m}\right)\left(\frac{x}{2}\right)^{2m} =: N_0(x).$$

Definition 16.23 *Die Funktion*

$$\boxed{N_0(x) := \frac{2}{\pi}\left(\ln\frac{x}{2}+C\right)J_0(x) - \frac{2}{\pi}\sum_{m=1}^{\infty}\frac{(-1)^m}{(m!)^2}\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{m}\right)\left(\frac{x}{2}\right)^{2m}}$$

heiße NEUMANN–**Funktion 0–ter Ordnung**. *Entsprechend wird für* $p \in \mathbf{N}$ *die* NEUMANN–**Funktion** p**–ter Ordnung** *oder* BESSEL*–Funktion zweiter Art der Ordnung* p *in der folgenden Weise definiert:*

$$\boxed{\begin{aligned} N_p(x) \;:=\; & \frac{2}{\pi}\left(\ln\frac{x}{2}+C\right)J_p(x) - \frac{1}{\pi}\left(\frac{x}{2}\right)^{-p}\sum_{m=0}^{p-1}\frac{(p-m-1)!}{m!}\left(\frac{x}{2}\right)^{2m} \\ & - \frac{1}{\pi}\left(\frac{x}{2}\right)^{p}\sum_{m=1}^{\infty}\frac{(-1)^m}{m!(m+p)!}\left(\frac{x}{2}\right)^{2m}\Big(\sum_{n=1}^{m}\frac{1}{n}+\sum_{n=p+1}^{m+p}\frac{1}{n}\Big). \end{aligned}}$$

Hierin ist $C := \lim\limits_{n\to\infty}\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}-\ln n\right) \doteq 0.577\,215$ *die* EULER–MASCHERONI–**Konstante**.

Abschließend zitieren wir noch ohne Beweis:

Satz 16.33 *Für* $p \in \mathbf{N}_0$ *bilden die beiden* BESSEL*–Funktionen* $J_p(x)$ *und* $N_p(x)$ *ein Fundamentalsystem für die* BESSEL*sche DGl* (9.9).

Bemerkung 16.19 Da die Differentialgleichung (9.9) invariant bleibt bei Übergang von p zu $-p$, genügt es, die Diskussion ausschließlich auf den Parameterbereich $p \geq 0$ zu beschränken. Dieser Bereich wird aber durch die beiden Sätze 16.32 und 16.33 vollständig abgedeckt. □

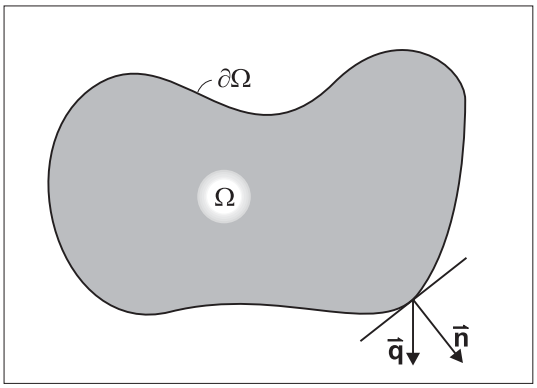
Kapitel 17

FOURIER–Reihen und ihre Bedeutung bei partiellen Differentialgleichungen

17.1 Der 1–dimensionale Wärmeleiter endlicher Länge

Ein wärmeleitendes Medium fülle einen Bereich $\Omega_0 \subset \mathbf{R}^3$ aus, der von einer geschlossenen Fläche $\partial\Omega_0$ berandet wird. Es bezeichne $u = u(x, y, z; t)$ die vom Ort $(x, y, z) \in \mathbf{R}^3$ und von der Zeit $t \geq 0$ abhängige **Temperaturverteilung** in Ω_0. Ist u nicht konstant, so gleichen sich nach der Erfahrung Temperaturunterschiede mit fortschreitender Zeit aus: Ein **Wärmestrom** fließt von den Stellen höherer Temperatur zu den Stellen niederer Temperatur. Ist das Medium homogen und isotrop, so gilt für diesen Wärmestrom $\vec{q}$ – pro Zeit– und Volumeneinheit gemessen – das von J.B. FOURIER empirisch gewonnene Gesetz

$$\vec{q} = -\kappa \operatorname{grad} u, \qquad \kappa > 0: \quad \text{Wärmeleitfähigkeit.} \tag{1.1}$$



Zur Herleitung der Wärmeleitungsgleichung

Der Wärmestrom ist dem Temperaturgradienten entgegengerichtet. Ist $\Omega \subset \Omega_0$ ein beliebiges Teilvolumen mit hinreichend glatter Berandung $\partial\Omega$, so ist der gesamte Wärmefluss Q, der pro Zeiteinheit durch Ω hindurchtritt, in der folgenden Weise durch ein **Oberflächenintegral** bestimmt:

$$Q = \int_{\partial\Omega} \langle \vec{n}, \vec{q} \rangle \, do, \qquad \vec{n}: \quad \text{äußere Einheitsnormale an } \partial\Omega.$$

Integrale von diesem Typ können mit Hilfsmitteln der **Vektoranalysis** in **Volumenintegrale** über das von $\partial\Omega$ eingeschlossene Volumen Ω umgeformt werden (GAUSSscher Integralsatz). Konkret folgt, indem man den Wärmestrom $\vec{q}$ aus (1.1) einsetzt:

$$Q = -\kappa \int_{\Omega} \Delta u(x, y, z; t) \, dx \, dy \, dz, \qquad \Delta := \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

Die Gesamtwärmemenge in Ω beträgt bei Ausschluss von Wärmequellen:

$$\int_\Omega c\rho u \, dx\, dy\, dz, \qquad \rho: \text{ Dichte}, \quad c: \text{ spezifische Wärme}.$$

Die zeitliche Änderung dieser Gesamtwärmemenge ist gerade der Wärmefluss Q, also

$$-\int_\Omega c\rho \frac{\partial u}{\partial t}\, dx\, dy\, dz = -\kappa \int_\Omega \Delta u \, dx\, dy\, dz.$$

Da diese Gleichung in jedem Teilvolumen $\Omega \subset \Omega_0$ richtig ist, muss in Ω_0 notwendigerweise die folgende **partielle Differentialgleichung der Wärmeleitung** gelten:

$$\boxed{u_t = k\,\Delta u, \qquad k := \frac{\kappa}{c\rho}: \quad \text{Temperaturleitfähigkeit.}} \tag{1.2}$$

Ist die Temperatur des Wärmeleiters nur in eindimensionaler Richtung veränderlich und wird diese Richtung als x–Richtung bezeichnet, so liegt der Sonderfall der **eindimensionalen Wärmeleitungsgleichung** vor:

$$\boxed{u_t - k u_{xx} = 0, \qquad 0 < x < L, \quad t > 0.} \tag{1.3}$$

Dieser partiellen DGl genügt zum Beispiel die Temperaturverteilung in einem Draht der Länge L, der gegen seitliche Wärmeabstrahlung isoliert ist.

Aus physikalischer Sicht möchte man gerne über den zeitlichen Temperaturverlauf an den Drahtenden $x = 0$ und $x = L$ verfügen. Werden diese Enden zum Beispiel in ein Eis–Wasser–Gemisch getaucht, so gelten dort die

- **Randbedingungen**: $u(0,t) = 0 = u(L,t)$ für $t > 0$. (RB)

Natürlich hat die Aufgabe (1.3), (RB) die triviale Lösung $u = 0$, die an sich uninteressant ist, da kein Temperaturausgleich stattfindet. Deshalb wird es sinnvoll sein, die Kenntnis der Temperaturverteilung zur Zeit $t = 0$ vorauszusetzen. Dies entspricht der Vorgabe einer

- **Anfangsbedingung**: $u(x,0) = f(x)$ für $0 \le x \le L$, (AB)

mit vorgelegter Funktion $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R})$. Die resultierende Aufgabe (1.3), (RB), (AB) heißt ein **Anfangs–Randwert–Problem** (ARWP) für die eindimensionale Wärmeleitungsgleichung, nämlich

- Finde ein $u = u(x,t)$ mit genügender Regularität, so dass gilt:

$$\text{(ARWP)} \quad \begin{cases} u_t - k u_{xx} = 0 & : \ 0 < x < L, \ \ t > 0, \qquad (1.3) \\ u(0,t) = 0 = u(L,t) & : \ t > 0, \qquad (\text{RB}) \\ u(x,0) = f(x) & : \ 0 \le x \le L. \qquad (\text{AB}) \end{cases}$$

Eine Lösung des ARWP gelingt durch Rückführung der partiellen DGl (1.3) auf eine Schar gewöhnlicher DGln. Man bewerkstelligt dies durch einen Bernoullischen **Separationsansatz** in der Form

$$\boxed{u(x,t) = X(x) \cdot T(t).}$$

Man sucht also partikuläre Lösungen als Produkte von Funktionen in jeweils nur einer der beiden Variablen x und t. Ein solcher Ansatz bedeutet **nicht**, dass die DGl (1.3) nur Lösungen in dieser Produktform zulässt. Wir werden in Abschnitt 17.5 noch andere Lösungen angeben. Wird das obige Produkt in die DGl (1.3) eingesetzt und werden danach die Variablen gemäß ihrer Abhängigkeit von t bzw. x getrennt, so resultiert

$$\frac{1}{k}\frac{\dot{T}}{T} = \frac{X''}{X} =: \alpha = const, \tag{1.4}$$

wobei wir wieder berücksichtigt haben, dass sich die beiden Gleichungsseiten wegen ihrer verschiedenen funktionalen Abhängigkeit nur in einer Konstanten treffen können. Wir erhalten das Paar gewöhnlicher DGln

$$\dot{T} - \alpha k T = 0, \qquad X'' - \alpha X = 0,$$

mit den allgemeinen Lösungen

$$T(t) = T_0 e^{\alpha k t}, \;\; t \geq 0, \qquad X(x) = a \cos x\sqrt{-\alpha} + b \sin x\sqrt{-\alpha}, \quad 0 \leq x \leq L.$$

Für $\alpha > 0$ würde $T(t)$ und somit auch $u(x,t)$ zeitlich unbeschränkt anwachsen. Dies steht im Widerspruch zur physikalischen Realität. Deshalb wird $\alpha := -\lambda^2 \leq 0$ ein realistischer Ansatz sein, aus dem sich nun partikuläre Lösungen der DGl (1.3) in folgender Form ergeben:

$$u_\lambda(x,t) = \begin{cases} a_0 + b_0 x & : \; \lambda = 0, \\ e^{-\lambda^2 k t}(a_\lambda \cos \lambda x + b_\lambda \sin \lambda x) & : \; \lambda \neq 0. \end{cases} \tag{1.5}$$

Diese Lösungen sind nun an die Bedingungen (RB) und (AB) anzupassen, wobei zuerst die **Anpassung der Randbedingungen** erfolgt.

- Erste Randbedingung:

$$u_\lambda(0,t) = 0 = \begin{cases} a_0 & : \; \lambda = 0, \\ a_\lambda e^{-\lambda^2 k t} & : \; \lambda \neq 0, \end{cases} \quad \Rightarrow \quad a_0 = 0 = a_\lambda.$$

- Zweite Randbedingung:

$$u_\lambda(L,t) = 0 = \begin{cases} b_0 L & : \; \lambda = 0, \\ b_\lambda e^{-\lambda^2 k t} \sin \lambda L & : \; \lambda \neq 0, \end{cases} \quad \Rightarrow \quad b_0 = 0 = \sin \lambda L.$$

Es wäre auch die Wahl $b_\lambda = 0$ möglich, die aber nur auf die triviale Lösung $u_\lambda = 0$ führt. Die Nullstellen der Funktion $\sin \lambda L$ werden für festes $L > 0$ gerade in den Werten

$$\lambda_n = \frac{n\pi}{L}, \qquad n \in \mathbf{N},$$

erreicht, und deswegen erhalten wir die folgende Familie partikulärer Lösungen

$$u_n(x,t) = b_n e^{-\left(\frac{n\pi}{L}\right)^2 k t} \sin \frac{n\pi x}{L}, \qquad n \in \mathbf{N}, \tag{1.6}$$

worin jede der Funktionen $u_n(x,t)$ die Randbedingungen (RB) erfüllt, während die Anfangsbedingung (AB) im allgemeinen von keinem u_n befriedigt wird. Da die DGl (1.3) **linear** ist, gilt jedoch das **Superpositionsprinzip**: Mit jedem einzelnen $u_n(x,t)$ ist auch jede Linearkombination der u_n Lösung der DGl (1.3), und es gelten die Randbedingungen (RB). Falls sogar die Reihe $\sum\limits_{n=1}^{\infty} |b_n n^2|$ konvergiert, so liefern die Weierstrass–Sätze die gleichmäßige Konvergenz der Funktionenreihe

$$\boxed{u(x,t) := \sum_{n=1}^{\infty} b_n e^{-\left(\frac{n\pi}{L}\right)^2 k t} \sin \frac{n\pi x}{L}} \tag{1.7}$$

und aller in der DGl (1.3) benötigten Ableitungen. Das heißt, man darf die Reihe (1.7) gliedweise differenzieren. Somit wird klar, dass die Funktion $u(x,t)$ selbst Lösung der DGl (1.3) und der Randbedingungen (RB) ist. Nun verbleibt das Problem der **Anpassung der Anfangsbedingung**. Können die Koeffizienten b_n in (1.7) so bestimmt werden, dass (AB) erfüllt ist? Das heißt, dass gilt:

$$\boxed{u(x,0) = f(x) = \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L}, \qquad 0 \leq x \leq L.} \tag{1.8}$$

Der französische Physiker J.B. FOURIER fand um 1822 die richtige Antwort. Er behauptete, dass die FOURIER–Koeffizienten b_n unter sehr allgemeinen Voraussetzungen an die Funktion $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R})$ gemäß der Vorschrift

$$\boxed{b_n = \frac{2}{L} \int_0^L f(\xi) \sin \frac{n\pi\xi}{L}\, d\xi, \qquad n \in \mathbf{N},} \tag{1.9}$$

zu bilden sind. Mit dieser Wahl ergibt sich nun die Lösung der Aufgabe ARWP in der Form

$$\boxed{u(x,t) = \frac{2}{L} \sum_{n=1}^{\infty} e^{-\left(\frac{n\pi}{L}\right)^2 kt} \sin \frac{n\pi x}{L} \Big(\int_0^L f(\xi) \sin \frac{n\pi\xi}{L}\, d\xi \Big).} \tag{1.10}$$

Im folgenden Abschnitt zeigen wir, wie man die Beziehungen (1.9) begründet.

17.2 FOURIER–Reihen und periodische Funktionen

Wir betrachten die Aufgabe, eine gegebene Funktion $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K})$ in einem festgelegten Intervall $I \subset \mathbf{R}$ durch eine Ersatzfunktion $g(x)$ im **quadratischen Mittel** zu approximieren. Diese Aufgabe tritt in den Anwendungen am häufigsten auf für **periodische Funktionen** $f(x)$, wobei $g(x)$ in der Klasse der FOURIER–Polynome zu bestimmen ist.

Definition 17.1 *Eine Funktion* $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K})$ *heiße* **periodisch mit der Periode** $T > 0$, *wenn gilt:*

$$\boxed{f(x+T) = f(x) \;\; \forall\, x \in \mathbf{R}.} \tag{2.1}$$

BSP. (17.2.1) Die konstante Funktion $f(x) := C = const$ ist periodisch mit **jeder** Periode T. Sie ist die einzige stetige Funktion mit dieser Eigenschaft.

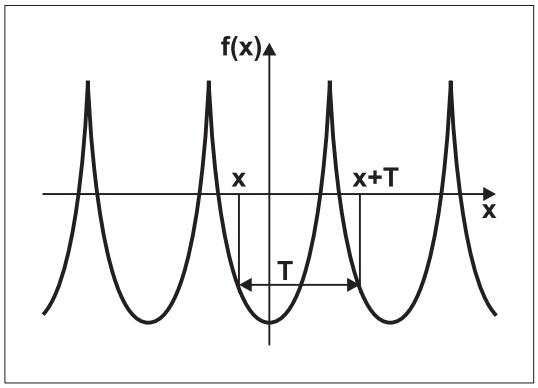
BSP. (17.2.2) Für festes $\omega > 0$ sind die Funktionen

$$f_1(x) := \cos \omega x, \quad f_2(x) := \sin \omega x, \quad f_3(x) := e^{i\omega x} = f_1(x) + i\, f_2(x)$$

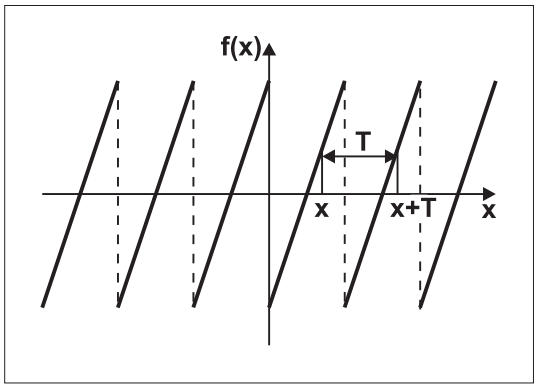
periodisch mit den Perioden

$$\boxed{T := \frac{2\pi n}{\omega}, \quad n \in \mathbf{N}.}$$

BSP. (17.2.3)



Stetige periodische Funktion



Unstetige periodische Funktion

Folgerung 17.1 *Ist $f \in \mathrm{Abb}(\mathbf{R}, \mathbf{K})$ periodisch mit der Periode $T > 0$, so auch mit jeder Periode $n \cdot T$, $n \in \mathbf{N}$.*

Folgerung 17.2 *Ist $f \in \mathrm{Abb}(\mathbf{R}, \mathbf{K})$ periodisch mit der Periode $T > 0$, so ist $\tilde{f}(x) := f(\omega x)$ für $\omega > 0$ periodisch mit der Periode $\tilde{T} := \frac{T}{\omega}$.*

Folgerung 17.3 *Ist $f \in \mathrm{Abb}(\mathbf{R}, \mathbf{K})$ periodisch mit der (kleinsten) Periode $T > 0$ und definiert man $\tilde{f}_j(x) := f(j\omega x)$ für festes $\omega > 0$ und $j = 0, 1, \ldots$, so hat das Funktionensystem*

$$f(0) =: \tilde{f}_0(x),\ \tilde{f}_1(x),\ \tilde{f}_2(x), \ldots$$

die kleinste gemeinsame Periode $\tilde{T} := \frac{T}{\omega}$.

Begründung: Die Funktion $\tilde{f}_j$ hat nach Folgerung 17.1 und 17.2 die Perioden $\tilde{T}_{jn} := n \cdot \frac{T}{j\omega}$ mit $n, j \in \mathbf{N}$, insbesondere also die Periode $\tilde{T}$. Dies ist aber die kleinste Periode für $\tilde{f}_1$. □

Folgerung 17.4 *Die Funktionensysteme*

$$\tilde{f}_0(x) := 1, \quad \tilde{f}_{2n}(x) := \cos(n\omega x), \quad \tilde{f}_{2n-1}(x) := \sin(n\omega x), \qquad n \in \mathbf{N}, \quad \omega > 0,$$

beziehungsweise

$$\tilde{f}_n(x) := e^{in\omega x}, \quad n \in \mathbf{Z},$$

haben jeweils die kleinste gemeinsame Periode $\tilde{T} := \frac{2\pi}{\omega}$.

Mit diesen Eigenschaften ist die folgende Definition hinreichend motiviert:

Definition 17.2 (a) *Eine unendliche Reihe der Form*

$$\boxed{T(x) := \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos(k\omega x) + b_k \sin(k\omega x)), \quad a_k,\, b_k \in \mathbf{R}, \quad \omega > 0,} \tag{2.2}$$

heiße **trigonometrische Reihe** *oder* FOURIER**–Reihe**. *Im Konvergenzfall ist die Grenzfunktion $T(x)$ periodisch mit der Periode $\tilde{T} = \frac{2\pi}{\omega}$. Die n–ten Partialsummen*

$$\boxed{T_n(x) := \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos(k\omega x) + b_k \sin(k\omega x)), \quad a_k,\, b_k \in \mathbf{R}, \quad \omega > 0,} \tag{2.3}$$

heißen **trigonometrische Polynome** *oder* FOURIER**–Polynome**.

(b) *Eine Reihe in der Form*

$$\boxed{E(x) := \sum_{k=-\infty}^{+\infty} \alpha_k e^{ik\omega x} := \sum_{k=0}^{\infty} \alpha_k e^{ik\omega x} + \sum_{k=1}^{\infty} \alpha_{-k} e^{-ik\omega x}, \quad \alpha_k \in \mathbf{C},} \tag{2.4}$$

heiße **komplexe** FOURIER**–Reihe.** *Sie heiße* **kovergent**, *wenn die beiden unendlichen Reihen auf der rechten Seite von (2.4) konvergieren. Im Konvergenzfall ist die Grenzfunktion $E(x)$ periodisch mit der Periode $\tilde{T} = \frac{2\pi}{\omega}$.*

Bemerkung 17.1 Da $e^{i\alpha} = \cos\alpha + i\sin\alpha$ gilt, können Real– und Imaginärteil einer komplexen FOURIER–Reihe jeweils als reelle FOURIER–Reihen dargestellt werden. Umgekehrt kann jede reelle FOURIER–Reihe als komplexe FOURIER–Reihe geschrieben werden. Dabei gelten die Koeffizientenrelationen □

$$\boxed{\alpha_k = \frac{1}{2}(a_k - ib_k), \quad \alpha_{-k} = \frac{1}{2}(a_k + ib_k), \qquad k \in \mathbf{N}_0, \quad b_0 := 0.} \tag{2.5}$$

Gegeben seien nun ein Intervall $I := [a, a+\tilde{T}]$ und eine Funktion $f : I \to \mathbf{K}$. Wir denken uns $f(x)$ in irgendeiner Weise $\tilde{T}$–periodisch auf ganz $\mathbf{R}$ fortgesetzt. Die Aufgabe, $f(x)$ in Form einer FOURIER–Reihe darzustellen, ist ein Spezialfall des folgenden Problems:

- **Allgemeines Reihenentwicklungsproblem:** Auf einem Intervall $I \subset \mathbf{R}$ sei ein Funktionensystem
$$\varphi_0(x),\ \varphi_1(x),\ \varphi_2(x), \ldots$$
gegeben. Gesucht ist für eine gegebene Funktion $f : I \to \mathbf{K}$ eine Reihenentwicklung nach dem vorgegebenen System in der Form
$$\boxed{f(x) = \sum_{k=0}^{\infty} a_k \varphi_k(x).} \tag{2.6}$$

BSP. (17.2.4) TAYLOR–Entwicklungen nach dem Funktionensystem 1, x, x^2, x^3, ...: Bekanntlich erhält man für reell–analytische Funktionen $f : I \to \mathbf{K}$ mit $0 \in I$ die Koeffizienten a_k der Entwicklung (2.6) aus den Beziehungen $a_k = \frac{f^{(k)}(0)}{k!}$, $k \in \mathbf{N}_0$.

Ein Nachteil der TAYLOR–Entwicklungen sind die sehr hohen Differenzierbarkeitsanforderungen an die Funktion $f(x)$. Hat das System $\varphi_0(x)$, $\varphi_1(x)$, $\varphi_2(x), \ldots$ die fundamentalen Eigenschaften eines **Orthogonalsystems**, so kann das allgemeine Reihenentwicklungsproblem unter weitaus schwächeren Voraussetzungen an $f(x)$ gelöst werden. Zur Formulierung eines Orthogonalbegriffes sei zunächst nochmals an den Vektorraum $L(I)$ der über dem Intervall I LEBESGUE–integrierbaren Funktionen $f \in \mathrm{Abb}(\mathbf{R}, \mathbf{K})$ erinnert, den wir in Abschnitt 8.6 eingeführt hatten. Für die Theorie der FOURIER–Reihen hat nun der folgende Unterraum von $L(I)$ eine ganz zentrale Bedeutung:

$$\boxed{L^2(I) := \{f \in L(I) \,:\, \int_I |f(x)|^2\, dx < \infty\}.}$$

Dieser Vektorraum über $\mathbf{K}$ heiße der Funktionenraum der **quadratisch Lebesgue–integrierbaren Funktionen** über I.

Bemerkung 17.2 (a) Im Sinne der LEBESGUEschen Integrationstheorie sind Funktionen f_1, $f_2 \in L(I)$ mit der Eigenschaft
$$\int_I |f_1(x) - f_2(x)|\, dx = 0$$
auf der Menge I **fast überall gleich**, siehe Satz 8.36(b). Das heißt, die Punktmenge $N := \{x \in I \,:\, f_1(x) \neq f_2(x)\} \subset \mathbf{R}$ ist eine **Nullmenge**. Wir schreiben in diesem Fall $f_1(x) \sim f_2(x)$, und man überprüft sehr einfach, dass durch "$\sim$" eine **Äquivalenzrelation** auf dem Funktionenraum $L(I)$ induziert wird. Somit zerfällt $L(I)$ in Klassen äquivalenter Funktionen: Je zwei Repräsentanten derselben Klasse unterscheiden sich nur auf einer Nullmenge. Die so

strukturierte Menge $L(I)$ der Äquivalenzklassen bezeichnen wir wieder mit demselben Symbol $L(I)$ und beachten, dass weiterhin ein Vektorraum über dem Körper $\mathbf{K}$ vorliegt. Addition und λ–Multiplikation der Klassen sind durch die entsprechenden Operationen ihrer Repräsentanten erklärt. In diesem Sinne ist nun auch die obige Definition des Teilraums $L^2(I)$ zu verstehen.

(b) Für $f,\, g \in L^2(I)$ gilt die SCHWARZsche Ungleichung

$$\Big|\int_I f(x)g(x)\,dx\Big| \le \int_I |f(x)||g(x)|\,dx \le \Big(\int_I |f(x)|^2\,dx\Big)^{1/2}\Big(\int_I |g(x)|^2\,dx\Big)^{1/2} < \infty. \tag{2.7}$$

Setzt man hier speziell $g = 1$, so folgt die Implikation $f \in L^2(I) \Rightarrow \int\limits_I |f(x)|\,dx < \infty$; das heißt, $L^2(I)$ ist tatsächlich ein Unterraum von $L(I)$.

(c) Wegen (2.7) existiert natürlich das **Skalarprodukt**

$$\boxed{\langle f, g\rangle_2 := \int_I f(x)\overline{g(x)}\,dx, \qquad f,\, g \in L^2(I),} \tag{2.8}$$

und somit wird auf $L^2(I)$ eine **Norm**

$$\boxed{L^2(I) \ni f \mapsto \|f\|_2 := \sqrt{\langle f, f\rangle_2}} \tag{2.9}$$

induziert. Dabei ist es wesentlich, dass Funktionen $f,\, g \in L^2(I)$, die sich nur auf einer Nullmenge unterscheiden, miteinander identifiziert werden. Das heißt, der Funktionenraum $L^2(I)$ ist nicht nur ein **Praehilbertraum** über dem Körper $\mathbf{K}$, sondern unter der induzierten Norm (2.9) auch ein **normierter Vektorraum**. In diesem Sinne kann sogar gezeigt werden, dass $L^2(I)$ **vollständig** ist. Man nennt Praehilberträume, die unter der induzierten Norm vollständig sind, auch **Hilberträume**. □

Definition 17.3 *Gegeben seien feste Zahlen $a \in \mathbf{R}$ sowie $T > 0$, und es gelte $I := [a, a+T]$. Ein System stetiger Funktionen $(\varphi_k(x))_{k\ge 0}$ heiße ein* **Orthogonalsystem** *auf I, falls gilt:*

$$\boxed{\langle \varphi_j, \varphi_k\rangle_2 = \int_a^{a+T} \varphi_j(x)\overline{\varphi_k(x)}\,dx \quad \begin{cases} = 0 \;:\; j \ne k, \\ > 0 \;:\; j = k, \end{cases} \qquad j,k = 0,1,\ldots} \tag{2.10}$$

BSP. (17.2.5) Das Funktionensystem

$$\varphi_0(x) := \frac{1}{\sqrt{2}}, \quad \varphi_{2k}(x) := \cos kx, \quad \varphi_{2k-1}(x) := \sin kx, \qquad k \in \mathbf{N},$$

bildet ein Orthogonalsystem auf **jedem** Intervall $I := [a, a+2\pi]$ der Länge $T := 2\pi$. Man rechnet nämlich leicht nach:

$$\boxed{\int_a^{a+2\pi} \cos kx \cdot \sin jx\,dx = 0 \;\; \forall\, j,k \in \mathbf{N}_0,} \tag{2.11}$$

$$\boxed{\int_a^{a+2\pi} \cos kx \cdot \cos jx\,dx = \int_a^{a+2\pi} \sin kx \cdot \sin jx\,dx = 0 \;\; \forall\, j \ne k, \quad j,k \in \mathbf{N}_0,} \tag{2.12}$$

$$\boxed{\int_a^{a+2\pi} \varphi_k^2(x)\,dx = \pi \;\; \forall\, k \in \mathbf{N}_0.} \tag{2.13}$$

BSP. (17.2.6) Das Funktionensystem

$$\varphi_0(x) := \frac{1}{\sqrt{2}}, \quad \varphi_{2k}(x) := \cos(k\omega x), \quad \varphi_{2k-1}(x) := \sin(k\omega x), \qquad k \in \mathbf{N}, \quad \omega > 0,$$

bildet ein Orthogonalsystem auf **jedem** Intervall $I := [a, a+T]$ der Länge $T := \frac{2\pi}{\omega}$. Es gilt hier:

$$\boxed{\int_a^{a+T} \varphi_k^2(x)\,dx = \frac{T}{2} = \frac{\pi}{\omega} \;\; \forall\, k \in \mathbf{N}_0.} \tag{2.14}$$

BSP. (17.2.7) Das Funktionensystem $\varphi_k(x) := e^{ikx}$, $k \in \mathbf{Z}$, bildet ein Orthogonalsystem auf **jedem** Intervall $I := [a, a+2\pi]$ der Länge $T := 2\pi$. Es gilt nämlich:

$$\boxed{\int_a^{a+2\pi} \varphi_k(x)\overline{\varphi_j(x)}\,dx = \int_a^{a+2\pi} e^{i(k-j)x}\,dx = \begin{cases} \frac{1}{i(k-j)} e^{i(k-j)a}(e^{2\pi i(k-j)} - 1) = 0 & : \; k \neq j, \\ 2\pi & : \; k = j. \end{cases}} \tag{2.15}$$

Die Frage, in welcher Weise die Koeffizienten a_k in (2.6) mit der gegebenen Funktion $f(x)$ zusammenhängen, kann für einen Spezialfall einfach beantwortet werden. Es gilt nämlich:

Satz 17.1 *Gegeben sei auf dem Intervall $I := [a, a+T]$ ein Orthogonalsystem $(\varphi_k(x))_{k\geq 0}$ stetiger Funktionen. Konvergiert die Reihe*

$$\boxed{\Phi(x) := \sum_{k=0}^{\infty} a_k \varphi_k(x)} \tag{2.16}$$

gleichmäßig *auf I, so ist die Grenzfunktion $\Phi : I \to \mathbf{K}$ stetig, und es gilt*

$$\boxed{a_k = \int_a^{a+T} \Phi(x)\overline{\varphi_k(x)}\,dx \Big/ \int_a^{a+T} |\varphi_k(x)|^2\,dx \;\; \forall\, k \in \mathbf{N}_0.} \tag{2.17}$$

Begründung: Aus dem Weierstrass–Kriterium folgt bei gleichmäßiger Konvergenz schon die Stetigkeit der Grenzfunktion $\Phi(x)$. Ferner darf die Reihe (2.16) gliedweise bestimmt integriert werden. Dazu multiplizieren wir (2.16) mit $\overline{\varphi_j(x)}$ und integrieren über das Intervall I:

$$\int_a^{a+T} \Phi(x)\overline{\varphi_j(x)}\,dx = \sum_{k=0}^{\infty} a_k \int_a^{a+T} \varphi_k(x)\overline{\varphi_j(x)}\,dx \overset{(2.10)}{=} a_j \,\langle \varphi_j, \varphi_j\rangle_2 \;\; \forall\, j \in \mathbf{N}_0.$$

Dies war behauptet worden. □

Die Zahl $a \in \mathbf{R}$, die in den Orthogonalsystemen der BSPe (17.2.5–7) auftritt, kann beliebig gewählt werden. Wir machen nachfolgend Gebrauch davon.

Folgerung 17.5 *Konvergiert die* Fourier*–Reihe*

$$T(x) := \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$$

auf dem Intervall $I := [-\pi, +\pi]$ **gleichmäßig**, *so gilt notwendig*

$$\boxed{a_k = \frac{1}{\pi}\int_{-\pi}^{+\pi} T(x)\cos kx\,dx, \quad b_k = \frac{1}{\pi}\int_{-\pi}^{+\pi} T(x)\sin kx\,dx \;\; \forall\, k \in \mathbf{N}_0.}$$

Konvergiert die FOURIER*–Reihe*

$$T(x) := \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos(k\omega x) + b_k \sin(k\omega x)\right)$$

auf dem Intervall $I := [0, T]$ *mit* $T := \frac{2\pi}{\omega}$ **gleichmäßig**, *so gilt notwendig*

$$\boxed{a_k = \frac{2}{T} \int_0^T T(x) \cos(k\omega x)\, dx, \quad b_k = \frac{2}{T} \int_0^T T(x) \sin(k\omega x)\, dx \;\; \forall\, k \in \mathbf{N}_0.}$$

Konvergiert die komplexe FOURIER*–Reihe*

$$E(x) := \sum_{k=-\infty}^{+\infty} \alpha_k e^{ikx}$$

auf dem Intervall $I := [-\pi, +\pi]$ **gleichmäßig**, *so gilt notwendig*

$$\boxed{\alpha_k = \frac{1}{2\pi} \int_{-\pi}^{+\pi} E(x) e^{-ikx}\, dx \;\; \forall\, k \in \mathbf{Z}.}$$

Bemerkung 17.3 Nach dem WEIERSTRASS–Kriterium liegt die oben geforderte gleichmäßige Konvergenz sicher vor, wenn die Reihen $\sum\limits_{k=1}^{\infty} (|a_k| + |b_k|)$ bzw. $\sum\limits_{k=-\infty}^{+\infty} |\alpha_k|$ konvergieren. □

17.3 FOURIER–Reihen und Konvergenz im quadratischen Mittel

Gegeben sei auf dem Intervall $I := [a, a+T]$ ein Orthogonalsystem $(\varphi_k(x))_{k\geq 0}$ stetiger Funktionen. Die Zahlen $a \in \mathbf{R}$ und $T > 0$ seien fest gewählt. Dann gibt Satz 17.1 Anlass zu folgender

Definition 17.4 *Gegeben sei eine integrable Funktion* $f \in L(I)$. *Dann heißen die Zahlen*

$$\boxed{a_k := \int_a^{a+T} f(x)\overline{\varphi_k(x)}\, dx \Big/ \int_a^{a+T} |\varphi_k(x)|^2\, dx, \qquad k \in \mathbf{N}_0,} \tag{3.1}$$

die FOURIER**–Koeffizienten** *der Funktion* $f(x)$ *bezüglich des vorgelegten Orthogonalsystems* $(\varphi_k(x))_{k\geq 0}$. *Die mit diesen Koeffizienten gebildete Funktionenreihe*

$$f(x) \sim \sum_{k=0}^{\infty} a_k \varphi_k(x), \qquad x \in I, \tag{3.2}$$

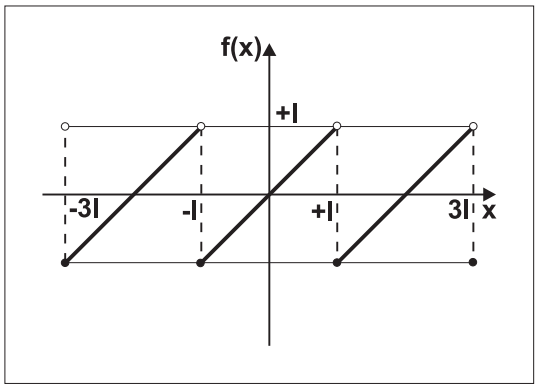
heiße die FOURIER**–Reihe** *der Funktion* $f(x)$ *bezüglich des Orthogonalsystems* $(\varphi_k(x))_{k\geq 0}$.

Es ist noch zu klären, wie das Symbol "$\sim$" in (3.2) zu verstehen ist. Es ist klar, dass "$\sim$" durch das Gleichheitszeichen "=" ersetzt werden kann, wenn die Funktion $f(x)$ stetig ist und wenn die Reihe (3.2) auf dem Intervall I gleichmäßig konvergiert.

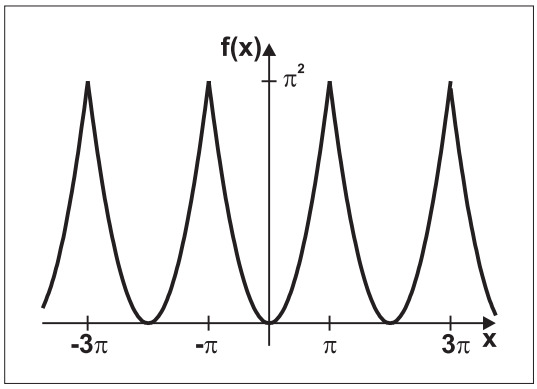
BSP. (17.3.1) Für eine feste Zahl $\ell > 0$ setzen wir

$$f(x) := \begin{cases} x & : \; x \in [-\ell, +\ell), \\ f(x - 2\ell) & : \; x \in \mathbf{R} \text{ sonst.} \end{cases}$$

Dann ist $f(x)$ periodisch mit der Periode $T = 2\ell$. Ferner ist $f(x)$ stetig auf ganz $\mathbf{R}$ mit Ausnahme der Punkte $x_j := (2j+1)\ell,\; j \in \mathbf{Z}$, in denen $f(x)$ Sprungstellen besitzt, siehe folgende Grafik.



Der Graph der 2ℓ–periodischen Funktion $f(x) := x,\ -\ell \le x < \ell$



Der Graph der 2π–periodischen Funktion $f(x) := x^2,\ -\pi \le x \le \pi$

Wählen wir in dem Orthogonalsystem aus BSP. (17.2.6) die Zahl ω gemäß $\omega := \frac{\pi}{\ell}$, ferner $a := -\ell$, so hat die FOURIER–Reihe

$$f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos\frac{k\pi x}{\ell} + b_k \sin\frac{k\pi x}{\ell}), \qquad x \in \mathbf{R},$$

die folgenden FOURIER–Koeffizienten:

$$a_k = \frac{2}{T}\int_{-\ell}^{-\ell+T} f(x)\cos\frac{k\pi x}{\ell}\,dx = \frac{1}{\ell}\int_{-\ell}^{\ell} x\cos\frac{k\pi x}{\ell}\,dx = 0 \ \ \forall\, k \in \mathbf{N}_0,$$

$$b_k = \frac{1}{\ell}\int_{-\ell}^{\ell} x\sin\frac{k\pi x}{\ell}\,dx = \frac{1}{\ell}\Big(-\frac{x\ell}{k\pi}\cos\frac{k\pi x}{\ell}\Big|_{-\ell}^{+\ell} + \frac{\ell^2}{k^2\pi^2}\sin\frac{k\pi x}{\ell}\Big|_{-\ell}^{+\ell}\Big) = \frac{2\ell}{k\pi}(-1)^{k+1} \ \ \forall\, k \in \mathbf{N}.$$

Die FOURIER–Reihe lautet also:

$$f(x) \sim \frac{2\ell}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}\sin\frac{k\pi x}{\ell}, \qquad x \in \mathbf{R},$$

und diese Reihe ist keineswegs gleichmäßig konvergent. Wir sehen nämlich:

- Für $x = 0$ ist der Summenwert der Reihe gleich 0, und das ist der Funktionswert $f(0)$.
- Für $x = -\ell$ ist der Summenwert der Reihe gleich 0, während der Funktionswert $f(-\ell) = -\ell$ beträgt.

Wir dürfen also keineswegs "$\sim$" durch "$=$" ersetzen. Man vermutet richtig, dass dieser "Defekt" nur in den Unstetigkeitsstellen von $f(x)$ auftritt.

BSP. (17.3.2) Wir betrachten die auf ganz $\mathbf{R}$ stetige und 2π–periodische Funktion

$$f(x) := \begin{cases} x^2 & : \ x \in [-\pi, +\pi], \\ f(x - 2\pi) & : \ x \in \mathbf{R} \text{ sonst.} \end{cases}$$

Wir wählen das Orthogonalsystem aus BSP. (17.2.5) mit $a := -\pi$ und bilden die FOURIER–Reihe

$$f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{\infty}(a_k\,\cos kx + b_k\,\sin kx), \qquad x \in \mathbf{R}.$$

Für die FOURIER–Koeffizienten ergibt sich durch partielle Integration:

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} x^2\,dx = \frac{2\pi^2}{3},$$

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} x^2\cos kx\,dx = \frac{1}{\pi}\Big(\frac{x^2\sin kx}{k}\Big|_{-\pi}^{\pi} - \frac{2}{k}\int_{-\pi}^{\pi} x\sin kx\,dx\Big) = \frac{2x\cos kx}{k^2\pi}\Big|_{-\pi}^{\pi} = (-1)^k\,\frac{4}{k^2},$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} x^2\sin kx\,dx = 0, \quad k \in \mathbf{N},$$

wobei das letzte Integral verschwindet, weil der Integrand eine ungerade Funktion ist. Es resultiert nun die FOURIER–Reihe

$$f(x) \sim \frac{\pi^2}{3} + 4\sum_{k=1}^{\infty} \frac{(-1)^k}{k^2} \cos kx, \qquad x \in \mathbf{R},$$

die die konvergente Majorante $\frac{\pi^2}{3} + \sum\limits_{k=1}^{\infty} \frac{4}{k^2}$ besitzt. Somit liegt gleichmäßige Konvergenz der FOURIER–Reihe vor, und da die Funktion $f(x)$ auf ganz $\mathbf{R}$ stetig ist, darf "$\sim$" durch das Gleichheitszeichen "$=$" ersetzt werden. Betrachten wir den speziellen Wert $x = \pi$, so muss also $\pi^2 = \frac{\pi^2}{3} + \sum\limits_{k=1}^{\infty} \frac{4}{k^2}$ gelten, und daraus erhält man den Reihenwert

$$\boxed{\frac{\pi^2}{6} = \sum_{k=1}^{\infty} \frac{1}{k^2}.}$$

In den BSPn (17.3.1) und (17.3.2) zeigte es sich, dass **Symmetrie–Eigenschaften** der Funktion $f(x)$ das Verschwinden der FOURIER–Koeffizienten a_k oder b_k verursachen können. Wir haben in der Tat:

Folgerung 17.6 *Die T–periodische Funktion $f \in \mathrm{Abb}(\mathbf{R}, \mathbf{K})$ sei auf dem Periodenintervall $[0, T]$ L–integrierbar.*

(a) *Ist f eine* **gerade** *Funktion: $f(x) = f(-x)\ \forall\, x \in \mathbf{R}$, so ist f als reine* **Cosinus–Reihe** *darstellbar:*

$$\boxed{f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos\frac{2\pi kx}{T} \quad mit \quad a_k = \frac{4}{T}\int_0^{T/2} f(x)\cos\frac{2\pi kx}{T}\,dx \ \ \forall\, k \in \mathbf{N}_0.}$$

(b) *Ist f eine* **ungerade** *Funktion: $f(x) = -f(-x)\ \forall\, x \in \mathbf{R}$, so ist f als reine* **Sinus–Reihe** *darstellbar:*

$$\boxed{f(x) \sim \sum_{k=1}^{\infty} b_k \sin\frac{2\pi kx}{T} \quad mit \quad b_k = \frac{4}{T}\int_0^{T/2} f(x)\sin\frac{2\pi kx}{T}\,dx \ \ \forall\, k \in \mathbf{N}.}$$

Begründungen: (a) Da f eine gerade Funktion ist, muss die Funktion $f(x)\sin\frac{2\pi kx}{T}$ ungerade sein. Deshalb gilt

$$b_k = \frac{2}{T}\int_{-T/2}^{T/2} f(x)\sin\frac{2\pi kx}{T}\,dx = 0 \ \ \forall\, k \in \mathbf{N}.$$

Weil $f(x)\cos\frac{2\pi kx}{T}$ eine gerade Funktion ist, gilt hingegen

$$a_k = \frac{2}{T}\int_{-T/2}^{T/2} f(x)\cos\frac{2\pi kx}{T}\,dx = 2\cdot\frac{2}{T}\int_0^{T/2} f(x)\cos\frac{2\pi kx}{T}\,dx \ \ \forall\, k \in \mathbf{N}_0.$$

Ganz analog zeigt man (b). □

Wir waren bisher davon ausgegangen, dass die T–periodische Funktion f auf ganz $\mathbf{R}$ oder zumindest auf einem Periodenintervall $[a, a+T]$ vorgegeben ist. Häufig wird man aber vor das Problem gestellt, die trigonometrische FOURIER–Reihe der Funktion f zu bestimmen, die nur auf einem Intervall $[a, a+L]$ vorgelegt ist. Da die Funktion f auf beliebig viele Arten auf ganz $\mathbf{R}$ als T–periodische Funktion fortgesetzt werden kann, ist das oben genannte Problem in dieser allgemeinen Form nicht lösbar. In vielen Fällen verfügt man jedoch über eine der folgenden Zusatzinformationen:

- L ist die Periodenlänge,
- $a = 0$, L ist die halbe Periodenlänge und $f(x)$ ist als ungerade Funktion auf das Intervall $[-L, 0)$ fortzusetzen. Man bestimme also die FOURIER–Reihe in der Form

$$f(x) \sim \sum_{k=1}^{\infty} b_k \sin\frac{k\pi x}{L}, \qquad x \in \mathbf{R},$$

- $a = 0$, L ist die halbe Periodenlänge und $f(x)$ ist als gerade Funktion auf das Intervall $[-L, 0)$ fortzusetzen. Man bestimme also die FOURIER–Reihe in der Form

$$f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos \frac{k\pi x}{L}, \qquad x \in \mathbf{R}.$$

BSP. (17.3.3) Wir betrachten auf dem Intervall $[0, L]$ die Funktion $f(x) := x$ und diskutieren verschiedene periodische Fortsetzungen von f auf die ganze reelle Achse.

(i) Die Funktion f wird mit der Periode $T := L$ auf ganz $\mathbf{R}$ fortgesetzt:

$$f(x + L) := f(x) \ \ \forall\, x \in \mathbf{R}.$$

Wir setzen die FOURIER–Reihe von f in der Form

$$f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos \frac{2\pi kx}{L} + b_k \sin \frac{2\pi kx}{L}), \qquad x \in \mathbf{R},$$

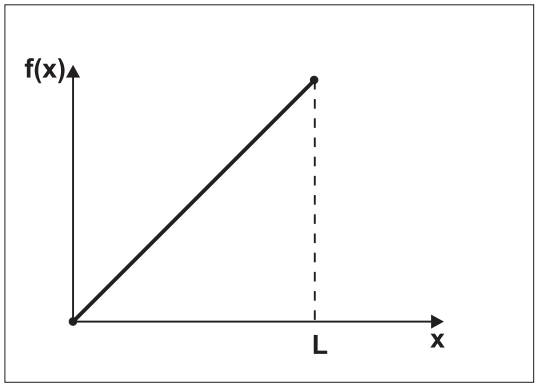
mit den folgenden FOURIER–Koeffizienten an:

$$a_k = \frac{2}{L} \int_0^L x \cos \frac{2\pi kx}{L}\, dx = \begin{cases} 0 & : \ k \in \mathbf{N}, \\ L & : \ k = 0, \end{cases}$$

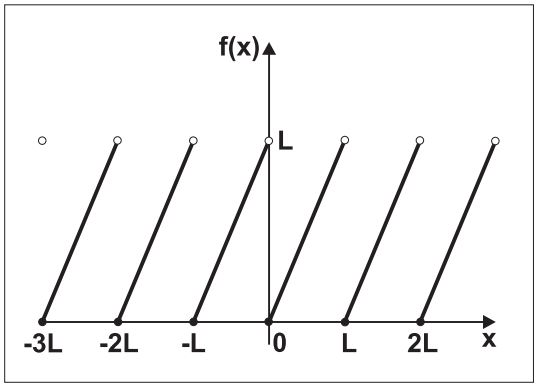
$$b_k = \frac{2}{L} \int_0^L x \sin \frac{2\pi kx}{L}\, dx = -\frac{2xL}{2k\pi L}\, x \cos \frac{2k\pi x}{L}\Big|_0^L = -\frac{L}{k\pi}, \quad k \in \mathbf{N}.$$

Daraus resultiert die Darstellung

$$f(x) \sim \frac{L}{2} - \frac{L}{\pi} \sum_{k=1}^{\infty} \frac{1}{k} \sin \frac{2k\pi x}{L}.$$



Die Funktion $f(x) := x$ auf dem Intervall $[0, L]$



Die Funktion $f(x)$ wird mit der Periode $T := L$ auf $\mathbf{R}$ fortgesetzt

Man erkennt an der Skizze, dass die Funktion $f(x) - \frac{L}{2}$ eine **ungerade** Funktion ist. Deshalb wird das hier erzielte Resultat auf Grund von Folgerung 17.6 plausibel.

(ii) Die Funktion f wird als **gerade** Funktion auf das Intervall $[-L, 0)$ fortgesetzt und danach als $2L$–periodische Funktion auf ganz $\mathbf{R}$ definiert:

$$f(x) := |x| \ \ \forall\, x \in [-L, L], \qquad f(x + 2L) := f(x) \ \ \forall\, x \in \mathbf{R}.$$

Nun kann die Funktion $f(x)$ mit der Periodenlänge $2L$ als reine Cosinus–Reihe dargestellt werden:

$$f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos \frac{k\pi x}{L}, \qquad x \in \mathbf{R}.$$

Wir berechnen die FOURIER–Koeffizienten gemäß

$$a_0 = \frac{2}{L}\int_0^L x\,dx = L,$$

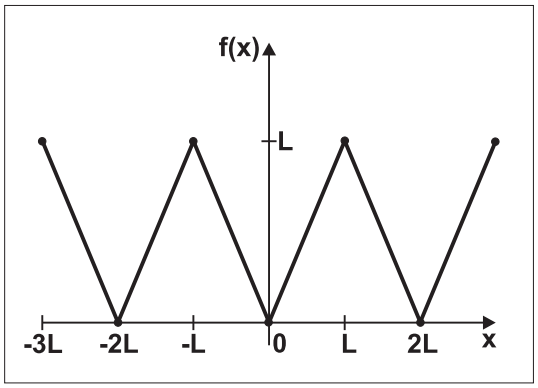
$$a_k = \frac{2}{L}\int_0^L x\cos\frac{k\pi x}{L}\,dx = \frac{2}{L}\left(\frac{xL}{k\pi}\sin\frac{k\pi x}{L} + \frac{L^2}{k^2\pi^2}\cos\frac{k\pi x}{L}\right)\Big|_0^L = \frac{2L}{k^2\pi^2}\left((-1)^k - 1\right).$$

Für $k = 2n$ resultiert $a_{2n} = 0$, und für $k = 2n+1$ erhalten wir $a_{2n+1} = -\frac{4L}{\pi^2}\frac{1}{(2n+1)^2}$. Somit folgt

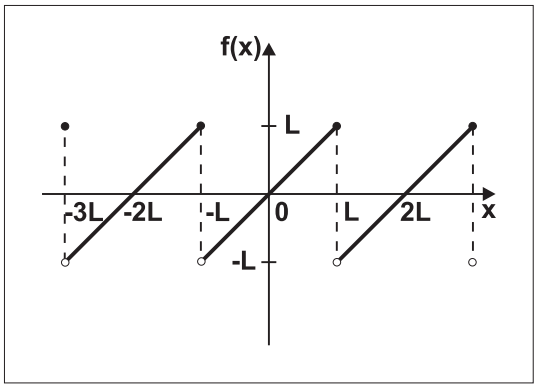
$$f(x) \sim \frac{L}{2} - \frac{4L}{\pi^2}\sum_{n=0}^{\infty}\frac{1}{(2n+1)^2}\cos\frac{(2n+1)\pi x}{L}.$$

Wegen der gleichmäßigen Konvergenz dieser FOURIER–Reihe und der Stetigkeit der Funktion $f(x)$ auf ganz $\mathbf{R}$ dürfen wir wieder "$\sim$" durch "$=$" ersetzen. Für $x = L$ erhalten wir wegen $\cos(2n+1)\pi = -1$ die Identität $\frac{L}{2} = \frac{4L}{\pi^2}\sum\limits_{n=0}^{\infty}\frac{1}{(2n+1)^2}$ und daraus den Summenwert

$$\boxed{\frac{\pi^2}{8} = \sum_{n=0}^{\infty}\frac{1}{(2n+1)^2},}$$



Die Funktion $f(x)$ wird mit der Periode $2L$ als gerade Funktion fortgesetzt



Die Funktion $f(x)$ wird mit der Periode $2L$ als ungerade Funktion fortgesetzt

(iii) Die Funktion f wird als **ungerade** Funktion auf das Intervall $(-L, 0)$ fortgesetzt und danach als $2L$–periodische Funktion auf ganz $\mathbf{R}$ definiert:

$$f(x) := x \;\; \forall\, x \in (-L, L], \qquad f(x+2L) := f(x) \;\; \forall\, x \in \mathbf{R}.$$

Es resultiert wiederum die FOURIER–Reihe aus BSP. (17.3.1).

In BSP. (17.3.1) ändert sich nichts an der Berechnung der FOURIER–Koeffizienten a_k, b_k, wenn die Funktion $f(x)$ gemäß $f(x) := x$ für $x \in (-\ell, +\ell]$ und $f(x+2\ell) := f(x)\;\; \forall\, x \in \mathbf{R}$ erklärt wird. Wir haben jetzt $f(+\ell) = \ell$, im Gegensatz zu $f(+\ell) = -\ell$ in BSP. (17.3.1). Dass für beide Funktionen dieselbe FOURIER–Reihe herauskommt, liegt an folgendem

Satz 17.2 *Sind die T–periodischen Funktionen $f_1(x)$ und $f_2(x)$ auf dem Periodenintervall $[0, T]$ L–integrierbar und gilt*

$$\int_0^T |f_1(x) - f_2(x)|\,dx = 0, \tag{3.3}$$

so haben $f_1(x)$ und $f_2(x)$ dieselben FOURIER*–Reihen.*

Begründung: Gemäß Satz 8.36(b) gilt wegen (3.3) $f_1(x) = f_2(x)$ fast überall auf $[0,T]$. Deshalb haben die Funktionen $f_1(x)$ und $f_2(x)$ dieselben FOURIER–Koeffizienten a_k, b_k und somit auch dieselben FOURIER–Reihen. □

Satz 17.2 trifft insbesondere zu, wenn wir Funktionen $f_1,\ f_2 \in L^2(I)$, $I := [a, a+T]$, betrachten, die in irgendeiner Form T–periodisch auf ganz $\mathbf{R}$ fortgesetzt werden. Wegen (3.3) unterscheiden sich aber f_1 und f_2 höchstens auf einer Nullmenge. Im Sinne von $L^2(I)$ werden solche Funktionen miteinander identifiziert; sie sind lediglich zwei *Repräsentanten derselben Äquivalenzklasse*. Wir können nun mit dieser Interpretation des Funktionenraumes $L^2(I)$ zeigen, dass die folgende Approximationsaufgabe für eine Funktion $f \in L^2(I)$ genau von dem FOURIER–Polynom der Funktion $f(x)$ mit den Koeffizienten (3.1) gelöst wird.

- **Approximationsaufgabe im quadratischen Mittel:** Bestimme zu gegebener Funktion $f \in L^2(I)$ und zu gegebenem Orthogonalsystem $(\varphi_k(x))_{k\geq 0}$ dasjenige

$$g_n(x) := \sum_{k=0}^{n} a_k \varphi_k(x),$$

  welches die folgende Extremaleigenschaft besitzt:

$$\boxed{\|f - g_n\|_2 = \Big(\int_I |f(x) - g_n(x)|^2 \, dx \Big)^{1/2} \stackrel{!}{=} \text{Min.}}$$

Satz 17.3 *Auf dem Intervall $I := [a, a+T]$ sei ein Orthogonalsystem $(\varphi_k(x))_{k\geq 0}$ stetiger Funktionen mit der folgenden Eigenschaft gegeben:*

$$\int_a^{a+T} |\varphi_k(x)|^2 \, dx =: \gamma = const \ \ \forall\ k \in \mathbf{N}_0. \tag{3.4}$$

Für eine gegebene Funktion $f \in L^2(I)$ seien

$$\boxed{a_k := \frac{1}{\gamma} \int_a^{a+T} f(x)\overline{\varphi_k(x)}\, dx, \qquad k \in \mathbf{N}_0,} \tag{3.5}$$

die FOURIER*–Koeffizienten von $f(x)$, und es sei für $\vec{\alpha} := (\alpha_0, \alpha_1, \ldots, \alpha_n)^T \in \mathbf{K}^{n+1}$*

$$g_n(x) := \sum_{k=0}^{n} \alpha_k \varphi_k(x) \tag{3.6}$$

gesetzt. Dann nimmt der **mittlere quadratische Fehler** *$F_n(\vec{\alpha})$ zwischen $f(x)$ und $g_n(x)$:*

$$\boxed{F_n(\vec{\alpha}) := \|f - g_n\|_2 = \Big(\int_a^{a+T} |f(x) - g_n(x)|^2 \, dx \Big)^{1/2}}$$

sein absolutes Minimum genau für die FOURIER*–Koeffizienten $\vec{\alpha}_0 := (a_0, a_1, \ldots, a_n)^T$ an. Darüber hinaus gilt die Relation*

$$0 \leq \frac{1}{\gamma} \cdot F_n^2(\vec{\alpha}_0) = \frac{1}{\gamma} \int_a^{a+T} |f(x)|^2 \, dx - \sum_{k=0}^{n} |a_k|^2 \ \ \forall\ n \in \mathbf{N}, \tag{3.7}$$

und daraus resultiert im Limes $n \to \infty$ die BESSEL**sche Ungleichung**

$$\boxed{\sum_{k=0}^{\infty} |a_k|^2 \leq \frac{1}{\gamma} \int_a^{a+T} |f(x)|^2 \, dx.} \tag{3.8}$$

Begründung: Es gilt unter Beachtung der Orthogonalitätsrelationen $\langle \varphi_j, \varphi_k \rangle_2 = \gamma \cdot \delta_{jk}$:

$$\begin{aligned} F_n^2(\vec{\alpha}) &= \langle f - g_n, f - g_n \rangle_2 = \langle f, f \rangle_2 - 2\,\mathrm{Re}\,\langle f, g_n \rangle_2 + \langle g_n, g_n \rangle_2 \\ &= \int_I |f(x)|^2\,dx - 2\gamma\,\mathrm{Re}\sum_{k=0}^{n} a_k \overline{\alpha_k} + \gamma \sum_{k=0}^{n} |\alpha_k|^2 = \int_I |f(x)|^2\,dx - 2\gamma\,\mathrm{Re}\,\langle \vec{\alpha}_0, \vec{\alpha} \rangle + \gamma\,\langle \vec{\alpha}, \vec{\alpha} \rangle. \end{aligned}$$

(Hier bezeichnet $\langle \cdot, \cdot \rangle$ das Standardskalarprodukt in $\mathbf{K}^{n+1}$.) Aus der obigen Relation folgt

$$F_n^2(\vec{\alpha}) - F_n^2(\vec{\alpha}_0) = \gamma\,\langle \vec{\alpha} - \vec{\alpha}_0, \vec{\alpha} - \vec{\alpha}_0 \rangle \geq 0 \quad \text{und} \quad = 0 \quad \text{genau für} \quad \vec{\alpha} = \vec{\alpha}_0.$$

Die Ungleichung (3.7) ergibt sich schließlich aus obiger Rechnung für $\vec{\alpha} = \vec{\alpha}_0$:

$$0 \leq \frac{1}{\gamma}\,F_n^2(\vec{\alpha}_0) = \frac{1}{\gamma}\int_I |f(x)|^2\,dx - \langle \vec{\alpha}_0, \vec{\alpha}_0 \rangle = \frac{1}{\gamma}\,\|f\|_2^2 - \sum_{k=0}^{n} |a_k|^2,$$

wie behauptet. □

Folgerung 17.7 *Vorgelegt sei das trigonometrische Orthogonalsystem*

$$\varphi_0(x) := \frac{1}{\sqrt{2}}, \quad \varphi_{2k}(x) := \cos(k\omega x), \quad \varphi_{2k-1}(x) := \sin(k\omega x), \quad k \in \mathbf{N}, \tag{3.9}$$

mit festem $\omega > 0$. In diesem Falle gilt

$$\boxed{\begin{aligned} &T = \frac{2\pi}{\omega}, \quad \gamma = \frac{T}{2}, \qquad f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos(k\omega x) + b_k \sin(k\omega x)\right), \\ &a_k = \frac{2}{T}\int_0^T f(x)\cos(k\omega x)\,dx, \quad b_k = \frac{2}{T}\int_0^T f(x)\sin(k\omega x)\,dx, \qquad k \in \mathbf{N}_0. \end{aligned}} \tag{3.10}$$

An die Stelle der Ungleichungen (3.7), (3.8) *treten hier:*

$$\boxed{0 \leq \frac{2}{T}\int_0^T |f(x)|^2\,dx - \frac{|a_0|^2}{2} - \sum_{k=1}^{n} \left(|a_k|^2 + |b_k|^2\right) \ \forall\, n \in \mathbf{N},} \tag{3.11}$$

bzw. die Bessel*–Ungleichung*

$$\boxed{\frac{|a_0|^2}{2} + \sum_{k=1}^{\infty} \left(|a_k|^2 + |b_k|^2\right) \leq \frac{2}{T}\,\|f\|_2^2 = \frac{2}{T}\int_0^T |f(x)|^2\,dx.} \tag{3.12}$$

Die Bessel–Ungleichung (3.12) besagt, dass die Reihe $\sum\limits_{k=1}^{\infty} \left(|a_k|^2 + |b_k|^2\right)$ der Quadrate der Fourier–Koeffizienten einer Funktion $f \in L^2(I)$ stets konvergiert. Deshalb muss **notwendig** gelten:

$$\lim_{k\to\infty} a_k = 0 = \lim_{k\to\infty} b_k, \quad \text{also} \quad a_k,\, b_k = \mathcal{O}(1) \ \ (k \to \infty).$$

Allgemeiner folgt daraus:

Satz 17.4 *Gegeben sei die T–periodische Funktion $f \in C^p(\mathbf{R})$ mit den* Fourier*–Koeffizienten* (3.10). *Dann gelten:*

$$a_k,\, b_k = \mathcal{O}\Big(\frac{1}{k^p}\Big) \ \ (k \to \infty), \quad \textit{also} \quad a_k = \frac{\alpha_k}{k^p},\ b_k = \frac{\beta_k}{k^p} \quad \textit{mit} \quad \alpha_k,\, \beta_k = \mathcal{O}(1) \ \ (k \to \infty).$$

Begründung: Mit $f \in C^p(\mathbf{R})$ folgt insbesondere $f^{(j)} \in L^2(I)$ für $I := [a, a+T]$ und für jedes $j = 0, 1, \ldots, p$. Seien $a_k[f]$, $b_k[f]$ bzw. $a_k[f^{(j)}]$, $b_k[f^{(j)}]$ die FOURIER–Koeffizienten der Funktionen $f(x)$ bzw. $f^{(j)}(x)$. Dann folgt durch partielle Integration für $k > 0$ und $T = \frac{2\pi}{\omega}$:

$$a_k[f] = \frac{2}{T} \int_0^T f(x) \cos(k\omega x)\, dx = -\frac{1}{k\omega} \int_0^T f'(x) \sin(k\omega x)\, dx = -\frac{1}{k\omega}\, b_k[f'].$$

Analog zeigt man $b_k[f] = \frac{1}{k\omega}\, a_k[f']$, und durch vollständige Induktion:

$$a_k[f] = \frac{\epsilon_p}{(k\omega)^p} \begin{cases} b_k[f^{(p)}] & : \ p \text{ ungerade}, \\ a_k[f^{(p)}] & : \ p \text{ gerade}, \end{cases} \qquad b_k[f] = \frac{\epsilon_p}{(k\omega)^p} \begin{cases} a_k[f^{(p)}] & : \ p \text{ ungerade}, \\ b_k[f^{(p)}] & : \ p \text{ gerade}, \end{cases}$$

wobei $\epsilon_p = \pm 1$ gilt. Wegen der Konvergenz der Reihe $\sum\limits_{k=1}^{\infty} \left(|a_k[f^{(p)}]|^2 + |b_k[f^{(p)}]|^2 \right)$ folgt dann schon die Behauptung. □

PARSEVAL–Gleichung.

Wegen (3.7) gilt $\lim\limits_{n\to\infty} F_n(\vec{\alpha}_0) = 0$ genau dann, wenn in der BESSEL–Ungleichung (3.8) das Gleichheitszeichen eintritt. In diesem Fall erhalten wir die PARSEVAL–Gleichung

$$\boxed{\sum_{k=0}^{\infty} |a_k|^2 = \frac{1}{\gamma} \int_a^{a+T} |f(x)|^2\, dx,} \tag{3.13}$$

beziehungsweise

$$\boxed{\frac{|a_0|^2}{2} + \sum_{k=1}^{\infty} \left(|a_k|^2 + |b_k|^2 \right) = \frac{2}{T} \int_0^T |f(x)|^2\, dx.} \tag{3.14}$$

Folgerung 17.8 *Genau dann gilt die* PARSEVAL*sche Gleichung für eine Funktion $f \in L^2(I)$, $I := [a, a+T]$, wenn die* FOURIER*–Reihe der Funktion $f(x)$* **im quadratischen Mittel** *gegen $f(x)$ konvergiert:*

$$\boxed{\lim_{n\to\infty} \int_a^{a+T} |f(x) - \sum_{k=0}^{n} a_k \varphi_k(x)|^2\, dx = 0.}$$

Definition 17.5 *Ein Orthogonalsystem $(\varphi_k(x))_{k\geq 0}$ auf dem Intervall $I := [a, a+T]$ heiße* **vollständig**, *wenn die* PARSEVAL*–Gleichung* (3.13) *bzw.* (3.14) *für* **jedes** *$f \in L^2(I)$ gilt. Die Gleichung* (3.13) *heißt häufig auch* **Vollständigkeitsrelation**.

Folgerung 17.9 *In einem* **vollständigen** *Orthogonalsystem $(\varphi_k(x))_{k\geq 0}$ konvergieren für jedes $f \in L^2(I)$ die* FOURIER*–Reihen im quadratischen Mittel gegen $f(x)$.*

Es gehört zu den (schwierigen) Aufgaben der Funktionalanalysis, die Vollständigkeit eines Orthogonalsystems zu zeigen. Für das trigonometrische Funktionensystem (3.9) kennt man allerdings die Vollständigkeit (→ Satz von FISCHER–RIESZ, Literatur).

17.4 Punktweise Konvergenz der trigonometrischen FOURIER–Reihen

Die im vorangegangenen Abschnitt 17.3 gezeigte Konvergenz der trigonometrischen FOURIER–Reihen im quadratischen Mittel gegen eine T–periodische Grenzfunktion $f \in L^2(I)$, $I := [a, a+T]$, ist für viele Anwendungen unbefriedigend, da zum Beispiel die FOURIER–Reihe der DIRICHLET–Funktion $f \in L^2(I)$ in **keinem rationalen** Punkt $x \in I$ den Funktionswert $f(x)$ liefert. Wir suchen nach Kriterien, mit deren Hilfe es einfach festzustellen ist, in welchen Punkten $x \in I$ die Relation "$\sim$" durch das Gleichheitszeichen "$=$" ersetzt werden kann. Wir gehen ohne Beschränkung der Allgemeinheit von folgenden Voraussetzungen aus:

$$\boxed{\begin{array}{c} T := 2\pi, \quad I := [-\pi, +\pi], \quad \text{und für } f \in L^2(I) \text{ gelte:} \\ g(x) := \dfrac{a_0}{2} + \displaystyle\sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx). \end{array}} \tag{4.1}$$

$$\boxed{a_k := \frac{1}{\pi} \int\limits_{-\pi}^{\pi} f(x) \cos kx \, dx, \quad b_k := \frac{1}{\pi} \int\limits_{-\pi}^{\pi} f(x) \sin kx \, dx, \quad k \in \mathbf{N}_0.} \tag{4.2}$$

Bei den weiteren Betrachtungen beschränken wir uns ausschließlich auf Funktionen $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K})$, die der folgenden Klassifizierung genügen:

Definition 17.6 (der Bedingung (F))
Eine Funktion $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K})$ erfülle die Bedingung (F) *genau dann, wenn gilt:*

(F1) $f(x + 2\pi) = f(x)\ \forall\ x \in \mathbf{R}$ *(2π-Periodizität) und* $f \in L^2(-\pi, +\pi)$.

(F2) $f(x)$ *hat in jedem Punkte* $x \in \mathbf{R}$ *einen rechts– und linksseitigen Grenzwert:*

$$f(x \pm 0) = \lim_{\epsilon \to 0+} f(x \pm \epsilon) \quad \textit{existiert} \quad \forall\ x \in \mathbf{R}.$$

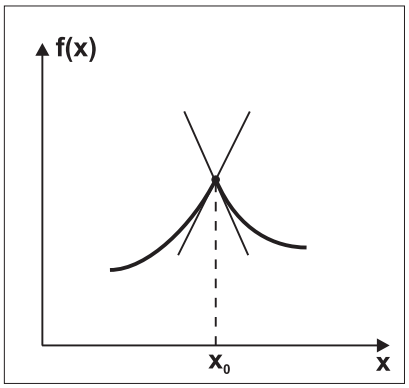
(F3) $f(x)$ *ist in jedem Punkte* $x \in \mathbf{R}$ *rechts– und linksseitig* LIPSCHITZ*–stetig:*

$$\forall\ x \in \mathbf{R}\ \ \exists\ L > 0\ \exists\ \epsilon_0 > 0\ :\quad \frac{1}{\epsilon} \cdot |f(x \pm \epsilon) - f(x \pm 0)| \le L\ \ \forall\ 0 < \epsilon \le \epsilon_0.$$

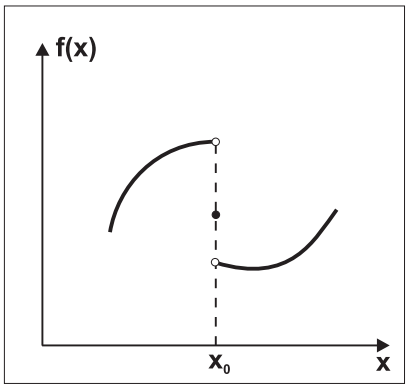
BSP. (17.4.1) Die folgenden Funktionen genügen der Bedingung (F):

- $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K})$ ist stetig, und $f' \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K})$ ist stetig mit höchstens endlich vielen Ausnahmepunkten in jedem Periodenintervall $[a, a + 2\pi]$, in denen $f'(x)$ Sprungstellen endlicher Höhe hat.

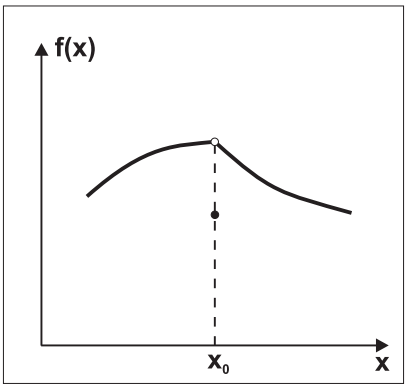
- $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K})$ ist stetig mit Ausnahme höchstens endlich vieler Sprungstellen endlicher Höhe in jedem Periodenintervall $[a, a+2\pi]$. In den Sprungstellen existieren endliche links– und rechtsseitige Ableitungen. Ansonsten ist $f'(x)$ überall stetig; *zum Beispiel:*



Knickpunkt



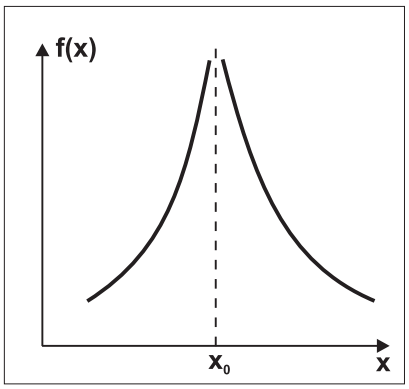
Sprungstelle



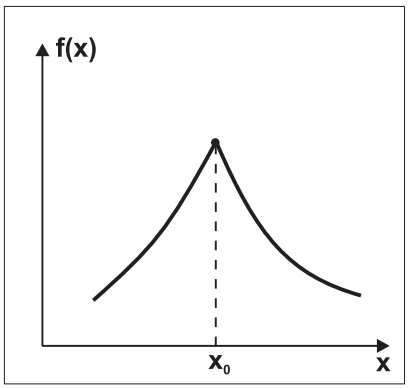
Hebbare Unstetigkeit

BSP. (17.4.2) Die folgenden Funktionen erfüllen **nicht** die Bedingung (F):

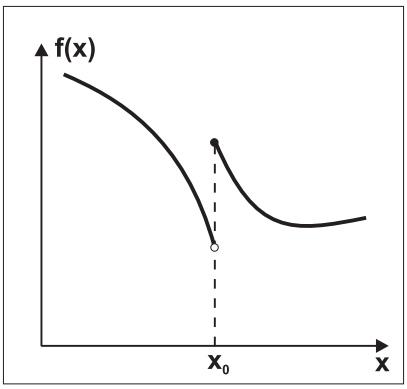
- $f(x) := \sin \frac{1}{x},\ x \in [-\pi, +\pi]$ und $f(x+2\pi) = f(x)\ \forall\, x \in \mathbf{R}$. Die Grenzwerte $f(x \pm 0)$ existieren in den Punkten $x_k := 2\pi k,\ k \in \mathbf{Z}$, nicht.
- $f(x) := \sqrt{|x|},\ x \in [-\pi, +\pi]$ und $f(x + 2\pi) = f(x)\ \forall\, x \in \mathbf{R}$. Der Graph von f hat in den Punkten $x_k := 2\pi k,\ k \in \mathbf{Z}$, eine vertikale Tangente.
- Andere Beispiele von Funktionen mit vertikaler Tangente:



Polstelle



Spitze



Sprungstelle

Bemerkung 17.4 Erfüllt die Funktion $f(x)$ die Bedingung (F), so folgt aus (F3), dass die Funktionen

$$g_{\pm}(x_0; t) := \frac{1}{t} \cdot (f(x_0 \pm t) - f(x_0 \pm 0)) \tag{4.3}$$

für jedes feste $x_0 \in \mathbf{R}$ auf dem Intervall $I := [-\pi, +\pi]$ quadratintegrabel sind, also zur Funktionenklasse $L^2(I)$ gehören. Dies wird uns in der weiteren Analyse von Nutzen sein. □

Zur Untersuchung der Konvergenzfrage der FOURIER–Reihe benötigen wir einige Vorbetrachtungen.

Satz 17.5 *Für $x \neq 2\pi j,\ j \in \mathbf{Z}$, gilt die Identität*

$$\boxed{\frac{1}{2} + \sum_{k=1}^{n} \cos kx = \frac{\sin\left((2n+1) \cdot \frac{x}{2}\right)}{2 \cdot \sin \frac{x}{2}}, \qquad x \in \mathbf{R}.} \tag{4.4}$$

Begründung: Für $k \geq 1$ haben wir $\cos kx \cdot \sin \frac{x}{2} = \frac{1}{2}\Big(\sin(k+\frac{1}{2})x - \sin(k-\frac{1}{2})x\Big)$. Summation von $k=1$ bis $k=n$ liefert

$$\sin\frac{x}{2} \cdot \sum_{k=1}^{n} \cos kx = \frac{1}{2} \cdot \Big(\sin(n+\frac{1}{2})x - \sin\frac{x}{2}\Big).$$

Division durch $\sin\frac{x}{2}$ führt auf die Beziehung (4.4). □

Die 2π–periodische Funktion $f \in \mathrm{Abb}(\mathbf{R},\mathbf{K})$ sei so gegeben, dass $f(x)$ über dem Intervall $[-\pi,+\pi]$ integrabel ist. Dann existiert das FOURIER–Polynom

$$g_n(x) := \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx), \;\; n \in \mathbf{N}. \tag{4.5}$$

Für $g_n(x)$ gilt folgende Darstellungsformel:

Satz 17.6 (vom DIRICHLET–Integral)
Ist die 2π–periodische Funktion $f \in \mathrm{Abb}(\mathbf{R},\mathbf{K})$ integrierbar über ein Periodenintervall der Länge 2π, so gestattet $g_n(x)$ die Darstellung

$$\boxed{g_n(x) = \frac{2}{\pi} \int_0^{\pi/2} \frac{1}{2} \cdot (f(x+2t) + f(x-2t)) \, \frac{\sin(2n+1)t}{\sin t}\, dt \;\; \forall\, x \in \mathbf{R}.} \tag{4.6}$$

Begründung: Es sei $x \in \mathbf{R}$ fest gewählt. Aus den Formeln für die FOURIER–Koeffizienten resultiert:

$$\begin{aligned}
g_n(x) &= \frac{1}{2\pi}\int_a^{a+2\pi} f(t)\,dt + \sum_{k=1}^{n} \frac{1}{\pi}\int_a^{a+2\pi} f(t)(\cos kx\cos kt + \sin kx \sin kt)\,dt \\
&= \frac{1}{\pi}\int_a^{a+2\pi} f(t)\Big(\frac{1}{2} + \sum_{k=1}^{n}\cos k(t-x)\Big)\,dt \\
&\overset{(4.4)}{=} \frac{1}{2\pi}\int_a^{a+2\pi} f(t)\cdot\frac{\sin\left((2n+1)\frac{t-x}{2}\right)}{\sin(\frac{t-x}{2})}\,dt \\
&\overset{s:=t-x}{=} \frac{1}{2\pi}\int_{a-x}^{a-x+2\pi} f(s+x)\,\frac{\sin\left((2n+1)\frac{s}{2}\right)}{\sin\frac{s}{2}}\,ds \\
&\overset{a:=x-\pi}{=} \frac{1}{2\pi}\Big\{\int_0^{\pi}\cdots ds + \int_{-\pi}^{0}\cdots ds\Big\}.
\end{aligned}$$

Im ersten Integral der letzten Zeile setzen wir $t := \frac{s}{2}$, im zweiten hingegen $t := -\frac{s}{2}$, und erhalten auf diese Weise:

$$g_n(x) = \frac{2}{\pi}\int_0^{\pi/2} \frac{1}{2}\cdot(f(x+2t)+f(x-2t))\,\frac{\sin(2n+1)t}{\sin t}\,dt,$$

wobei wegen

$$\lim_{t\to 0}\frac{\sin(2n+1)t}{\sin t} = 2n+1$$

die Existenz des Integrals auch bei $t=0$ gesichert ist. □

Wir setzen nun

$$K_n(t) := \begin{cases} \dfrac{\sin(2n+1)t}{\sin t} & : \; t \in [-\frac{\pi}{2},+\frac{\pi}{2}],\;\; t \neq 0, \\ 2n+1 & : \; t = 0. \end{cases} \tag{4.7}$$

Die spezielle Funktion $f(x) \equiv 1$ hat als 2π–periodische Funktion die FOURIER–Koeffizienten $a_0 = 2$, $a_k = b_k = 0,\ k \geq 1$. Also liefert Satz 17.6 die Identität:

$$\boxed{1 = \frac{2}{\pi} \int\limits_0^{\pi/2} \frac{\sin(2n+1)t}{\sin t}\, dt = \frac{2}{\pi} \int\limits_0^{\pi/2} K_n(t)\, dt \ \ \forall\, n \in \mathbf{N}.} \tag{4.8}$$

Wir sind jetzt in der Lage, den **Hauptsatz** über die punktweise Konvergenz einer FOURIER–Reihe zu formulieren:

Satz 17.7 *Die Funktion $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{K})$ genüge der Bedingung* (F)*. Dann gilt für die zugehörige* FOURIER*–Reihe $g(x)$ aus* (4.1) *in jedem Punkte $x = x_0 \in \mathbf{R}$:*

$$\boxed{\frac{1}{2}\left(f(x_0+0) + f(x_0-0)\right) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos kx_0 + b_k \sin kx_0\right) \sim f(x_0).} \tag{4.9}$$

Das heißt, die FOURIER*–Reihe $g(x)$ konvergiert im Punkte $x = x_0$ gegen das arithmetische Mittel von rechts– und linksseitigem Funktionenlimes $f(x_0 \pm 0)$. Ist $f(x)$ stetig in $x = x_0$, so konvergiert $g(x)$ im Punkte $x = x_0$ gegen $f(x_0)$.*

Begründung: Sei $x_0 \in \mathbf{R}$ fest gewählt. Aus den Beziehungen (4.6) und (4.8) resultiert:

$$\begin{aligned}
\Delta_n \quad &:= \quad g_n(x_0) - \frac{1}{2} \cdot \left(f(x_0+0) + f(x_0-0)\right) \cdot 1 \\
&= \quad \frac{2}{\pi} \int_0^{\pi/2} \frac{1}{2} \Big(\left(f(x_0+2t) - f(x_0+0)\right) + \left(f(x_0-2t) - f(x_0-0)\right) \Big) K_n(t)\, dt \\
&\overset{(4.3)}{=} \quad \frac{2}{\pi} \int_0^{\pi/2} \left(g_+(x_0;2t) + g_-(x_0;2t)\right) \frac{t}{\sin t} \left(\sin(2nt)\cos t + \cos(2nt)\sin t\right) dt.
\end{aligned}$$

Man sieht sofort, dass die Funktion

$$U(t) := \left(g_+(x_0;2t) + g_-(x_0;2t)\right) \frac{t\cos t}{\sin t} = \frac{f(x_0+2t) - f(x_0+0) + f(x_0-2t) - f(x_0-0)}{2 \cdot \sin t} \cos t$$

ungerade ist. Sie ist quadratintegrabel über das Periodenintervall $[-\frac{\pi}{2}, +\frac{\pi}{2}]$. Ferner ist $U(t)$ π–periodisch. Gemäß Satz 17.4 gilt daher:

$$U(t) \sim \sum_{k=1}^{\infty} b_k(U) \sin(2kt) \quad \text{mit} \quad b_k(U) = \frac{4}{\pi} \int_0^{\pi/2} U(t) \sin(2kt)\, dt = \mathcal{O}(1) \ \ (k \to \infty).$$

Ganz analog ergründet man, dass die Funktion

$$G(t) := \left(g_+(x_0;2t) + g_-(x_0;2t)\right) \cdot t = \frac{1}{2}\left(f(x_0+2t) - f(x_0+0) + f(x_0-2t) - f(x_0-0)\right)$$

gerade ist; sie ist π–periodisch und quadratintegrabel über das Periodenintervall $[-\frac{\pi}{2}, +\frac{\pi}{2}]$. Also gilt ebenfalls gemäß Satz 17.4:

$$G(t) \sim \frac{a_0(G)}{2} + \sum_{k=1}^{\infty} a_k(G) \cos(2kt) \quad \text{mit} \quad a_k(G) = \frac{4}{\pi} \int_0^{\pi/2} G(t) \cos(2kt)\, dt = \mathcal{O}(1) \ \ (k \to \infty).$$

Wir haben hiermit

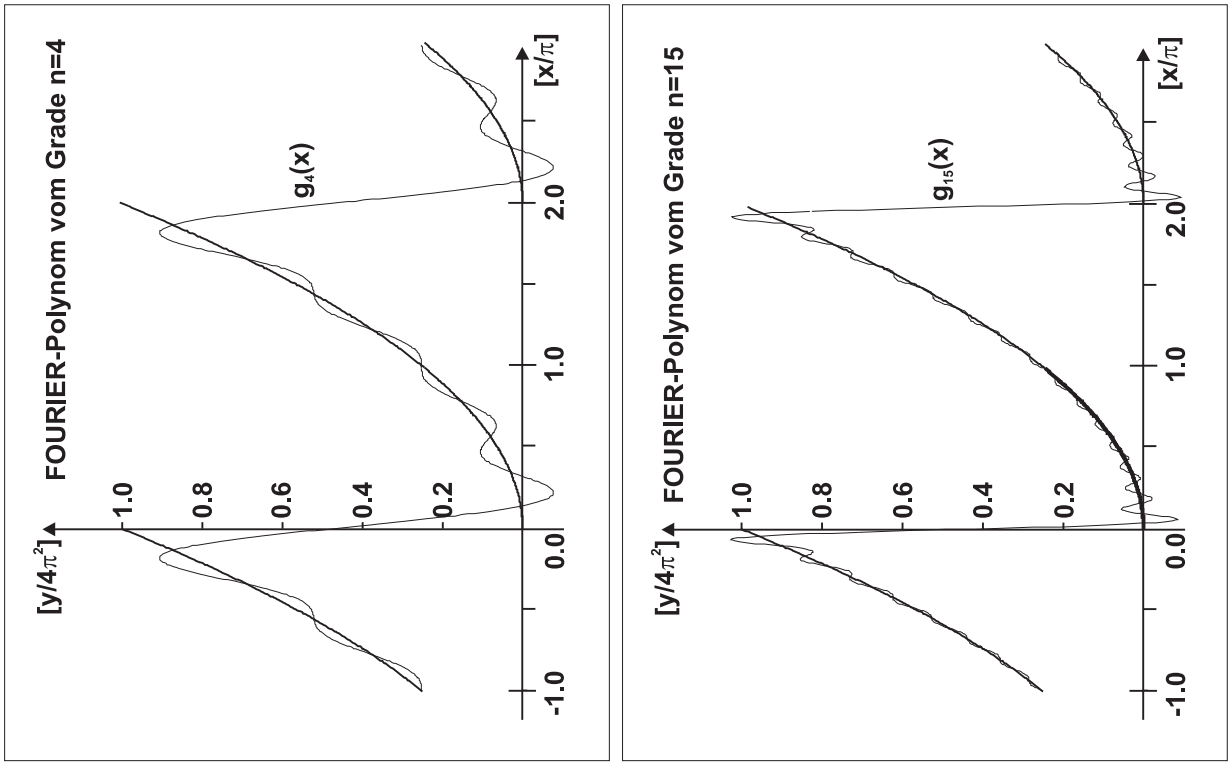
$$\Delta_n = \frac{1}{2}\Big(a_n(G) + b_n(U)\Big) \to 0 \ \ (n \to \infty),$$

und das ist die behauptete punktweise Konvergenz. □

Bemerkung 17.5 Funktionen $f(x)$ mit der Bedingung (F) sind sicher auf dem Periodenintervall $[-\pi, +\pi]$ beschränkt, da sonst die Funktionenlimites $f(x \pm 0)$ nicht überall existieren würden. Gemäß (4.9) ist dann auch die FOURIER–Reihe $g(x)$ beschränkt. Man kann ferner zeigen, dass die Gleichung $f(x) = g(x)$ für fast alle $x \in \mathbf{R}$ gilt. Unter diesen Eigenschaften hat man nach Satz 17.3:

$$\lim_{n\to\infty} F_n^2(\vec{\alpha}_0) = \lim_{n\to\infty} \int_{-\pi}^{\pi} |f(x) - g_n(x)|^2 \, dx = \int_{-\pi}^{\pi} \lim_{n\to\infty} |f(x) - g_n(x)|^2 \, dx = \int_{-\pi}^{\pi} 0 \, dx = 0.$$

Die hier durchgeführte Vertauschung von "lim" und "$\int$" ist auf Grund des LEBESGUEschen Satzes 8.37 von der dominierten Konvergenz gestattet. Damit wird offenkundig, dass Funktionen mit der Bedingung (F) die PARSEVAL–Gleichung erfüllen, so dass die Konvergenz der FOURIER–Reihe $g(x)$ im quadratischen Mittel manifestiert ist. □



BSP. (17.4.3) Es sei $f \in \mathrm{Abb}(\mathbf{R}, \mathbf{R})$ in der folgenden Weise definiert:

$$f(x) := \begin{cases} x^2 & : \ 0 \le x < 2\pi, \\ f(x - 2\pi) & : \ x \quad \text{sonst.} \end{cases}$$

Offensichtlich erfüllt $f(x)$ die Bedingung (F), da nur in den Punkten $x_j := 2\pi j, \ j \in \mathbf{Z}$, Sprungstellen mit endlichen (rechts– und linksseitigen) Ableitungen auftreten. Die FOURIER–Koeffizienten gewinnt man mit elementarer Integration:

$$a_k = \frac{1}{\pi} \int_0^{2\pi} x^2 \cos kx \, dx = \begin{cases} \frac{8}{3}\pi^2 & : \ k = 0, \\ \frac{4}{k^2} & : \ k \in \mathbf{N}, \end{cases}$$

$$b_k = \frac{1}{\pi} \int_0^{2\pi} x^2 \sin kx \, dx = -\frac{4\pi}{k}, \quad k \in \mathbf{N}.$$

Hieraus resultiert die FOURIER–Reihe

$$g(x) = \frac{4\pi^2}{3} + 4 \sum_{k=1}^{\infty} \left(\frac{\cos kx}{k^2} - \frac{\pi \cdot \sin kx}{k} \right) = \begin{cases} f(x) & : \ x \ne 2\pi j, \quad j \in \mathbf{Z}, \\ 2\pi^2 & : \ x = 2\pi j, \quad j \in \mathbf{Z}. \end{cases}$$

Die obigen Skizzen zeigen die Approximation von $f(x)$ durch die FOURIER–Polynome $g_4(x)$ und $g_{15}(x)$. Das Auftreten der "Zipfel" ($\rightarrow$ GIBBSsches Phänomen) in den Sprungstellen $x_j := 2\pi j$ ist typisch. Die Zipfel verschwinden keinesfalls, wenn man mit $g_n(x)$ für $n \gg 1$ approximiert.

17.5 Die diskrete FOURIER–Transformation

Ist die 2π–periodische Funktion $f(x)$ nur numerisch vorgegeben oder ist $f(x)$ nicht elementar, so lassen sich in der Regel die Integrale (4.2) für die FOURIER–Koeffizienten a_k, b_k von $f(x)$ nicht mehr analytisch berechnen. In diesem Falle ist es angezeigt, numerische Integration mit Hilfe geeigneter **Quadraturformeln** durchzuführen. Die einfachste und für die vorliegende Problemstellung angemessenste Quadraturformel ist die **verallgemeinerte Trapezregel**. Das Integral $\int_a^b F(x)\,dx$ wird bei äquidistanter Zerlegung des Intervalls $I := [a,b]$ mit $h := (b-a)/N,\ x_j := a + jh,\ j = 0,1,\ldots,N \in \mathbf{N}$, sehr gut durch den folgenden Näherungswert beschrieben:

$$\boxed{\int_a^b F(x)\,dx \approx \frac{h}{2}\left(F(x_0) + 2\sum_{j=1}^{N-1} F(x_j) + F(x_N)\right) \equiv T_N(F).} \tag{5.1}$$

Die Relation (5.1) heißt **verallgemeinerte Trapezregel**.

Gegeben sei nun die 2π–periodische Funktion $f \in \mathrm{Abb}\,(\mathbf{R},\mathbf{C})$. Wir wählen $I := [0,2\pi]$ als Periodenintervall und setzen:

$$\boxed{h := \frac{2\pi}{N}, \quad x_j := jh = \frac{2\pi}{N}\,j, \quad f_j := \frac{1}{2}\,(f(x_j+0) + f(x_j-0)), \quad j = 0,1,\ldots,N \in \mathbf{N},} \tag{5.2}$$

$$\boxed{a_k^* := \frac{2}{N}\sum_{j=1}^{N} f_j\,\cos kx_j, \quad b_k^* := \frac{2}{N}\sum_{j=1}^{N} f_j\,\sin kx_j, \quad k \in \mathbf{N}_0.} \tag{5.3}$$

Mit $C_k(x) := f(x)\cos kx$ bzw. $S_k(x) := f(x)\sin kx$ hat man $C_k(x_0) = C_k(x_N)$ sowie $S_k(x_0) = S_k(x_N)$, so dass

$$a_k^* = \frac{1}{\pi}\,T_N(C_k), \qquad b_k^* = \frac{1}{\pi}\,T_N(S_k)\ \ \forall\, k \in \mathbf{N}_0$$

gilt. Das heißt, die Koeffizienten a_k^*, b_k^* sind **Näherungswerte** der FOURIER–Koeffizienten a_k, b_k (4.2) von $f(x)$. Wir zeigen, dass man mit Hilfe des trigonometrischen Polynoms

$$g_m^*(x) := \frac{a_0^*}{2} + \sum_{k=1}^{m}\,(a_k^*\cos kx + b_k^*\sin kx) \tag{5.4}$$

ein *Approximationsproblem im diskreten quadratischen Mittel* lösen kann. Wir benötigen dazu einen Orthogonalitätsbegriff und führen deshalb auf dem Raum der **Gitterfunktionen**

$$\Pi^h := \left\{f^h\ :\ f^h := (f(x_1), f(x_2), \ldots, f(x_N))\right\}$$

ein Skalarprodukt ein:

$$\langle f^h, g^h\rangle_N := \sum_{j=1}^{N} f(x_j)\overline{g(x_j)}\ \ \forall\, f^h,\, g^h \in \Pi^h. \tag{5.5}$$

Wir zeigen zunächst ein Hilfsresultat:

Satz 17.8 *Auf den Gitterpunkten x_j aus (5.2) gilt:*

$$\boxed{\sum_{j=1}^{N} \sin kx_j = 0 \ \ \forall\, k \in \mathbf{Z}, \qquad \sum_{j=1}^{N} \cos kx_j = \begin{cases} 0 & : \ \frac{k}{N} \notin \mathbf{Z}, \\ N & : \ \frac{k}{N} \in \mathbf{Z}. \end{cases}} \tag{5.6}$$

Begründung: Die komplexe Summe

$$S_k := \sum_{j=1}^{N} (\cos kx_j + i \sin kx_j) = \sum_{j=1}^{N} (e^{2\pi ik/N})^j = \begin{cases} N & : \ \frac{k}{N} \in \mathbf{Z}, \\ e^{ikh}\, \dfrac{e^{2\pi ik} - 1}{e^{ikh} - 1} = 0 & : \ \frac{k}{N} \notin \mathbf{Z}, \end{cases}$$

liefert durch Zerlegung in Real– und Imaginärteil bereits die Behauptung. □

Wir zeigen jetzt mit Hilfe dieses Satzes die Gültigkeit **diskreter Orthogonalitätsrelationen**.

Satz 17.9 *Auf den Gitterpunkten x_j aus (5.2) gelten folgende diskrete Orthogonalitätsrelationen:*

$$\sum_{j=1}^{N} \cos kx_j \cdot \sin lx_j \ = \ 0 \ \ \forall\, k, l \in \mathbf{N}_0 \tag{5.7}$$

$$\sum_{j=1}^{N} \cos kx_j \cdot \cos lx_j \ = \ \begin{cases} 0, & \text{falls } \frac{k+l}{N} \notin \mathbf{Z} \text{ und } \frac{k-l}{N} \notin \mathbf{Z}, \\ \frac{N}{2}, & \text{falls entweder } \frac{k+l}{N} \in \mathbf{Z} \text{ oder } \frac{k-l}{N} \in \mathbf{Z}, \\ N, & \text{falls } \frac{k+l}{N} \in \mathbf{Z} \text{ und } \frac{k-l}{N} \in \mathbf{Z}, \end{cases} \tag{5.8}$$

$$\sum_{j=1}^{N} \sin kx_j \cdot \sin lx_j \ = \ \begin{cases} 0, & \text{falls } \frac{k+l}{N} \begin{cases} \notin \mathbf{Z} \\ \in \mathbf{Z} \end{cases} \text{ und } \frac{k-l}{N} \begin{cases} \notin \mathbf{Z}, \\ \in \mathbf{Z}, \end{cases} \\ -\frac{N}{2}, & \text{falls } \frac{k+l}{N} \in \mathbf{Z} \text{ und } \frac{k-l}{N} \notin \mathbf{Z}, \\ \frac{N}{2}, & \text{falls } \frac{k+l}{N} \notin \mathbf{Z} \text{ und } \frac{k-l}{N} \in \mathbf{Z}. \end{cases} \tag{5.9}$$

Begründung: Man verwende die Identitäten

$$\begin{aligned} \cos kx_j \cdot \sin lx_j &= \frac{1}{2} \cdot (\sin(k+l)x_j - \sin(k-l)x_j), \\ \cos kx_j \cdot \cos lx_j &= \frac{1}{2} \cdot (\cos(k+l)x_j + \cos(k-l)x_j), \\ \sin kx_j \cdot \sin lx_j &= \frac{1}{2} \cdot (\cos(k+l)x_j - \cos(k-l)x_j), \end{aligned}$$

zusammen mit den Relationen (5.6) aus Satz 17.8. □

Wir suchen jetzt die Lösung des folgenden **Interpolationsproblems**:

$$\boxed{\begin{array}{l} \text{Gegeben seien die äquidistanten Stützstellen } x_j \text{ aus (5.2) und dazu Stützwerte } f_j := \\ \frac{1}{2}\,(f(x_j{+}0){+}f(x_j{-}0)),\ j = 0, 1, \ldots, N := 2n,\ n \in \mathbf{N}. \text{ Gesucht ist ein trigonometrisches} \\ \text{Polynom } g_n(x) \text{ mit } g_n(x_j) = f_j \ \ \forall\, j = 0, 1, \ldots, N. \end{array}} \tag{5.10}$$

Satz 17.10 *Sei $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{C})$ eine 2π–periodische Funktion. Dann ist die eindeutig bestimmte Lösung des Interpolationsproblems (5.10) gegeben durch das spezielle trigonometrische Polynom*

$$\boxed{g_n^*(x) := \frac{a_0^*}{2} + \sum_{k=1}^{n-1} (a_k^* \cos kx + b_k^* \sin kx) + \frac{a_n^*}{2} \cos nx,} \tag{5.11}$$

worin die Koeffizienten a_k^*, b_k^* gemäß (5.3) definiert sind.

Begründung: Den N Interpolationsbedingungen $g_n(x_j) = f_j$ wird ein trigonometrisches Polynom mit entsprechend vielen Koeffizienten in der Form (beachte: $f_0 = f_N$)

$$g_n(x) = \frac{\alpha_0}{2} + \sum_{k=1}^{n-1} (\alpha_k \cos kx + \beta_k \sin kx) + \frac{\alpha_n}{2} \cos nx \tag{5.12}$$

gerecht. Das inhomogene lineare Gleichungssystem

$$\frac{\alpha_0}{2} + \sum_{k=1}^{n-1} (\alpha_k \cos kx_j + \beta_k \sin kx_j) + \frac{\alpha_n}{2} \cos nx_j = f_j, \quad j = 1, 2, \ldots, N,$$

ist eindeutig lösbar, da die Spaltenvektoren der Koeffizientenmatrix nicht verschwinden und wegen (5.6)–(5.9) paarweise orthogonal sind. Wir setzen jetzt

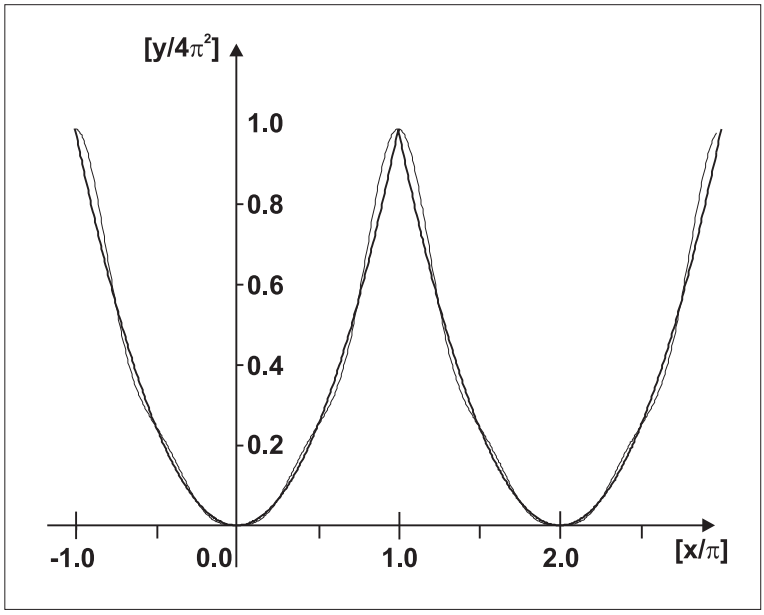
$$c_l(x) := \cos lx, \qquad s_l(x) := \sin lx, \quad l \in \mathbf{Z}.$$

Dann folgt aus dem Ansatz (5.12) zusammen mit (5.6)–(5.9):

$$\frac{N}{2} \cdot \alpha_l = \langle g_n^h, c_l^h \rangle_N = \langle f^h, c_l^h \rangle_N = \sum_{j=1}^{N} f_j \cos lx_j = \frac{N}{2} \cdot a_l^*, \quad l = 0, 1, \ldots, n,$$

$$\frac{N}{2} \cdot \beta_l = \langle g_n^h, s_l^h \rangle_N = \langle f^h, s_l^h \rangle_N = \sum_{j=1}^{N} f_j \sin lx_j = \frac{N}{2} \cdot b_l^*, \quad l = 1, 2, \ldots, n-1.$$

Also gilt $a_l^* = \alpha_l$ und $b_l^* = \beta_l$, so dass (5.11) resultiert. □



Interpolierendes trigonometrisches Polynom

BSP. (17.5.1) Es sei $f \in \mathrm{Abb}(\mathbf{R}, \mathbf{R})$ definiert durch

$$f(x) := x^2 \quad \text{für} \quad -\pi \le x \le +\pi \quad \text{und} \quad f(x + 2\pi) := f(x) \ \ \forall\, x \in \mathbf{R}.$$

Für $N := 8$ gilt $h := \frac{\pi}{4}$ und $x_j := \frac{\pi}{4} \cdot j$, $j = 0, 1, \ldots, 8$. Da $f(x)$ eine gerade Funktion ist, gilt $b_k^* = 0$ für die hier relevanten Werte $k = 1, 2, 3$. Die numerisch berechneten Koeffizienten a_k^* aus (5.3) sind in folgender Tabelle aufgelistet:

k	0	1	2	3	4
a_k^*	6.785 353	−4.212 117	1.233 701	−0.722 685	0.616 850

Insbesondere gilt $a_0^* = \frac{11\pi^2}{16}$. Die obige Grafik zeigt die beiden Graphen der Funktionen $f(x)$ und

$$g_4^*(x) := \frac{a_0^*}{2} + \sum_{k=1}^{3} a_k^* \cos kx + \frac{a_4^*}{2} \cos 4x.$$

Definiert man den **diskreten mittleren quadratischen Fehler** $F_n^h(\vec{\alpha})$ zwischen der 2π–periodischen Funktion $f \in \mathrm{Abb}(\mathbf{R}, \mathbf{C})$ und dem trigonometrischen Polynom $g_n(x) := \frac{\alpha_0}{2} + \sum\limits_{k=1}^{N} (\alpha_k \cos kx + \beta_k \sin kx)$ gemäß

$$\boxed{F_n^h(\vec{\alpha}) := \Big(\sum_{j=1}^{N} |f_j - g_n(x_j)|^2 \Big)^{1/2}, \quad \vec{\alpha} := (\alpha_0, \alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n)^T \in \mathbf{C}^{2n+1},} \tag{5.13}$$

so besagt Satz 17.10, dass das interpolierende trigonometrische Polynom (5.11) den Fehler F_n^h genau zu Null macht. In diesem Zusammenhang ist es sinnvoll, nach demjenigen trigonometrischen Polynom $g_m(x)$ vom Grade $m < n$ zu fragen, welches den Fehler $F_m^h(\vec{\alpha})$ minimiert. Als Antwort auf diese Frage erhalten wir:

Satz 17.11 *Die äquidistanten Stützstellen x_j, $j = 0, 1, \ldots, N := 2n$, $n \in \mathbf{N}$, seien wie in (5.2) erklärt. Für die 2π–periodische Funktion $f \in \mathrm{Abb}(\mathbf{R}, \mathbf{C})$ seien a_k^* und b_k^* gemäß (5.3) definiert. Ferner gelte für $\vec{\alpha} := (\alpha_0, \alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m)^T \in \mathbf{C}^{2m+1}$:*

$$g_m(x) := \frac{\alpha_0}{2} + \sum_{k=1}^{m} (\alpha_k \cos kx + \beta_k \sin kx). \tag{5.14}$$

Ist $m < n$, so nimmt der diskrete mittlere quadratische Fehler $F_m^h(\vec{\alpha})$ aus (5.13) sein absolutes Minimum genau für die diskreten FOURIER*–Koeffizienten $\vec{\alpha}^* := (a_0^*, a_1^*, \ldots, a_m^*, b_1^*, \ldots, b_m^*)^T \in \mathbf{C}^{2m+1}$ an. Das heißt, das trigonometrische Polynom*

$$\boxed{g_m^*(x) := \frac{a_0^*}{2} + \sum_{k=1}^{m} (a_k^* \cos kx + b_k^* \sin kx)} \tag{5.15}$$

löst das Ausgleichsproblem

$$F_m^h(\vec{\alpha}) = \Big(\sum_{j=1}^{N} |f_j - g_m(x_j)|^2 \Big)^{1/2} \stackrel{!}{=} \mathrm{Min}, \qquad \vec{\alpha} \in \mathbf{C}^{2m+1}.$$

Begründung: Mit Hilfe des Skalarproduktes (5.5) erschließt man aus den Orthogonalitätsrelationen (5.6)–(5.9):

$$\begin{aligned}
(F_m^h(\vec{\alpha}))^2 &= \langle f^h - g_m^h, f^h - g_m^h \rangle_N = \langle f^h, f^h \rangle_N - 2 \cdot \mathrm{Re}\, \langle f^h, g_m^h \rangle_N + \langle g_m^h, g_m^h \rangle_N \\
&= \sum_{j=1}^{N} |f_j|^2 - N \cdot \mathrm{Re}\, \Big(\frac{a_0^* \overline{\alpha_0}}{2} + \sum_{k=1}^{m} (a_k^* \overline{\alpha_k} + b_k^* \overline{\beta_k}) \Big) + \frac{N}{2} \Big(\frac{|\alpha_0|^2}{2} + \sum_{k=1}^{m} (|\alpha_k|^2 + |\beta_k|^2) \Big) \\
&= \sum_{j=1}^{N} |f_j|^2 - \frac{N}{2} \Big(2 \cdot \mathrm{Re}\, \langle \vec{\alpha}^*, \vec{\alpha} \rangle - \langle \vec{\alpha}, \vec{\alpha} \rangle \Big).
\end{aligned}$$

Hier haben wir im Vektorraum $\mathbf{C}^{2m+1}$ das Skalarprodukt $\langle \vec{\xi}, \vec{\eta} \rangle := \frac{1}{2} \xi_0 \overline{\eta_0} + \sum\limits_{k=1}^{2m} \xi_k \overline{\eta_k}$ verwendet. Aus der obigen Rechnung folgt:

$$(F_m^h(\vec{\alpha}))^2 - (F_m^h(\vec{\alpha}^*))^2 = \frac{N}{2} \langle \vec{\alpha} - \vec{\alpha}^*, \vec{\alpha} - \vec{\alpha}^* \rangle \geq 0,$$

wobei $= 0$ genau für $\vec{\alpha} = \vec{\alpha}^*$ eintritt. □

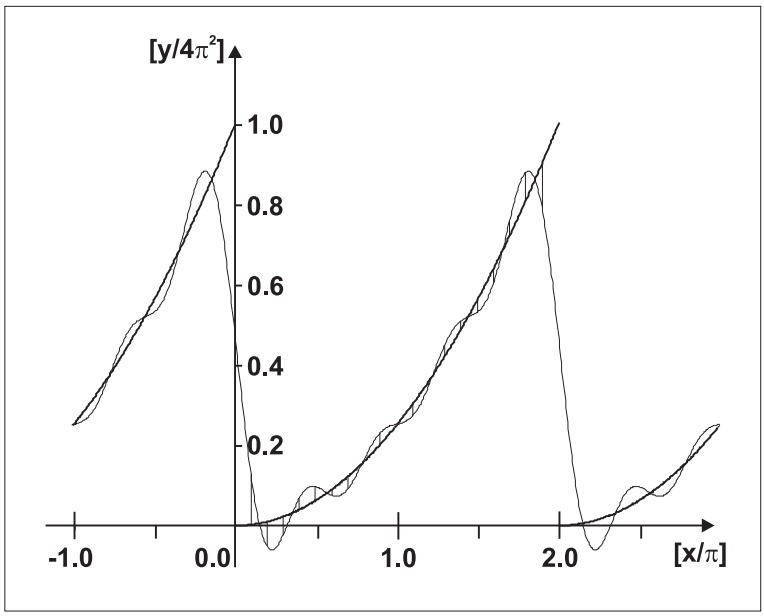
BSP. (17.5.2) Sei $f \in \mathrm{Abb}(\mathbf{R}, \mathbf{R})$ wie folgt definiert:

$$f(x) := x^2 \quad \text{für} \quad 0 \le x < 2\pi \quad \text{und} \quad f(x + 2\pi) := f(x) \ \ \forall\, x \in \mathbf{R}.$$

Auf dem Periodenintervall $[0, 2\pi]$ hat $f(x)$ Sprungstellen bei $x_0 = 0$ und $x_N = 2\pi$. Wir setzen deshalb $f_0 = f_N := \frac{1}{2}\, x_N^2 = 2\pi^2$. Für $N := 20$ haben die numerisch berechneten Koeffizienten a_k^*, b_k^* die folgenden Werte:

k	0	1	2	3	4
a_k^*	26.351 844	4.033 062	1.033 558	0.478 857	0.285 669
b_k^*	0.000 000	−12.462 846	−6.075 104	−3.874 038	−2.716 869

In der folgenden Skizze sind die Graphen von $f(x)$ und $g_4^*(x)$ auf dem Periodenintervall $[0, 2\pi]$ angedeutet. Die Residuen $r_j := g_4^*(x_j) - f_j$ sind durch vertikale Striche markiert.



Approximation im diskreten quadratischen Mittel

Wir gehen jetzt der Frage nach, wie groß der Fehler zwischen den diskreten FOURIER–Koeffizienten a_k^*, b_k^* und den kontinuierlichen FOURIER–Koeffizienten a_k, b_k ist. Eine Antwort gibt der folgende

Satz 17.12 *Die Funktion $f \in \mathrm{Abb}(\mathbf{R}, \mathbf{C})$ genüge der Bedingung* (F). *Es sei $N := 2n$, $n \in \mathbf{N}$, und an einer Sprungstelle x_j werde f_j durch das arithmetische Mittel $\frac{1}{2}\,(f(x_j + 0) + f(x_j - 0))$ definiert. Dann gelten für die diskreten* FOURIER*–Koeffizienten a_k^*, b_k^* aus* (5.3) *die Darstellungen:*

$$\begin{aligned} a_k^* &= a_k + \sum_{l=1}^{\infty} (a_{lN-k} + a_{lN+k}), \quad k = 0, 1, \ldots, n, \\ b_k^* &= b_k - \sum_{l=1}^{\infty} (b_{lN-k} - b_{lN+k}), \quad k = 1, 2, \ldots, n-1. \end{aligned} \tag{5.16}$$

Begründung: Wir drücken in (5.3) den Funktionswert f_j durch die FOURIER–Reihe (4.9) aus:

$$f_j = \frac{a_0}{2} + \sum_{l=1}^{\infty} (a_l \cos l x_j + b_l \sin l x_j), \quad j = 1, 2, \ldots, N.$$

Somit gilt

$$\begin{aligned} a_k^* &= \frac{2}{N}\sum_{j=1}^{N}\Big(\frac{a_0}{2}+\sum_{l=1}^{\infty}(a_l\cos lx_j+b_l\sin lx_j)\Big)\cos kx_j \\ &= \frac{2}{N}\Big(\frac{a_0}{2}\sum_{j=1}^{N}\cos kx_j+\sum_{l=1}^{\infty}(a_l\sum_{j=1}^{N}\cos lx_j\cdot\cos kx_j+b_l\sum_{j=1}^{N}\sin lx_j\cdot\cos kx_j)\Big). \end{aligned}$$

Aus den diskreten Orthogonalitätsrelationen (5.6)–(5.8) folgt daher:

$$a_k^* = \begin{cases} \frac{2}{N}\Big(\frac{N}{2}a_0+N(a_N+a_{2N}+a_{3N}+\cdots)\Big) = a_0+2(a_N+a_{2N}+a_{3N}+\cdots), & k=0, \\ \frac{2}{N}\Big(\frac{N}{2}a_k+\frac{N}{2}(a_{N-k}+a_{N+k}+a_{2N-k}+a_{2N+k}+\cdots)\Big) & \\ = a_k+\sum_{l=1}^{\infty}(a_{lN-k}+a_{lN+k}) & , \ 1\le k<n, \\ \frac{2}{N}\Big(N(a_n+a_{3n}+a_{5n}+\cdots)\Big) = 2(a_n+a_{3n}+a_{5n}+\cdots) & , \ k=n. \end{cases}$$

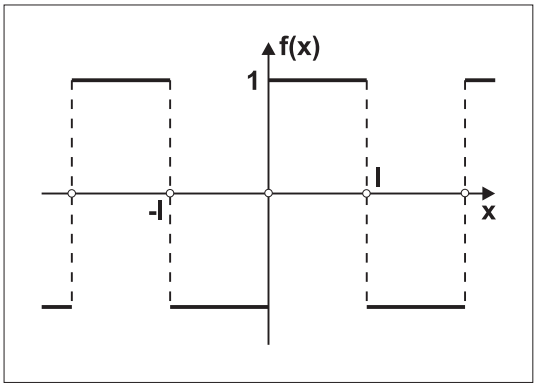
Diese drei Relationen können in der Form (5.16) zusammengefasst werden. Für b_k^* gilt eine analoge Rechnung. □

Folgerung 17.10 *Es gelten die Fehlerabschätzungen*

$$\boxed{\begin{aligned} |a_k^*-a_k| &\le \sum_{l=1}^{\infty}(|a_{lN-k}|+|a_{lN+k}|), \quad 0\le k\le n, \\ |b_k^*-b_k| &\le \sum_{l=1}^{\infty}(|b_{lN-k}|+|b_{lN+k}|), \quad 1\le k\le n-1. \end{aligned}} \tag{5.17}$$

Man kann (5.17) zur Schätzung der Anzahl N der Gitterpunkte x_j verwenden, die benötigt werden, um zu vorgegebenem $\epsilon>0$ die Fehlerschranken $|a_k^*-a_k|\le\epsilon$ und $|b_k^*-b_k|\le\epsilon$ für $k=0,1,\ldots,m<n$ zu erreichen. Dies zeigen wir in folgendem Beispiel.

BSP. (17.5.3) Es sei $f\in \mathrm{Abb}(\mathbf{R},\mathbf{R})$ durch $f(x):=1$ für $0<x<\pi$, $f(x):=0$ für $x=0$, $x=\pi$ sowie durch $f(-x):=-f(x)$ und $f(x+2\pi):=f(x)\ \forall\, x\in\mathbf{R}$ erklärt, siehe folgende Skizze.



Der Graph der Funktion $f(x)$

Offensichtlich erfüllt $f(x)$ die Bedingung (F). Da $f(x)$ eine ungerade Funktion ist, verschwinden die Koeffizienten a_k für alle $k\in\mathbf{N}_0$. Es gilt ferner

$$b_k=\frac{2}{\pi}\int_0^{\pi}1\cdot\sin kx\,dx=-\frac{2}{k\pi}\cos kx\Big|_0^{\pi}=\frac{2}{k\pi}(1-(-1)^k).$$

Wir haben somit $b_k=\frac{4}{k\pi}$ für ungerades k sowie $b_k=0$ für gerades k. Ist $\epsilon>0$ gegeben (*zum Beispiel* $\epsilon:=10^{-6}$), so bestimme man nun N so, dass $|b_k^*-b_k|\le\epsilon$ für $1\le k\le m\ll N$ gilt. Zur Lösung dieser

Aufgabe verwenden wir (5.16), wobei $N = 2n$ gerade sei. Für ungerades k sind dann auch $lN - k$ und $lN + k$ für alle $l \in \mathbf{N}$ ungerade, und wir erhalten:

$$b_k^* - b_k \quad = \quad -\sum_{l=1}^{\infty} (b_{lN-k} - b_{lN+k}) = -\frac{4}{\pi} \sum_{l=1}^{\infty} \Big(\frac{1}{lN-k} - \frac{1}{lN+k} \Big) = -\frac{8k}{\pi} \sum_{l=1}^{\infty} \frac{1}{(lN)^2 - k^2}$$

$$\overset{k \ll N}{\approx} \quad -\frac{8k}{\pi N^2} \sum_{l=1}^{\infty} \frac{1}{l^2} = -\frac{4k\pi}{3N^2}.$$

Aus der Bedingung $|b_k^* - b_k| \leq 10^{-6}$ für $k = 1, 2, \ldots, m$, folgt $N > 2 \cdot 10^3 \sqrt{\frac{m\pi}{3}}$. *Zum Beispiel* erhält man im Falle $m = 10$ die Abschätzung

$$\boxed{N > 6\,472.}$$

Das obige BSP. (17.5.3) lehrt, dass für die Erzielung einer brauchbaren Approximation der FOURIER–Koeffizienten a_k, b_k durch die diskreten Koeffizienten a_k^*, b_k^* die Zahl N oft sehr groß gewählt werden muss. Es ist deshalb besonders wichtig, die Summen (5.3) mit möglichst geringem Rechenaufwand zu bestimmen. Die **Aufwandsbilanz** bei direkter Auswertung von (5.3) zeigt, dass man im Falle $N = 2n$ zur Berechnung eines jeden Koeffizienten a_k^*, $k = 0, 1, \ldots, n$, und b_k^*, $k = 1, 2, \ldots, n-1$, jeweils $N+1$ Multiplikationen/Divisionen und N trigonometrische Funktionensauswertungen benötigt. Insgesamt also:

$$\boxed{Z_{\text{FOUR}} = N^2 + N \text{ w.a.O. und } N^2 \text{ trigonometr. Funktionsauswertungen.}}$$

17.6 Die schnelle FOURIER–Transformation (FFT; Algorithmus von COOLY und TUKEY)

Die Berechnung der Summen (5.3) kann man im Falle $N := 2^m,\ m \in \mathbf{N}$, äußerst effizient durchführen, wenn man zur **diskreten komplexen** FOURIER–**Transformation** übergeht. Wie vorher sollen für gegebene Zahlen $f_j := \frac{1}{2}(f(x_j + 0) + f(x_j - 0))$ mit $x_j := \frac{2\pi}{N} j,\ j = 0, 1, \ldots, N$, die trigonometrischen Summen

$$\boxed{a_k^* := \frac{1}{n} \sum_{j=0}^{N-1} f_j \cos j x_k \quad \text{für} \quad k = 0, 1, \ldots, n := \frac{N}{2},} \tag{6.1}$$

$$\boxed{b_k^* := \frac{1}{n} \sum_{j=1}^{N-1} f_j \sin j x_k \quad \text{für} \quad k = 1, 2, \ldots, n-1} \tag{6.2}$$

berechnet werden. Hier haben wir die Identität $k x_j = \frac{2\pi}{N} jk = j x_k$ verwendet, und wir haben wegen der vorausgesetzten 2π–Periodizität von f die Relation $f_0 = f_N$ eingearbeitet. Wir führen die diskrete komplexe FOURIER–Transformation wie folgt ein. Unter der Annahme

$$\boxed{f \text{ reellwertig}}$$

setzen wir:

$$\boxed{\begin{aligned} y_j &:= f_{2j} + i\, f_{2j+1} \quad \text{für} \quad j = 0, 1, \ldots, n-1, \\ w_n &:= \exp\big(-\tfrac{2\pi i}{n}\big) = \cos\tfrac{2\pi}{n} - i \sin\tfrac{2\pi}{n}. \end{aligned}} \tag{6.3}$$

$$\boxed{c_k^* := \frac{1}{n} \sum_{j=0}^{n-1} y_j w_n^{jk} \quad \text{für} \quad k = 0, 1, \ldots, n-1.} \tag{6.4}$$

Definition 17.7 *Die Abbildung* $\mathbf{C}^n \ni (y_0, y_1, \ldots, y_{n-1})^T \mapsto (c_0^*, c_1^*, \ldots, c_{n-1}^*)^T \in \mathbf{C}^n$ *mit*

$$\boxed{c_k^* := \frac{1}{n} \sum_{j=0}^{n-1} y_j \exp\Big(- \tfrac{2\pi i}{n}\, jk\Big) \quad \text{für} \quad k = 0, 1, \ldots, n-1,}$$

heiße **diskrete** FOURIER–**Transformation**. *Ihre Umkehrabbildung ist gegeben durch*

$$\boxed{y_j = \sum_{k=0}^{n-1} c_k^* \exp\Big(\tfrac{2\pi i}{n}\, jk\Big) \quad \text{für} \quad j = 0, 1, \ldots, n-1.}$$

(*Man verwende die Relation* $\frac{1}{n} \sum\limits_{j=0}^{n-1} \exp\Big(\frac{2\pi i}{n}\,(l-k)j\Big) = \delta_{lk}$.)

Der *Zusammenhang* zwischen den komplexen FOURIER–Transformierten c_k^* und den reellen Koeffizienten a_k^*, b_k^* wird durch den folgenden Satz hergestellt.

Satz 17.13 *Es seien* $a_k' := n a_k^*$, $b_k' := n b_k^*$ *und* $c_k := n c_k^*$ *gesetzt. Dann gilt für* $k = 0, 1, \ldots, n$:

$$a_k' - i\, b_k' = \frac{1}{2}\,(c_k + \overline{c_{n-k}}) + \frac{1}{2i}\,(c_k - \overline{c_{n-k}}) \cdot \exp\,(\,-\tfrac{k\pi i}{n}), \tag{6.5}$$

$$a_{n-k}' - i\, b_{n-k}' = \frac{1}{2}\,(\overline{c_k} + c_{n-k}) + \frac{1}{2i}\,(\overline{c_k} - c_{n-k}) \cdot \exp\,(+\tfrac{k\pi i}{n}). \tag{6.6}$$

Dabei sind $b_0' = b_n' = 0$ *und* $c_n = c_0$ *zu setzen.*

Begründung: Aus den Relationen

$$\frac{1}{2}\,(c_k + \overline{c_{n-k}}) = \frac{1}{2} \sum_{j=0}^{n-1} (y_j w_n^{jk} + \overline{y_j} \cdot \overline{w_n}^{j(n-k)}) = \frac{1}{2} \sum_{j=0}^{n-1} (y_j + \overline{y_j}) w_n^{jk},$$

$$\frac{1}{2i}\,(c_k - \overline{c_{n-k}}) = \frac{1}{2i} \sum_{j=0}^{n-1} (y_j w_n^{jk} - \overline{y_j} \cdot \overline{w_n}^{j(n-k)}) = \frac{1}{2i} \sum_{j=0}^{n-1} (y_j - \overline{y_j}) w_n^{jk}$$

resultiert unter Verwendung von (6.3):

$$\frac{1}{2}\,(c_k + \overline{c_{n-k}}) + \frac{1}{2i}\,(c_k - \overline{c_{n-k}}) \cdot e^{-\frac{k\pi i}{n}} = \sum_{j=0}^{n-1} \Big(f_{2j} e^{-\frac{2\pi i}{n}\, jk} + f_{2j+1} e^{-\frac{\pi i}{n}\,(2j+1)k} \Big)$$

$$= \sum_{j=0}^{n-1} \Big(f_{2j}(\,\cos(2jx_k) - i\,\sin(2jx_k)) + f_{2j+1}(\,\cos(2j+1)x_k - i\,\sin(2j+1)x_k) \Big)$$

$$= a_k' - i\, b_k'.$$

Also gilt (6.5), während (6.6) durch Substitution von k durch $n-k$ aus (6.5) folgt. □

Folgerung 17.11 *Werden* (6.5) *und* (6.6) *in Real– und Imaginärteil zerlegt, so erhält man die vier* FOURIER–*Koeffizienten* $a_k^* = \frac{1}{n}\, a_k'$, $b_k^* = \frac{1}{n}\, b_k'$, $a_{n-k}^* = \frac{1}{n}\, a_{n-k}'$ *und* $b_{n-k}^* = \frac{1}{n}\, b_{n-k}'$ *in Abhängigkeit von* c_k *und* c_{n-k}. *Setzt man nämlich*

$$R_k^{\pm} := \mathrm{Re}\,(c_k \pm \overline{c_{n-k}}), \quad \Delta_k^1 := R_k^- \cdot \sin\tfrac{k\pi}{n} - I_k^- \cdot \cos\tfrac{k\pi}{n},$$

$$I_k^{\pm} := \mathrm{Im}\,(c_k \pm \overline{c_{n-k}}), \quad \Delta_k^2 := R_k^- \cdot \cos\tfrac{k\pi}{n} + I_k^- \cdot \sin\tfrac{k\pi}{n},$$

so erhält man nach einfacher Rechnung:

$$\boxed{\begin{array}{ll} a_k^* = \frac{1}{N}\,(R_k^+ - \Delta_k^1), & b_k^* = \frac{1}{N}\,(-I_k^+ + \Delta_k^2), \\ a_{n-k}^* = \frac{1}{N}\,(R_k^+ + \Delta_k^1), & b_{n-k}^* = \frac{1}{N}\,(I_k^+ + \Delta_k^2). \end{array}} \tag{6.7}$$

Das heißt, zur Berechnung der vier Koeffizienten a_k^*, b_k^*, a_{n-k}^*, b_{n-k}^* aus den zwei komplexen FOURIER*–Transformierten c_k^*, c_{n-k}^* sind nur* **acht** *reelle wesentliche arithmetische Operationen nötig.*

Es bleibt also noch die Aufgabe zu lösen, die Koeffizienten $c_k = nc_k^*$ auf effiziente Weise aus den gegebenen Zahlen y_j (6.3) zu berechnen. Dazu beachten wir, dass die Relation (6.4) eine lineare Transformation

$$\vec{c} = W_n \vec{y}, \quad \vec{c} := (c_0, c_1, \ldots, c_{n-1})^T, \quad \vec{y} := (y_0, y_1, \ldots, y_{n-1})^T \tag{6.8}$$

des $\mathbf{C}^n$ in sich beschreibt. Die Koeffizienten der Matrix $W_n \in \mathbf{C}^{(n,n)}$, nämlich

$$\omega_{jk} := w_n^{jk}, \quad j,k = 0,1,\ldots,n-1,$$

sind Potenzen einer n–ten **Einheitswurzel** $w_n = \exp\,(-\frac{2\pi i}{n})$. Sie bilden daher auf der Einheitskreislinie in $\mathbf{C}$ die Eckpunkte eines regelmäßigen n–Ecks. Das heißt, von den n^2 Koeffizienten ω_{jk} sind nur n Stück voneinander verschieden, nämlich die Potenzen w_n^l, $l = 0,1,\ldots,n-1$, und die restlichen müssen sich auf diese n Potenzen zurückführen lassen.

BSP. (17.6.1) Im Falle $n := 4$ setzen wir $w := w_4 := e^{-i\pi/2} = -i$. Die Transformationsmatrix W_4 lautet hier:

$$W_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & w^1 & w^2 & w^3 \\ 1 & w^2 & 1 & w^2 \\ 1 & w^3 & w^2 & w^1 \end{bmatrix}. \tag{6.9}$$

Verwenden wir die Permutationsmatrix

$$P_{23} := \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

so ist die folgende Faktorisierung sehr einfach nachzurechnen:

$$P_{23}W_4 = \left[\begin{array}{cc|cc} 1 & 1 & 1 & 1 \\ 1 & w^2 & 1 & w^2 \\ \hline 1 & w^1 & w^2 & w^3 \\ 1 & w^3 & w^2 & w^1 \end{array}\right] = \left[\begin{array}{cc|cc} 1 & 1 & 0 & 0 \\ 1 & w^2 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & w^2 \end{array}\right] \left[\begin{array}{cc|cc} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ \hline 1 & 0 & -1 & 0 \\ 0 & w^1 & 0 & -w^1 \end{array}\right] =: D_4 S_4.$$

Hier haben wir die Beziehungen $w^2 = -1$ und $w^3 = -w^1$ eingebaut. Anstelle der Transformation (6.8) haben wir jetzt äquivalent:

$$\vec{c}^{**} := P_{23}\vec{c} = D_4 S_4 \vec{y}. \tag{6.10}$$

Das heißt, für den neuen Vektor

$$\vec{z} := S_4\vec{y} \quad \Leftrightarrow \quad \begin{cases} z_0 := y_0 + y_2, & z_1 := y_1 + y_3, \\ z_2 := (y_0 - y_2)w^0, & z_3 := (y_1 - y_3)w^1, \end{cases} \tag{6.11}$$

mit $w^0 := 1$ hat man jetzt **zweimal eine** FOURIER**–Transformation der Ordnung zwei** durchzuführen, nämlich

$$\vec{c}^{**} = D_4\vec{z} \quad \Leftrightarrow \quad \begin{cases} c_0^{**} = c_0 = z_0 + z_1, & c_1^{**} = c_2 = z_0 + (w_4^2)z_1, \\ c_2^{**} = c_1 = z_2 + z_3, & c_3^{**} = c_3 = z_2 + (w_4^2)z_3. \end{cases} \tag{6.12}$$

Dabei gilt $w_4^2 = -1 = w_2^1$.

Wir zeigen jetzt, dass diesem Beispiel eine allgemeinere Struktur zugrundeliegt: Die diskrete komplexe FOURIER–Transformation der Ordnung $n = 2m,\ m \in \mathbf{N}$, kann zerlegt werden in zwei gleiche FOURIER–Transformationen der Ordnung m.

Satz 17.14 *Es sei $n = 2m,\ m \in \mathbf{N}$. Für $w_n := \exp(-\frac{2\pi i}{n})$ setzen wir*

$$\boxed{\left.\begin{array}{lll} z_j & := & y_j + y_{m+j}, \\ z_{m+j} & := & (y_j - y_{m+j}) w_n^j, \end{array}\right\}\ j = 0, 1, \ldots, m-1.} \tag{6.13}$$

Dann kann die diskrete komplexe FOURIER*–Transformation* (6.8) *der Ordnung n zerlegt werden in die zwei diskreten komplexen* FOURIER*–Transformationen der Ordnung $m = \frac{n}{2}$:*

$$\boxed{\left.\begin{array}{lll} c_{2l} & = & \displaystyle\sum_{j=0}^{m-1} z_j w_m^{jl}, \\ c_{2l+1} & = & \displaystyle\sum_{j=0}^{m-1} z_{m+j} w_m^{jl}, \end{array}\right\}\ l = 0, 1, \ldots, m-1.} \tag{6.14}$$

Hier haben wir $w_m := w_n^2$ zu setzen.

Begründung: (i) Für $k = 2l,\ l = 0, 1, \ldots, m-1$, erschließen wir aus (6.4):

$$c_{2l} = \sum_{j=0}^{2m-1} y_j w_n^{2jl} = \sum_{j=0}^{m-1} (y_j + y_{m+j})(w_n^2)^{jl} = \sum_{j=0}^{m-1} z_j w_m^{jl}.$$

Dabei wurde die Identität $w_n^{2l(m+j)} = w_n^{2jl} \cdot w_n^{nl} = w_n^{2jl}$ verwendet.

(ii) Für $k = 2l+1,\ l = 0, 1, \ldots, m-1$, folgern wir ebenso aus (6.4):

$$\begin{aligned} c_{2l+1} &= \sum_{j=0}^{2m-1} y_j w_n^{j(2l+1)} = \sum_{j=0}^{m-1} \left(y_j w_n^{j(2l+1)} + y_{m+j} w_n^{j(2l+1)} \cdot w_n^{m(2l+1)} \right) \\ &= \sum_{j=0}^{m-1} (y_j - y_{m+j}) w_n^j \cdot (w_n^2)^{jl} = \sum_{j=0}^{m-1} z_{m+j} w_m^{jl}. \end{aligned}$$

Dies war zu zeigen. □

Bemerkung 17.6 (a) Durch die oben gezeigte Zerlegung $FT_n \to 2\,(FT_{n/2})$ entstehen als wesentlicher Rechenaufwand die in (6.13) durchzuführenden $m = \frac{n}{2}$ Multiplikationen.

(b) Ist m wiederum eine gerade Zahl, also $n = 4q$, so kann jede der beiden FOURIER–Transformationen (6.14) nach demselben Verfahren zerlegt werden in je zwei Transformationen der Ordnung $\frac{m}{2} = q$: $FT_n \to 2\,(FT_{n/2}) \to 4\,(FT_{n/4})$. Der Rechenaufwand beträgt dann gemäß (a): $1 \cdot \frac{n}{2} + 2 \cdot \frac{n}{4} = 2 \cdot \frac{n}{2}$ Multiplikationen.

(c) Gilt allgemein $n = 2^p,\ p \in \mathbf{N}$, so kann in p Schritten eine Reduktion auf n FOURIER–Transformationen der Ordnung 1 vorgenommen werden:

$$FT_{2^p} \to 2\,(FT_{2^{p-1}}) \to 2^2\,(FT_{2^{p-2}}) \to \cdots \to 2^{p-1}\,(FT_2) \to 2^p\,(FT_1).$$

In der eindimensionalen FOURIER–Transformation bestehen die Summen (6.4) nur aus einem einzigen Summanden mit Index $j = 0$. Ferner ist $w_1 = 1$, so dass in FT_1 keine Multiplikation durchgeführt werden muss. Nach (b) ergibt sich für $n = 2^p$ folglich ein Rechenaufwand von

$$\boxed{Z_{FT_n} = \frac{n}{2} \cdot p = \frac{n}{2} \log_2 n = \frac{n \cdot \ln n}{2 \cdot \ln 2}} \tag{6.15}$$

komplexen Multiplikationen. Der Rechenaufwand steigt etwa linear mit der Ordnung n an. Man bezeichnet deshalb die hier vorgestellte Methode als *schnelle* FOURIER–*Transformation* (Fast FOURIER Transform ≡ FFT). In der Literatur wird diese Methode den Autoren COOLEY und TUKEY zugewiesen, obwohl eine frühere Version bereits auf GOOD zurückgeht.

(d) Wir stellen in der folgenden Tabelle den Rechenaufwand für einige Zahlen p gemäß (6.15) zusammen. Dabei wird der *Reduktionsfaktor* als Quotient n^2/Z_{FT_n} definiert. Denn wie wir bereits früher festgestellt haben, benötigt man für reellwertiges $f(x)$ bei direkter Auswertung der Formeln (4.3) ca. $N^2 + N$ reelle Multiplikationen. Da jede der in (6.13) auftretenden komplexen Multiplikationen gleichwertig ist mit **vier** reellen Multiplikationen, sind n^2 komplexe Multiplikationen gleichwertig mit $4n^2 = (2n)^2 = N^2$ reellen Multiplikationen. □

p	5	6	8	9	10	11	12
n	32	64	256	512	1024	2048	4096
n^2	1024	4096	65536	$2.62 \cdot 10^5$	$1.05 \cdot 10^6$	$4.19 \cdot 10^6$	$1.68 \cdot 10^7$
Z_{FT_n}	80	192	1024	2304	5120	11264	24576
Red.–Faktor	12.8	21.3	64	114	205	372	683

BSP. (17.6.2) Wir zeigen die algorithmische Realisierung der FFT am Beispiel $n = 8 = 2^3$. In diesem Falle gilt $w := \exp(-\frac{2\pi i}{8}) = \cos\frac{\pi}{4} - i\sin\frac{\pi}{4}$. Die zu berechnenden Hilfsgrößen z_j und z_{m+j} werden wieder mit y_j und y_{m+j} bezeichnet.

y_0	$y_0 := y_0 + y_4$	c_0	$y_0 := y_0 + y_2$	c_0	$y_0 := y_0 + y_1$	$= c_0$
y_1	$y_1 := y_1 + y_5$	c_2	$y_1 := y_1 + y_3$	c_4	$y_1 := (y_0 - y_1)w^0$	$= c_4$
y_2	$y_2 := y_2 + y_6$	c_4	$y_2 := (y_0 - y_2)w^0$	c_2	$y_2 := y_2 + y_3$	$= c_2$
y_3	$y_3 := y_3 + y_7$	c_6	$y_3 := (y_1 - y_3)w^2$	c_6	$y_3 := (y_2 - y_3)w^0$	$= c_6$
y_4	$y_4 := (y_0 - y_4)w^0$	c_1	$y_4 := y_4 + y_6$	c_1	$y_4 := y_4 + y_5$	$= c_1$
y_5	$y_1 := (y_1 - y_5)w^1$	c_3	$y_5 := y_5 + y_7$	c_5	$y_5 := (y_4 - y_5)w^0$	$= c_5$
y_6	$y_6 := (y_2 - y_6)w^2$	c_5	$y_6 := (y_4 - y_6)w^0$	c_3	$y_6 := y_6 + y_7$	$= c_3$
y_7	$y_7 := (y_3 - y_7)w^3$	c_7	$y_7 := (y_5 - y_7)w^2$	c_7	$y_7 := (y_6 - y_7)w^0$	$= c_7$
FT_8	$\to 2\,FT_4$		$\to 4\,FT_2$		$\to 8\,FT_1$	

Beachte **Indizes der Koeffizienten** c_k: Die Buchhaltung der Indizes der gesuchten Koeffizienten c_k geschieht in jedem Transformationsschritt am einfachsten durch eine **binäre Darstellung der Indexwerte**. Für $n := 2^p$ benötigt man zur Indizierung von c_k genau p Binärstellen. Die folgende Tabelle zeigt die Binärdarstellung der Indexwerte nach jedem Transformationsschritt im Fallbeispiel $p = 3$:

j	y_j	$c_k^{(1)}$	$c_k^{(2)}$	$c_k^{(3)}$	$k =$
0	000	000	000	000	0
1	001	010	100	100	4
2	010	100	010	010	2
3	011	110	110	110	6
4	100	001	001	001	1
5	101	011	101	101	5
6	110	101	011	011	3
7	111	111	111	111	7

Die Binärdarstellungen der letzten Spalte sind genau die in umgekehrter Reihenfolge aufgeschriebenen Binärdarstellungen der ersten Spalte. Die eindeutige Zuordnung $y_j \rightleftharpoons c_k$ nach Durchführung aller

Rechenschritte geschieht also durch **Bitumkehr der Binärdarstellungen** von j. Eine Begründung dieser Tatsache erhält man wie folgt:

1.Reduktionsschritt: $FT_{2^p} \to 2\,(FT_{2^{p-1}})$. Alle p Binärstellen des Index j werden zyklisch vertauscht.

2.Reduktionsschritt: $2\,(FT_{2^{p-1}}) \to 4\,(FT_{2^{p-2}})$. Die ersten $p-1$ Binärstellen werden zyklisch vertauscht.

l–ter Reduktionsschritt: $2^{l-1}\,(FT_{2^{p+1-l}}) \to 2^l\,(FT_{2^{p-l}})$. Die ersten $p+1-l$ Binärstellen werden zyklisch vertauscht.

p–ter Reduktionsschritt: $2^{p-1}\,(FT_2) \to 2^p\,(FT_1)$. Die Bitumkehr ist bereits vollständig durchgeführt.

Die Bitumkehr einer ganzen Zahl j mit $0 \le j < n := 2^p,\ p \in \mathbf{N}$, wird durch den folgenden **Algorithmus** bewerkstelligt, der dem Zahlenpaar (j,p) die durch Bitumkehr gewonnene ganze Zahl $k := inv(j,p),\ 0 \le k < n$, zuordnet.

```
1:  Eingabe von  j, p;  0 ≤ j < 2^p;
2:  i := j; k := 0;
3:  für  l := 1, 2, ..., p :
4:      i1 := i div 2;
5:      k := 2 * k + i − 2 * i1;
6:      i := i1;  (Ende l)
7:  inv := k.
```

Hierbei wird mit $i\ div\ 2$ die ganzzahlige Division bezeichnet, die den Wert $\frac{i}{2}$ auf eine ganze Zahl abrundet.

Mit Hilfe der Funktionsroutine $inv(j,p)$ können wir nun den **FFT–Algorithmus** (FFT) formulieren, der aus der Vorgabe einer reellen 2π–periodischen Funktion $f(x)$ mit den diskreten Funktionswerten $f_j := \frac{1}{2}\left(f(\frac{2\pi}{N}\,j+0) + f(\frac{2\pi}{N}\,j-0)\right),\ j = 0,1,\ldots,N-1$, für eine Zahl $N := 2n = 2\cdot 2^p,\ p \in \mathbf{N}$, die diskreten FOURIER–Koeffizienten $a_k^*,\ b_k^*,\ k = 0,1,\ldots,n$, berechnet; siehe nächste Seite.

Erweiterungen: Eine Variante der schnellen FOURIER–Transformation kann auch im Falle $n = mp,\ p$ Primzahl, $m \in \mathbf{N}$, hergeleitet werden:

Satz 17.15 *Es sei $n := mp$ mit p prim und $m \in \mathbf{N}$. Für $w_n := \exp(-\frac{2\pi i}{n})$ setzen wir bei festem $\mu \in \{0,\ldots,p-1\}$:*

$$z_{j+m\mu} := \sum_{l=0}^{p-1} y_{j+lm} w_n^{\mu(j+lm)}, \quad j = 0,1,\ldots,m-1. \tag{6.16}$$

Dann kann die diskrete komplexe FOURIER*–Transformation* (6.8) *der Ordnung $n = mp$ zerlegt werden in die p diskreten komplexen* FOURIER*–Transformationen der Ordnung m:*

$$c_{kp+\mu} = \sum_{j=0}^{m-1} z_{j+m\mu} w_m^{jk}, \quad k = 0,1,\ldots,m-1. \tag{6.17}$$

Begründung: Ersetzen wir in (6.4) k durch $kp+\mu$ für festes $\mu = 0,1,\ldots,p-1$ und $k = 0,1,\ldots,m-1$, so erschließen wir:

$$c_{kp+\mu} = \sum_{j=0}^{pm-1} y_j w_n^{j(kp+\mu)} = \sum_{j=0}^{m-1}\Big(\sum_{l=0}^{p-1} y_{j+lm} w_n^{(j+lm)(kp+\mu)}\Big) = \sum_{j=0}^{m-1} z_{j+m\mu} w_m^{jk}.$$

Dabei haben wir die Identität

$$w_n^{(j+lm)kp} = w_n^{jkp}\cdot w_n^{lkpm} = (w_n^p)^{jk}(w_n^n)^{lk} = w_m^{jk}$$

verwendet. □

Algorithmus der FFT

1:	Eingabe von $n, p, f_j;\ n = 2^p,\ 0 \le j \le 2n-1;$
2:	**für** $j := 0, 1, \ldots, n-1:$
3:	$y_{j+1} := f_{2j} + i\, f_{2j+1}$;(Ende j)
4:	$n2 := n\ div\ 2; m := n2; k := 0;$
5:	**für** $q := 1, 2, \ldots, p:$
6:	**wiederhole:**
7:	**für** $j := 1, 2, \ldots, m:$
8:	$k1 := k\ div\ m;$
9:	$a := inv(k1, p) * 2 * \pi / n;$
10:	$w := \cos a - i\ \sin a;$
11:	$l := k + 1; lm := l + m;$
12:	$t := w * y_{lm};$
13:	$y_{lm} := y_l - t; y_l := y_l + t;$
14:	$k := k + 1;$ (Ende j)
15:	$k := k + m;$
16:	**bis** $k \ge n;$
17:	$k := 0; m := m\ div\ 2;$ (Ende q)
18:	**für** $k := 1, 2, \ldots, n:$
19:	$j := inv(k-1, p) + 1;$
20:	**falls** $j > k$ **dann**
21:	$t := y_k; y_k := y_j; y_j := t;$ (Ende falls, Ende k)
22:	$y_{n+1} := y_1; a := 0.5/n$
23:	**für** $k := 0, 1, \ldots, n2:$
24:	$R^+ := \mathrm{Re}\,(y_{k+1} + y_{n-k+1}); R^- := \mathrm{Re}\,(y_{k+1} - y_{n-k+1});$
25:	$I^+ := \mathrm{Im}\,(y_{k+1} - y_{n-k+1}); I^- := \mathrm{Im}\,(y_{k+1} + y_{n-k+1});$
26:	$c := \cos \frac{k\pi}{n}; s := \sin \frac{k\pi}{n};$
27:	$D1 := s * R^- - c * I^-; D2 := c * R^- + s * I^-;$
28:	$a_k^* := a * (R^+ - D1); a_{n-k}^* := a * (R^+ + D1);$
29:	$b_k^* := a * (D2 - I^+); b_{n-k}^* := a * (D2 + I^+).$ (Ende k)

Bemerkung 17.7 (a) Wir können in (6.16) für festes $j = 0, 1, \ldots, m-1$ wegen $w_n^m = w_p$ auch folgende Darstellung herleiten:

$$\boxed{z_{j+m\mu} = \Big(\sum_{l=0}^{p-1} y_{j+lm} w_p^{\mu l} \Big) \cdot w_n^{\mu j}, \quad \mu = 0, 1, \ldots, p-1.} \tag{6.18}$$

Man erkennt hier in der runden Klammer eine diskrete komplexe FOURIER–Transformation der Ordnung p. Beachtet man, dass in (6.18) für $\mu = 0$ nur Summationen durchzuführen sind, so erfordert die Gesamtauswertung von (6.18) $p(p-1)m$ komplexe Multiplikationen. Zusammen mit der Auswertung von (6.17) benötigen wir

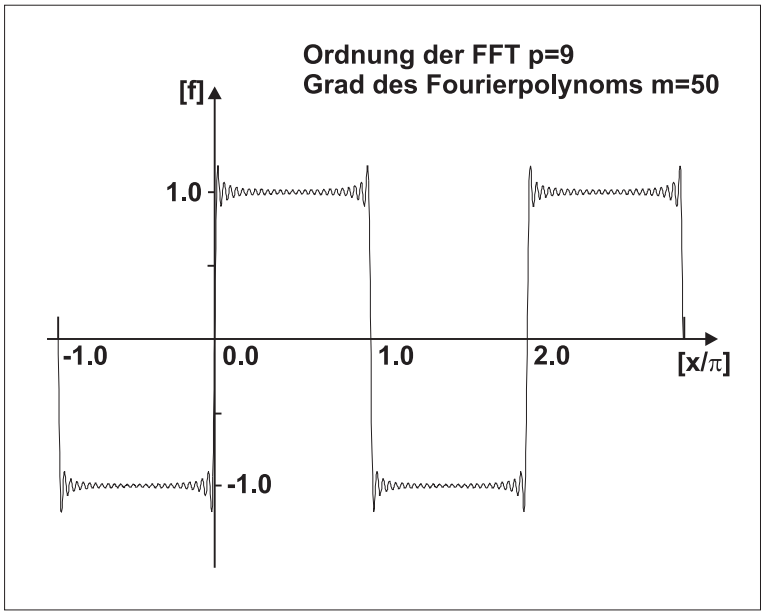
$$\boxed{Z_n = Z_{mp} = pZ_m + n(p-1) \quad \text{w.a.O.}} \tag{6.19}$$

(b) Ist m keine Primzahl, so können die p diskreten komplexen FOURIER–Transformationen (6.17) der Ordnung m weiter reduziert werden. Falls eine Primzahlzerlegung

$$n = p_1^{q_1} \cdot p_2^{q_2} \cdots p_r^{q_r},\ p_j \ \text{ prim},\ q_j \in \mathbf{N},$$

vorliegt, so wird eine Reduktion auf FOURIER–Transformationen der Ordnung 1 in $q_1 + q_2 + \cdots + q_r$ Schritten erreicht. □

BSP. (17.6.3) Sei $n = 240 = 2^4 \cdot 3 \cdot 5$. Dann sind gemäß (6.19) total $n(1+1+1+1+2+4) = 2400$ komplexe Multiplikationen durchzuführen, falls die Reduktion auf FOURIER–Transformationen der Ordnung 1 in $4 + 1 + 1 = 6$ Schritten erfolgt ist. Bei der direkten Auswertung hat man hingegen $n^2 = 240^2 = 57\,600$ Multiplikationen durchzuführen. Wir weisen darauf hin, dass die Aufwandsformel (6.19) etwas zu pessimistisch ist. Die tatsächlich erforderliche Zahl von Multiplikationen liegt darunter. Hat man nämlich $n = 2^q$, so liefert (2.19) $Z_n = nq$. Das ist der doppelte Wert, der sich nach (6.15) ergäbe.



FOURIER–**Polynom vom Grad** $m = 50$**, berechnet mit der FFT**

BSP. (17.6.4) Wir greifen hier nochmals das Beispiel (17.5.3) auf. Die 2π–periodische, ungerade Funktion $f(x) := 1$ für $0 < x < \pi$ und $f(x) := -1$ für $\pi < x < 2\pi$ hat die nichtverschwindenden FOURIER–Koeffizienten $b_k := \frac{4}{k\pi}$, $k \geq 1$, ungerade. Die ersten 50 mit dem FFT–Algorithmus berechneten diskreten FOURIER–Koeffizienten b_k^* haben im Falle $p = 9$ (d.h. für $n = 512$) die in der folgenden Tabelle aufgelisteten numerischen Werte. Im Vergleich zu den in der darunterstehenden Tabelle angegebenen analytisch berechneten Werten erkennt man, dass der Fehler $|b_k - b_k^*|$ etwa die Größenordnung $2 \cdot 10^{-4}$ hat. Mit $m := 50$ berechnet man aus der in BSP. (17.5.3) angegebenen Fehlerabschätzung die Mindeststützstellenzahl $N > 1023.33$. Wir haben exakt mit der Stützstellenzahl $N = 1024$ gerechnet.

k	b_k^*	k	b_k^*	k	b_k^*	k	b_k^*	k	b_k^*	k	b_k^*
0	0.000 00	1	1.273 236	2	0.000 00	3	0.424 401	4	0.000 00	5	0.254 628
6	0.000 00	7	0.181 863	8	0.000 00	9	0.141 435	10	0.000 00	11	0.115 705
12	0.000 00	13	0.097 890	14	0.000 00	15	0.084 823	16	0.000 00	17	0.074 829
18	0.000 00	19	0.066 937	20	0.000 00	21	0.060 547	22	0.000 00	23	0.055 266
24	0.000 00	25	0.050 830	26	0.000 00	27	0.047 049	28	0.000 00	29	0.043 789
30	0.000 00	31	0.040 948	32	0.000 00	33	0.038 451	34	0.000 00	35	0.036 238
36	0.000 00	37	0.034 264	38	0.000 00	39	0.032 491	40	0.000 00	41	0.030 891
42	0.000 00	43	0.029 438	44	0.000 00	45	0.028 114	46	0.000 00	47	0.026 902
48	0.000 00	49	0.025 788	50	0.000 00						

k	b_k	k	b_k	k	b_k	k	b_k	k	b_k	k	b_k
0	0.000 00	1	1.273 240	2	0.000 00	3	0.424 413	4	0.000 00	5	0.254 648
6	0.000 00	7	0.181 891	8	0.000 00	9	0.141 471	10	0.000 00	11	0.115 749
12	0.000 00	13	0.097 942	14	0.000 00	15	0.084 883	16	0.000 00	17	0.074 896
18	0.000 00	19	0.067 013	20	0.000 00	21	0.060 630	22	0.000 00	23	0.055 358
24	0.000 00	25	0.050 930	26	0.000 00	27	0.047 157	28	0.000 00	29	0.043 905
30	0.000 00	31	0.041 072	32	0.000 00	33	0.038 583	34	0.000 00	35	0.036 378
36	0.000 00	37	0.034 412	38	0.000 00	39	0.032 647	40	0.000 00	41	0.031 055
42	0.000 00	43	0.029 610	44	0.000 00	45	0.028 294	46	0.000 00	47	0.027 090
48	0.000 00	49	0.025 984	50	0.000 00						

Die obige Grafik zeigt das mit den Koeffizienten b_k^* berechnete FOURIER–Polynom $g_m^*(x)$ aus (5.4) vom Grade $m = 50$.

17.7 Das Anfangs–Randwert–Problem für die eindimensionale Wärmeleitungsgleichung

Wir haben in Abschnitt 17.1 gezeigt, dass die **Wärmeleitungs–** oder **Diffusionsgleichung**

$$\boxed{u_t = k\,\Delta u, \qquad k := \frac{\kappa}{c\rho}: \quad \text{Temperaturleitfähigkeit},} \tag{7.1}$$

den zeit– und ortsabhängigen Temperaturverlauf $u = u(x, y, z; t)$ in einem homogenen isotropen Wärmeleiter der Dichte $\rho > 0$ (konstant), der spezifischen Wärme c und der Wärmeleitfähigkeit $\kappa > 0$ beschreibt. Treten im Inneren des Wärmeleiters **Wärmequellen** auf, so ist die partielle Differentialgleichung (7.1) durch die inhomogene DGl

$$u_t - k\,\Delta u = g(x, y, z; t) \tag{7.2}$$

bei vorgegebener Funktion g zu ersetzen. Wir werden hier nicht eine allgemeine Lösungstheorie dieser partiellen DGl entwickeln, sondern lediglich auf einige spezielle Problemstellungen im eindimensionalen Ortsraum eingehen. Wir haben bereits in (1.3) den Sonderfall der **eindimensionalen Wärmeleitungsgleichung** andiskutiert:

$$\boxed{u_t - k\,u_{xx} = g(x, t).} \tag{7.3}$$

Diese kann häufig mit der Methode der FOURIER–Reihen behandelt werden, was insbesondere im Fall des **eindimensionalen Wärmeleiters endlicher Länge** zutrifft.

Definition 17.8 Das Anfangs–Randwert–Problem (ARWP) für den eindimensionalen Wärmeleiter der Länge $L > 0$*: Zu vorgegebenen Funktionen* $f(x)$, $\varphi_1(t)$, $\varphi_2(t)$ *und* $g(x, t)$ *sowie zu gegebenen Zahlen* α_j, $\beta_j \in \mathbf{R}$ *mit* $\alpha_j^2 + \beta_j^2 > 0$ *für* $j = 1, 2$ *ist eine Funktion* $u = u(x, t)$ *auf* $[0, L] \times [0, +\infty)$ *gesucht, so dass gilt:*

$$\begin{cases} u_t - k\,u_{xx} = g(x,t) & : \ 0 < x < L, \ \ t > 0, \qquad \text{(DGl)} \\ \alpha_1 u_x(0,t) + \beta_1 u(0,t) = \varphi_1(t), \quad \alpha_2 u_x(L,t) + \beta_2 u(L,t) = \varphi_2(t) & : \ t \geq 0, \qquad \text{(RB)} \\ u(x,0) = f(x) & : \ 0 \leq x \leq L. \qquad \text{(AB)} \end{cases}$$

Die partielle Differentialgleichung (DGl) heiße **homogen**, wenn $g = 0$ gilt. Die Randbedingungen (RB) heißen **homogen**, wenn $\varphi_1 = \varphi_2 = 0$ gelten. Es ist nicht sinnvoll, das obige Anfangs–Randwert–Problem in seiner vollen Allgemeinheit zu diskutieren. Wir betrachten hier nur Spezialfälle, die aber den Charakter der Vorgehensweise bereits offenlegen.

Teil (A): Homogene Randbedingungen

BSP. (17.7.1) Wir betrachten hier das folgende Anfangs–Randwert–Problem:

$$\text{(ARWP)} \quad \begin{cases} u_t - k\,u_{xx} = 0 & : \ 0 < x < L, \ \ t > 0, \qquad \text{(DGl)} \\ u(0,t) = u(L,t) = 0 & : \ t \geq 0, \qquad \text{(RB)} \\ u(x,0) = f(x) & : \ 0 \leq x \leq L. \qquad \text{(AB)} \end{cases}$$

Dies ist das Beispiel eines quellenfreien eindimensionalen Wärmeleiters, dessen Enden auf konstanter Temperatur 0 gehalten werden und dessen Anfangstemperatur gemäß $f(x)$ vorgegeben ist. Wir haben dieses Problem bereits in Abschnitt 17.1 behandelt. Genügt die Funktion f der Bedingung (F) aus Definition 17.6, so lautet die **formale Lösung** der Aufgabe ARWP gemäß (1.10):

$$\boxed{u(x,t) = \frac{2}{L} \sum_{n=1}^{\infty} e^{-\left(\frac{n\pi}{L}\right)^2 kt} \sin\frac{n\pi x}{L} \Big(\int_0^L f(\xi) \sin\frac{n\pi\xi}{L}\, d\xi \Big).} \tag{7.4}$$

Bemerkung 17.8 Wir sprechen hier von einer **formalen Lösung**, da wir nicht untersuchen wollen, ob die Funktion (7.4) tatsächlich in jedem Punkt (x,t) die DGl und die Anfangsbedingungen (AB) erfüllt. Zum erstgenannten Problem muss die gliedweise Differenzierbarkeit der Funktionenreihe überprüft werden, zum Beispiel mit den bekannten Weierstrass–Kriterien. Zum zweitgenannten Problem können wir sicher feststellen, dass gilt:

$$u(x,0) = \sum_{n=1}^{\infty} \Big(\frac{2}{L} \int_0^L f(\xi) \sin\frac{n\pi\xi}{L}\, d\xi \Big) \sin\frac{n\pi x}{L} = \frac{1}{2}(f(x+0) + f(x-0)).$$

Da das trigonometrische Funktionensystem $\varphi_n(x) := \sin\frac{n\pi x}{L}$, $n \in \mathbf{N}$, mit der kleinsten gemeinsamen Periode $T = 2L$ vorgegeben war, muss die Funktion $f \in \mathrm{Abb}\,([0,L], \mathbf{R})$ zunächst als **ungerade** Funktion auf das Intervall $[-L, 0)$ fortgesetzt werden und danach als $2L$–periodische Funktion auf ganz $\mathbf{R}$ definiert werden: □

$$f(x) := -f(-x) \ \ \forall\, x \in [-L, 0), \qquad f(x + 2L) := f(x) \ \ \forall\, x \in \mathbf{R}.$$

BSP. (17.7.2) Wir betrachten nun das folgende Anfangs–Randwert–Problem:

$$\text{(ARWP)} \quad \begin{cases} u_t - k\,u_{xx} = 0 & : \ 0 < x < L, \ \ t > 0, \qquad \text{(DGl)} \\ u_x(0,t) = u_x(L,t) = 0 & : \ t \geq 0, \qquad \text{(RB)} \\ u(x,0) = f(x) & : \ 0 \leq x \leq L. \qquad \text{(AB)} \end{cases}$$

Dies ist das Beispiel eines quellenfreien eindimensionalen Wärmeleiters, dessen Enden wärmeisoliert sind und dessen Anfangstemperatur gemäß $f(x)$ vorgegeben ist. Wir gehen wie in Abschnitt 17.1 nach der Methode des Produktansatzes vor:

- Bernoulli**scher Separationsansatz und TdV:**

$$u(x,t) = X(x) \cdot T(t) \quad \Rightarrow \quad \frac{1}{k}\frac{\dot{T}}{T} = \frac{X''}{X} =: -\lambda^2, \quad \lambda \geq 0.$$

- **Integration der resultierenden DGln:**

$$T(t) = e^{-k\lambda^2 t}, \ \ t \geq 0, \qquad X(x) = \begin{cases} \alpha_0 + b_0 x & : \ \lambda = 0, \\ a_\lambda \cos\lambda x + b_\lambda \sin\lambda x & : \ \lambda > 0. \end{cases}$$

- **Anpassen der Randbedingungen:**
  - Erste Randbedingung:

$$u_x(0,t)=0 \quad \Leftarrow \quad X'(0)=0=\begin{cases} b_0 & : \ \lambda=0, \\ \lambda b_\lambda & : \ \lambda>0, \end{cases} \quad \Rightarrow \quad b_0=0=b_\lambda.$$

  - Zweite Randbedingung:

$$u_x(L,t)=0 \quad \Leftarrow \quad X'(L)=0=\begin{cases} 0 & : \ \lambda=0, \\ \lambda a_\lambda \sin \lambda L & : \ \lambda>0, \end{cases} \quad \Rightarrow \quad \lambda L=n\pi, \ \ n\in\mathbf{N}.$$

  Wir erhalten partikuläre Lösungen in der Form

$$u_n(x,t)=\begin{cases} \frac{a_0}{2} & : \ n=0, \\ a_n e^{-\left(\frac{n\pi}{L}\right)^2 kt}\cos\frac{n\pi x}{L} & : \ n\in\mathbf{N}, \end{cases}$$

  die bereits die geforderten Randbedingungen erfüllen.

- **Superposition:**

$$u(x,t)=\frac{a_0}{2}+\sum_{n=1}^{\infty} a_n e^{-\left(\frac{n\pi}{L}\right)^2 kt}\cos\frac{n\pi x}{L}.$$

- **Anpassen der Anfangsbedingung:**

$$u(x,0)=\frac{a_0}{2}+\sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{L}\sim f(x).$$

  Da das trigonometrische Funktionensystem $\varphi_0(x):=\frac{1}{\sqrt{2}}$, $\varphi_n(x):=\cos\frac{n\pi x}{L}$, $n\in\mathbf{N}$, mit der kleinsten gemeinsamen Periode $T=2L$ vorgelegt ist, muss die Funktion $f\in \mathrm{Abb}\,([0,L],\mathbf{R})$ zunächst als **gerade** Funktion auf das Intervall $[-L,0)$ fortgesetzt werden und danach als $2L$–periodische Funktion auf ganz $\mathbf{R}$ definiert werden:

$$f(x):=f(-x) \ \ \forall\, x\in[-L,0), \qquad f(x+2L):=f(x) \ \ \forall\, x\in\mathbf{R}.$$

Aus den Formeln für die FOURIER–Koeffizienten erhalten wir nun die formale Lösung

$$\boxed{u(x,t)=\frac{1}{L}\int_0^L f(\xi)\,d\xi+\frac{2}{L}\sum_{n=1}^{\infty} e^{-\left(\frac{n\pi}{L}\right)^2 kt}\cos\frac{n\pi x}{L}\Big(\int_0^L f(\xi)\cos\frac{n\pi\xi}{L}\,d\xi\Big).} \tag{7.5}$$

Bemerkung 17.9 Für $t\to\infty$ resultiert aus (7.5) die asymptotische Temperaturverteilung $u_\infty=\frac{1}{L}\int_0^L f(\xi)\,d\xi$. Es stellt sich also asymptotisch der Integralmittelwert der Anfangstemperaturverteilung ein, und dies ist aus energetischen Gründen völlig korrekt: Wegen der Isolierung des Wärmeleiters können im Zeitverlauf keine Wärmeverluste auftreten. □

BSP. (17.7.3) Wir betrachten nun das folgende Anfangs–Randwert–Problem:

$$\text{(ARWP)} \quad \begin{cases} u_t-k\,u_{xx}=0 & : \ 0<x<L, \ \ t>0, & \text{(DGl)} \\ u(0,t)=0=u_x(L,t)+h\,u(L,t) & : \ t\geq 0, \quad h>0 \text{ fest}, & \text{(RB)} \\ u(x,0)=f(x) & : \ 0\leq x\leq L. & \text{(AB)} \end{cases}$$

Diese Problemstellung entspricht dem Beispiel eines quellenfreien eindimensionalen Wärmeleiters, dessen eines Ende $x=0$ auf konstanter Temperatur 0 gehalten wird und dessen anderes Ende $x=L$ Wärmeenergie durch Abstrahlung an das umgebende Medium abgibt. Wir gehen wie in BSP. (17.7.2) nach der Methode des Produktansatzes vor:

- **Bernoullischer Separationsansatz und TdV:**

$$u(x,t) = X(x)\cdot T(t) \quad \Rightarrow \quad \frac{1}{k}\frac{\dot{T}}{T} = \frac{X''}{X} =: -\lambda^2, \quad \lambda \geq 0.$$

- **Integration der resultierenden DGln:**

$$T(t) = e^{-k\lambda^2 t}, \;\; t \geq 0, \qquad X(x) = \begin{cases} \alpha_0 + b_0 x & : \; \lambda = 0, \\ a_\lambda \cos\lambda x + b_\lambda \sin\lambda x & : \; \lambda > 0. \end{cases}$$

- **Anpassen der Randbedingungen:**

  - Erste Randbedingung:

$$u(0,t) = 0 \quad \Leftarrow \quad X(0) = 0 = \begin{cases} \alpha_0 & : \; \lambda = 0, \\ a_\lambda & : \; \lambda > 0, \end{cases} \quad \Rightarrow \quad \alpha_0 = 0 = a_\lambda.$$

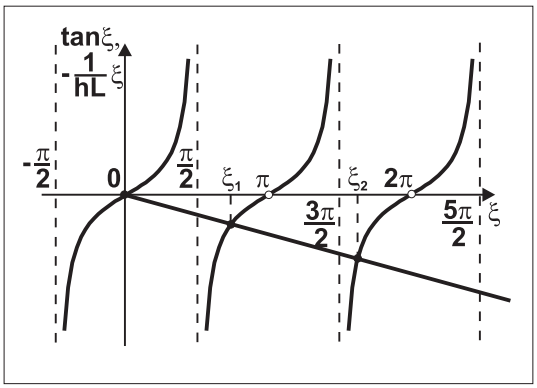
  - Zweite Randbedingung:

$$u_x(L,t) + hu(L,t) = 0 \quad \Leftarrow \quad X'(L) + hX(L) = 0 = \begin{cases} b_0(1+hL) & : \; \lambda = 0, \\ b_\lambda(\lambda\cos\lambda L + h\sin\lambda L) & : \; \lambda > 0, \end{cases}$$

$$\Rightarrow \quad b_0 = 0 \quad \text{und}$$

$$\boxed{\tan\lambda L = -\frac{\lambda L}{hL} \quad \overset{\xi := \lambda L}{\Leftrightarrow} \quad \tan\xi = -\frac{\xi}{hL}.} \tag{7.6}$$

Man erkennt an der folgenden Skizze, dass die transzendente Gleichung (7.6) abzählbar unendlich viele Lösungen $\xi_n > 0$, $n \in \mathbf{N}$, hat mit $\pi(n - \frac{1}{2}) < \xi_n < \pi(n + \frac{1}{2})$ und $\xi_n \approx \pi(n - \frac{1}{2})$ für $n \gg 1$.



Die Lösungen der transzendenten Gleichung $\tan\xi = -\frac{\xi}{hL}$

Dementsprechend hat die Gleichung $\tan\lambda L = -\frac{\lambda L}{hL}$ abzählbar unendlich viele Lösungen $\lambda_n = \frac{\xi_n}{L} > 0$, $n \in \mathbf{N}$, denen die **Eigenlösungen** $\varphi_n(x) := \sin\lambda_n x$, $n \in \mathbf{N}$, zugeordnet sind. Wir erhalten nun partikuläre Lösungen in der Form

$$u_n(x,t) = b_n e^{-\lambda_n^2 kt}\sin\lambda_n x, \quad n \in \mathbf{N},$$

die bereits die geforderten Randbedingungen erfüllen.

- **Superposition:**

$$u(x,t) = \sum_{n=1}^{\infty} b_n e^{-\lambda_n^2 kt} \sin \lambda_n x.$$

- **Anpassen der Anfangsbedingung:**

$$u(x,0) = \sum_{n=1}^{\infty} b_n \sin \lambda_n x \overset{?}{\sim} f(x). \tag{7.7}$$

Wegen $\lambda_n \neq n\omega$ haben wir **keine** FOURIER–Reihe vorliegen. Wir zeigen aber, dass das Funktionensystem $\varphi_n(x) := \sin \lambda_n x$ auf dem Intervall $I := [0, L]$ ein Orthogonalsystem ist. Es gilt nämlich:

$$\left.\begin{array}{l|l} \varphi_n'' + \lambda_n^2 \varphi_n = 0 & \cdot \varphi_m \\ \varphi_m'' + \lambda_m^2 \varphi_m = 0 & \cdot \varphi_n \end{array}\right\} (-) \quad \Rightarrow \quad \underbrace{\varphi_m \varphi_n'' - \varphi_m'' \varphi_n}_{=(\varphi_m \varphi_n' - \varphi_m' \varphi_n)'} = (\lambda_m^2 - \lambda_n^2) \varphi_n \varphi_m.$$

Die Integration dieser letzten Gleichung über das Intervall $[0, L]$ liefert

$$(\lambda_m^2 - \lambda_n^2) \int_0^L \varphi_n(x)\varphi_m(x)\,dx = (\varphi_m(x)\varphi_n'(x) - \varphi_m'(x)\varphi_n(x))\Big|_0^L = \varphi_m(L)\varphi_n'(L) - \varphi_m'(L)\varphi_n(L).$$

Verwenden wir die Gleichung (7.6), so muss $\varphi_m'(L) = \lambda_m \cos \lambda_m L = -h \sin \lambda_m L = -h\varphi_m(L)$ gelten und somit auch $\varphi_m(L)\varphi_n'(L) - \varphi_m'(L)\varphi_n(L) = 0$. Das heißt, für $n \neq m$ erhalten wir

$$\boxed{\int_0^L \varphi_n(x)\varphi_m(x)\,dx = 0, \qquad n \neq m,}$$

während für $n = m$ gilt:

$$\boxed{\int_0^L \varphi_m^2(x)\,dx = \frac{1}{2}\int_0^L (1 - \cos 2\lambda_m x)\,dx = \frac{L}{2} - \frac{1}{2\lambda_m} \sin \lambda_m L \cdot \cos \lambda_m L.}$$

Man kann nachweisen, dass das Funktionensystem $(\varphi_n(x))_{n\in\mathbf{N}}$ auf dem Intervall $[0, L]$ ein **vollständiges Orthogonalsystem** ist, so dass die PARSEVAL–Gleichung gilt. Dementsprechend erhält man aus dem Ansatz (7.7) unter Verwendung des Satzes 17.1 die FOURIER–Koeffizienten

$$b_n = \Big(\int_0^L \varphi_n^2(x)\,dx \Big)^{-1} \int_0^L f(x)\varphi_n(x)\,dx, \qquad n \in \mathbf{N}.$$

Nun resultiert die formale Lösung des ARWP in der Form

$$\boxed{u(x,t) = 2 \sum_{n=1}^{\infty} \frac{\lambda_n \int_0^L f(\xi) \sin \lambda_n \xi\,d\xi}{\lambda_n L - \sin \lambda_n L \cdot \cos \lambda_n L}\, e^{-\lambda_n^2 kt} \sin \lambda_n x.} \tag{7.8}$$

BSP. (17.7.4) Wir betrachten das folgende Anfangs–Randwert–Problem mit inhomogener DGl:

$$\text{(ARWP)} \quad \begin{cases} u_t - k\,u_{xx} = g(x,t) & : \ 0 < x < L, \ \ t > 0, & \text{(DGl)} \\ u(0,t) = 0 = u(L,t) & : \ t \geq 0, & \text{(RB)} \\ u(x,0) = f(x) & : \ 0 \leq x \leq L. & \text{(AB)} \end{cases}$$

Diese Problemstellung entspricht dem Beispiel eines eindimensionalen Wärmeleiters, der im Inneren Wärmequellen der Intensität $g(x,t)$ besitzt und dessen Enden auf konstanter Temperatur 0 gehalten

werden. Das **inhomogene** Problem wird zunächst wie das **homogene** Problem behandelt, welches gemäß BSP. (17.7.1) die Lösung

$$u_h(x,t) = \sum_{n=1}^{\infty} b_n e^{-\left(\frac{n\pi}{L}\right)^2 kt} \sin\frac{n\pi x}{L}$$

besitzt, die bereits die geforderten Randbedingungen (RB) erfüllt. Zur Lösung der **inhomogenen** DGl macht man nun einen **Ansatz** in der Form

$$\boxed{u(x,t) = \sum_{n=1}^{\infty} p_n(t) \sin\frac{n\pi x}{L}.} \tag{7.9}$$

Durch formales Einsetzen in die inhomogene DGl resultiert

$$\sum_{n=1}^{\infty} \Big(\dot p_n(t) + k(\frac{n\pi}{L})^2 p_n(t)\Big) \sin\frac{n\pi x}{L} = g(x,t).$$

Für festgehaltenes $t > 0$ ist dies eine FOURIER–Reihe für die Funktion $g(\cdot,t)$. Die Koeffizientenformeln liefern dementsprechend die folgende Familie von linearen gewöhnlichen DGln 1.Ordnung:

$$\dot p_n(t) + k(\frac{n\pi}{L})^2 p_n(t) = g_n(t) := \frac{2}{L}\int_0^L g(x,t)\sin\frac{n\pi x}{L}\,dx, \qquad n \in \mathbf{N}.$$

Es bezeichne $q_n(t)$ diejenige partikuläre Lösung, die der Anfangsbedingung $q(0) = 0$ genügt. Dann ist die Funktion

$$u(x,t) = \sum_{n=1}^{\infty} \Big(q_n(t) + b_n e^{-\left(\frac{n\pi}{L}\right)^2 kt}\Big) \sin\frac{n\pi x}{L}$$

eine formale Lösung der inhomogenen (DGl), die bereits die erforderlichen Randbedingungen (RB) erfüllt. Das Anpassen der Anfangsbedingung (AB) geschieht nun wiederum durch FOURIER–Entwicklung

$$u(x,0) = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L} \sim f(x).$$

Wir betrachten konkret das folgende

$$\text{(ARWP)} \quad \begin{cases} u_t - k\,u_{xx} = (1-\frac{x}{L})T_0 e^{-t} & : \ 0 < x < L, \ \ t > 0, \qquad \text{(DGl)} \\ u(0,t) = 0 = u(L,t) & : \ t \ge 0, \qquad \text{(RB)} \\ u(x,0) = x \sim \dfrac{2L}{\pi}\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \sin\frac{n\pi x}{L} & : \ 0 \le x \le L, \qquad \text{(AB)} \end{cases}$$

man vergleiche BSP. (17.3.1). Der Ansatz (7.9) führt hier auf die gewöhnlichen DGln

$$\dot p_n(t) + k(\frac{n\pi}{L})^2 p_n(t) = T_0 e^{-t}\,\frac{2}{L}\int_0^L (1-\frac{x}{L})\sin\frac{n\pi x}{L}\,dx = \frac{2T_0}{n\pi}\,e^{-t}, \quad n \in \mathbf{N},$$

mit den allgemeinen Lösungen (wir nehmen hier $k(\frac{n\pi}{L})^2 \neq 1 \ \forall\, n \in \mathbf{N}$ an, um nicht auch noch den Resonanzfall diskutieren zu müssen!)

$$p_n(t) = b_n e^{-\left(\frac{n\pi}{L}\right)^2 kt} + \frac{2T_0 e^{-t}}{n\pi(k(\frac{n\pi}{L})^2 - 1)}, \qquad n \in \mathbf{N}.$$

Die partikuläre Lösung $q_n(t)$ mit dem Anfangswert $q_n(0) = 0$ lautet somit

$$q_n(t) = \frac{2T_0(e^{-t} - e^{-\left(\frac{n\pi}{L}\right)^2 kt})}{n\pi(k(\frac{n\pi}{L})^2 - 1)},$$

und hieraus resultiert die folgende formale Lösung des ARWP

$$\boxed{u(x,t) = 2L \sum_{n=1}^{\infty} \frac{1}{n\pi} \Big(\frac{2T_0}{L} \, \frac{e^{-t} - e^{-\left(\frac{n\pi}{L}\right)^2 kt}}{k(\frac{n\pi}{L})^2 - 1} - (-1)^n e^{-\left(\frac{n\pi}{L}\right)^2 kt} \Big) \sin \frac{n\pi x}{L}.} \tag{7.10}$$

Teil (B): **Inhomogene Randbedingungen**

BSP. (17.7.5) Wir betrachten exemplarisch das folgende Anfangs–Randwert–Problem:

$$\text{(ARWP)} \quad \begin{cases} u_t - k\, u_{xx} = g(x,t) & : \ 0 < x < L, \ \ t > 0, & \text{(DGl)} \\ u(0,t) = \varphi_1(t), \quad u(L,t) = \varphi_2(t) & : \ t \geq 0, & \text{(RB)} \\ u(x,0) = f(x) & : \ 0 \leq x \leq L. & \text{(AB)} \end{cases}$$

Bei inhomogenen Randbedingungen führt der Bernoullische Separationsansatz nicht direkt auf das gesuchte orthogonale Funktionensystem. Es muss vorher stets eine **Transformation auf homogene Randbedingungen** durchgeführt werden. Ist $w(x,t)$ eine Funktion mit ausreichenden Differenzierbarkeitseigenschaften, die die **inhomogenen Randbedingungen** (RB) erfüllt, so gelingt eine solche Transformation mit dem folgenden Ansatz:

$$u(x,t) = v(x,t) + w(x,t).$$

Für die gesuchte Funktion $v(x,t)$ ergibt sich nun ein Anfangs–Randwert–Problem mit **homogenen** Randbedingungen. Das Auffinden einer geeigneten Funktion $w(x,t)$ gelingt oft mit einem (in der Variablen x) polynomialen Ansatz

$$w(x,t) = a(t)x^2 + b(t)x + c(t). \tag{7.11}$$

Treten die partiellen Ableitungen u_x in den Randbedingungen (RB) **nicht** auf, so kann stets $a(t) = 0$ gesetzt werden. Im vorliegenden BSP. führt der Ansatz (7.11) auf die Bedingungen $w(0,t) = c(t) \stackrel{!}{=} \varphi_1(t)$ und $w(L,t) = L\, b(t) + \varphi_1(t) \stackrel{!}{=} \varphi_2(t)$, also

$$\boxed{w(x,t) = \frac{x}{L} \left(\varphi_2(t) - \varphi_1(t)\right) + \varphi_1(t).} \tag{7.12}$$

Wird jetzt $u(x,t) = v(x,t) + w(x,t)$ in das Ausgangsproblem eingesetzt, so resultiert für $v(x,t)$ das folgende Anfangs–Randwert–Problem, wenn wir $\varphi_1,\, \varphi_2 \in C^1(\mathbf{R}_+)$ voraussetzen:

$$\begin{cases} v_t - k\, v_{xx} = g(x,t) - \frac{x}{L} \left(\dot\varphi_2(t) - \dot\varphi_1(t)\right) - \dot\varphi_1(t) =: \tilde g(x,t) & : \ 0 < x < L, \ \ t > 0, & \text{(DGl)} \\ v(0,t) = v(L,t) = 0 & : \ t \geq 0, & \text{(RB)} \\ v(x,0) = f(x) - \frac{x}{L} \left(\varphi_2(0) - \varphi_1(0)\right) - \varphi_1(0) =: \tilde f(x) & : \ 0 \leq x \leq L. & \text{(AB)} \end{cases}$$

Dieses ARWP ist wieder vom Problemtyp des BSPs.(17.7.4). Wir wollen konkret werden und betrachten das folgende

$$\text{(ARWP)} \quad \begin{cases} u_t - k\, u_{xx} = \frac{x}{L} T_1 \cos t & : \ 0 < x < L, \ \ t > 0, & \text{(DGl)} \\ u(0,t) = T_0 e^{-t}, \quad u(L,t) = T_1 \sin t & : \ t \geq 0, & \text{(RB)} \\ u(x,0) = x - T_0(\frac{x}{L} - 1) & : \ 0 \leq x \leq L. & \text{(AB)} \end{cases}$$

Setzen wir hier also

$$u(x,t) = v(x,t) + (1 - \frac{x}{L}) T_0 e^{-t} + \frac{x}{L} T_1 \sin t,$$

so folgt für die neue Funktion $v(x,t)$ das folgende Anfangs–Randwert–Problem:

$$\begin{cases} v_t - k\,v_{xx} = \dfrac{x}{L}\,T_1\cos t - \dfrac{x}{L}\,(T_1\cos t + T_0 e^{-t}) + T_0 e^{-t} = (1-\dfrac{x}{L})T_0 e^{-t} & : 0<x<L,\ \ t>0, \qquad \text{(DGl)} \\ v(0,t) = v(L,t) = 0 & : t \geq 0, \qquad \text{(RB)} \\ v(x,0) = x - T_0(\dfrac{x}{L} - 1) - \dfrac{x}{L}(0 - T_0) - T_0 = x & : 0 \leq x \leq L. \qquad \text{(AB)} \end{cases}$$

Dieses Problem haben wir bereits in BSP. (17.7.4) gelöst. Wir erhalten also aus der Lösung (7.10):

$$\boxed{u(x,t) = 2L \sum_{n=1}^{\infty} \frac{1}{n\pi}\Big(\frac{2T_0}{L}\,\frac{e^{-t} - e^{-\left(\frac{n\pi}{L}\right)^2 kt}}{k\left(\frac{n\pi}{L}\right)^2 - 1} - (-1)^n e^{-\left(\frac{n\pi}{L}\right)^2 kt}\Big)\sin\frac{n\pi x}{L} + (1-\frac{x}{L})T_0 e^{-t} + \frac{x}{L}\,T_1 \sin t.}$$

Eine von der Produktform $u(x,t) = T(t)\cdot X(x)$ verschiedene Lösungsklasse der eindimensionalen Wärmeleitungsgleichung bilden die sogenannten **Wärmepole**

$$\boxed{w_1(x,\xi;t,\tau) := \frac{1}{\sqrt{4\pi(t-\tau)}}\,\exp\Big(-\frac{(x-\xi)^2}{4k(t-\tau)}\Big), \qquad t > \tau \in \mathbf{R}, \quad x,\,\xi \in \mathbf{R}.} \tag{7.13}$$

In der Tat, es gilt

$$w_t = k w_{xx} = \frac{1}{\sqrt{4\pi}}\,\exp\Big(-\frac{(x-\xi)^2}{4k(t-\tau)}\Big)\Big(\frac{(x-\xi)^2}{4k(t-\tau)^{5/2}} - \frac{1}{2(t-\tau)^{3/2}}\Big), \qquad t>\tau, \quad x,\,\xi \in \mathbf{R},$$

so dass die Wärmepole $w_1(x,\xi;t,\tau)$ für $t>\tau$ und für $x \in \mathbf{R}$ Lösungen der eindimensionalen homogenen Wärmeleitungsgleichung sind. Ihr Name begründet sich aus den folgenden Eigenschaften:

$$\lim_{t\to\tau+0} w_1(x,\xi;t,\tau) = \begin{cases} 0 & : \ \xi \neq x, \\ +\infty & : \ \xi = x, \end{cases} \qquad \lim_{|x|\to\infty} w_1(x,\xi;t,\tau) = 0 \ \ \forall\, t > \tau,$$

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{k}}\, w_1(x,\xi;t,\tau)\,d\xi = 1 \ \ \forall\, t>\tau \ \ \forall\, x \in \mathbf{R}.$$

Diese Eigenschaften besagen, dass der Wärmepol $w_1(x,\xi;t,\tau)$ zum Zeitpunkt $t=\tau$ an der Stelle $x=\xi$ auf das durch die DGl (7.1) beschriebene wärmeleitende Medium aufgebracht wurde. Der Pol zerfließt mit der Zeit $\tau < t \to +\infty$; seine Gesamtintensität bleibt jedoch stets konstant $\sqrt{k}$.

Wird ein in x–Richtung ausgedehnter eindimensionaler Wärmeleiter betrachtet, der nach beiden Seiten unendlich lang ist, so ist es aus energetischen Gründen plausibel, dass die Temperaturverteilung $u(x,t)$ den **natürlichen Randbedingungen**

$$\lim_{|x|\to\infty} u(x,t) = 0$$

genügen muss. Diese Eigenschaft wird gerade von den Wärmepolen erfüllt. Man verwendet sie deshalb, um das folgende **Anfangswert–Problem** oder CAUCHY–**Problem** für den eindimensionalen, beidseitig unendlich langen Wärmeleiter zu lösen: Bestimme die Funktion $u = u(x,t)$ so, dass gilt:

$$\text{(CP)} \quad \begin{cases} u_t - k\,u_{xx} = 0 \ : \ x \in \mathbf{R}, \ \ t>0, & \text{(DGl)} \\ u(x,0) = f(x) \ : \ x \in \mathbf{R}. & \text{(AB)} \end{cases}$$

Ohne Beweis teilen wir das folgende Existenzergebnis mit:

Satz 17.16 *Es gelte $f \in C(\mathbf{R}) \cap L(\mathbf{R})$. Dann hat das* Cauchy*–Problem* (CP) *für den eindimensionalen, beidseitig unendlich langen Wärmeleiter die folgende Lösung*

$$\boxed{u(x,t) = \frac{1}{\sqrt{k}} \int\limits_{-\infty}^{+\infty} f(\xi) w_1(x,\xi;t,0)\, d\xi = \frac{1}{\sqrt{4\pi k t}} \int\limits_{-\infty}^{+\infty} f(\xi) \exp\Big(-\frac{(x-\xi)^2}{4kt}\Big)\, d\xi.} \tag{7.14}$$

Bemerkung 17.10 Das Integral (7.14) heißt **singuläres Integral von** Gauss–Weierstrass. Im n–dimensionalen Ortsraum wird das Cauchy–Problem (CP) durch die folgende Funktion gelöst: □

$$u(\vec{x},t) = (4\pi k t)^{-n/2} \int\limits_{\mathbf{R}^n} f(\vec{\xi}) \exp\Big(-\frac{\|\vec{x}-\vec{\xi}\|^2}{4kt}\Big)\, d\xi_1\, d\xi_2 \cdots d\xi_n.$$

17.8 Parameterintegrale

Wir haben in Satz 17.16 behauptet, dass das Cauchy–Problem für den eindimensionalen, beidseitig unendlich langen Wärmeleiter durch das singuläre Integral (7.14) von Gauss–Weierstrass gelöst wird, nämlich

$$u(x,t) = \frac{1}{\sqrt{k}} \int\limits_{-\infty}^{+\infty} f(\xi) w_1(x,\xi;t,0)\, d\xi, \quad x \in \mathbf{R}, \quad t \geq 0.$$

Im Integranden treten dabei die Variablen (x,t) als **Parameter** auf. Dass die Funktion $u(x,t)$ der eindimensionalen Wärmeleitungsgleichung $u_t - k\, u_{xx} = 0$ genügt, prüft man nach, indem man die Differentialgleichung

$$\frac{\partial}{\partial t} w_1(x,\xi;t,0) - k \frac{\partial^2}{\partial x^2} w_1(x,\xi;t,0) = 0$$

verifiziert. Dabei wird stillschweigend vorausgesetzt, dass die Vertauschung von Differentiation und uneigentlicher Integration erlaubt ist, dass also zum Beispiel gilt:

$$\frac{\partial}{\partial t} \int\limits_{-\infty}^{+\infty} f(\xi) w_1(x,\xi;t,0)\, d\xi = \int\limits_{-\infty}^{+\infty} f(\xi) \frac{\partial}{\partial t} w_1(x,\xi;t,0)\, d\xi.$$

Wir wollen uns mit der Rechtfertigung dieses Vertauschungsprozesses in dem einfacheren Fall auseinandersetzen, dass der Integrand nur von einem einzigen Parameter abhängt. Wir betrachten zunächst auch nur den Fall eines **eigentlichen** Riemann–Integrals.

Definition 17.9 *Die Funktion $f \in \mathrm{Abb}\,([a,b] \times [c,d], \mathbf{K})$ sei so gegeben, dass das folgende* Riemann*–Integral für alle Werte $y \in [c,d]$ existiere:*

$$\boxed{F(y) := \int_a^b f(x,y)\, dx, \quad y \in [c,d].} \tag{8.1}$$

Dann heiße (8.1) *ein* **Parameterintegral mit dem Parameter** y.

BSP. (17.8.1) Die Ausbreitung ebener Wellen in einem zweidimensionalen elastischen Medium wird durch die **zweidimensionale Wellengleichung**

$$\boxed{u_{tt} = c^2(u_{xx} + u_{yy}) \equiv c^2 \Delta_2 u, \qquad c > 0: \quad \text{Wellenausbreitungsgeschwindigkeit},} \tag{8.2}$$

beschrieben. Man kann hier wie bei der eindimensionalen Wärmeleitungsgleichung einen BERNOULLIschen **Separationsansatz** $u(x,y;t) = T(t) \cdot W(x,y)$ durchführen, der nach TdV auf die Beziehung

$$\frac{\ddot{T}}{c^2 T} = \frac{\Delta_2 W}{W} \stackrel{!}{=} -\lambda^2, \qquad \lambda \geq 0,$$

führt. Man erhält die beiden DGln

$$\ddot{T} + \lambda^2 c^2 T = 0, \qquad \Delta_2 W + \lambda^2 W = 0. \tag{8.3}$$

Im n–dimensionalen Ortsraum $\mathbf{R}^n$, $n \geq 2$, heißt die partielle Differentialgleichung

$$\Delta_n W + \lambda^2 W = 0$$

die HELMHOLTZ–**DGl**. Wir betrachten speziell das Problem der freien Schwingungen einer ebenen **Rechteck–Membran**, die im Ortsraum $\mathbf{R}^2$ den Rechteckbereich $[0,L] \times [0,M]$ einnimmt und deren Rand fest eingespannt ist. Es sollen keine äußeren Kräfte auf die Membran einwirken, während ihre transversale Anfangsauslenkung $u(x,y;0)$ und ihre Anfangsgeschwindigkeit $u_t(x,y;0)$ vorgegeben seien. In der mathematischen Formulierung führt die Bestimmung der resultierenden transversalen Auslenkung $u(x,y;t)$ auf das folgende Anfangs–Randwert–Problem:

$$\text{(ARWP)} \quad \begin{cases} u_{tt} - c^2 \Delta_2 u = 0 & : 0 < x < L,\ 0 < y < M,\ t > 0, \qquad \text{(DGl)} \\ u(0,y;t) = 0 = u(L,y;t) & : 0 \leq y \leq M,\ t \geq 0, \qquad (\text{RB})_x \\ u(x,0;t) = 0 = u(x,M;t) & : 0 \leq x \leq L,\ t \geq 0, \qquad (\text{RB})_y \\ u(x,y;0) = f(x,y),\ \ u_t(x,y;0) = h(x,y) & : 0 \leq x \leq L,\ 0 \leq y \leq M. \qquad \text{(AB)} \end{cases}$$

Der *erste* Separationsansatz $u(x,y;t) = T(t) \cdot W(x,y)$ führt auf die beiden DGln (8.3), wobei die DGl für $T(t)$ die folgende allgemeine Lösung hat:

$$T(t) = \begin{cases} a_0 + b_0 t & : \ \lambda = 0, \\ a_\lambda \cos \lambda c t + b_\lambda \sin \lambda c t & : \ \lambda > 0. \end{cases}$$

Die HELMHOLTZ–DGl separieren wir wieder mit einem Ansatz $W(x,y) := X(x) \cdot Y(y)$. Nach Einsetzen in die DGl und TdV resultiert:

$$\frac{X''}{X} = -\frac{Y'' + \lambda^2 Y}{Y} \stackrel{!}{=} -\mu^2, \qquad \mu \geq 0.$$

Das heißt, wir haben die beiden gewöhnlichen DGln

$$X'' + \mu^2 X = 0, \qquad Y'' + k^2 Y = 0, \quad k^2 := \lambda^2 - \mu^2, \tag{8.4}$$

zu integrieren. Deren allgemeine Lösungen lauten

$$X(x) = \begin{cases} \alpha_0 + \beta_0 x & : \ \mu = 0, \\ \alpha_\mu \cos \mu x + \beta_\mu \sin \mu x & : \ \mu > 0, \end{cases} \qquad Y(y) = \begin{cases} c_0 + d_0 y & : \ k = 0, \\ c_k \cos k y + d_k \sin k y & : \ k > 0. \end{cases}$$

- **Anpassen der Randbedingungen:**
  - Die Randbedingung $(\text{RB})_x$ ist sicher erfüllt für

    $$X(0) = X(L) = 0 \quad \Rightarrow \quad \alpha_0 = \beta_0 = \alpha_\mu = 0 \ \text{ sowie } \ \sin \mu L = 0 \quad \Rightarrow \quad \mu = \frac{n\pi}{L}, \ \ n \in \mathbf{N}.$$

  - Die Randbedingung $(\text{RB})_y$ ist sicher erfüllt für

    $$Y(0) = Y(M) = 0 \quad \Rightarrow \quad c_0 = d_0 = c_k = 0 \ \text{ sowie } \ \sin kM \quad \Rightarrow \quad k = \frac{m\pi}{M}, \ \ m \in \mathbf{N}.$$

Aus der Beziehung (8.4) erhalten wir nun die Separationskonstante

$$0 \neq \lambda_{mn}^2 = k^2 + \mu^2 = \pi^2 (\frac{n^2}{L^2} + \frac{m^2}{M^2}), \qquad m,\, n \in \mathbf{N}.$$

Somit liegt die folgende Schar partikulärer Lösungen vor, die bereits die geforderten Randbedingungen erfüllen:

$$u_{mn}(x,y;t) = (a_{mn} \cos \lambda_{mn} ct + b_{mn} \sin \lambda_{mn} ct) \sin \frac{n\pi x}{L} \cdot \sin \frac{m\pi y}{M}.$$

- **Superposition:**

$$u(x,y;t) = \sum_{m,n=1}^{\infty} (a_{mn} \cos \lambda_{mn} ct + b_{mn} \sin \lambda_{mn} ct) \sin \frac{n\pi x}{L} \cdot \sin \frac{m\pi y}{M}. \tag{8.5}$$

Hierin sind die Koeffizienten a_{mn}, b_{mn} durch

- **Anpassung der Anfangsbedingungen** (AB) zu bestimmen. Diese führen auf die FOURIER–**Doppelreihen**

$$\begin{aligned} u(x,y;0) &= \sum_{m,n=1}^{\infty} a_{mn} \sin \frac{n\pi x}{L} \cdot \sin \frac{m\pi y}{M} \sim f(x,y), \\ u_t(x,y;0) &= \sum_{m,n=1}^{\infty} b_{mn} \lambda_{mn} c \sin \frac{n\pi x}{L} \cdot \sin \frac{m\pi y}{M} \sim h(x,y), \end{aligned}$$

deren Koeffizienten in Analogie zum Fall der einfachen FOURIER–Reihen aus den Beziehungen

$$\left.\begin{aligned} a_{mn} &= \frac{4}{LM} \int_0^M \Big(\int_0^L f(\xi,\eta) \sin \frac{n\pi\xi}{L}\, d\xi \Big) \sin \frac{m\pi\eta}{M}\, d\eta, \\ b_{mn} &= \frac{4}{LM} \frac{1}{\lambda_{mn} c} \int_0^M \Big(\int_0^L h(\xi,\eta) \sin \frac{n\pi\xi}{L}\, d\xi \Big) \sin \frac{m\pi\eta}{M}\, d\eta \end{aligned}\right\} \tag{8.6}$$

folgen.

Somit ist die obige Anfangs–Randwert–Aufgabe für die ebene Rechteckmembran formal gelöst.

Ist zum Beispiel $f \in \mathrm{Abb}\,([0,L] \times [0,M], \mathbf{R})$ eine **stetige** Funktion, so existiert sicher in jedem Punkt $y \in [0,M]$ das Integral

$$F(y) := \frac{4}{LM} \int_0^L f(x,y) \sin \frac{n\pi x}{L}\, dx.$$

In den obigen Koeffizientenformeln (8.6) wurde stillschweigend vorausgesetzt, dass die Funktion $F(y) \sin \frac{m\pi y}{M}$ nun auch auf dem Intervall $[0,M]$ integrierbar ist, so dass die Koeffizienten a_{mn} sinnvoll erklärt sind. Ist diese Prämisse richtig?

Wegen der Gleichwertigkeit der beiden Variablen x und y sollte man erwarten, dass man die Koeffizienten a_{mn} in gleicher Weise durch

$$a_{mn} = \frac{4}{LM} \int_0^L \Big(\int_0^M f(x,y) \sin \frac{m\pi y}{M}\, dy \Big) \sin \frac{n\pi x}{L}\, dx$$

definieren kann. Ob dies ebenso richtig ist, beantworten wir in dem folgenden Satz.

Satz 17.17 *Die Funktion $f \in \mathrm{Abb}\,([a,b] \times [c,d], \mathbf{K})$ sei stetig. Dann gelten für das Parameterintegral*

$$F(y) := \int_a^b f(x,y)\, dx$$

die folgenden Aussagen:

(a) *Die Funktion* $F \in \mathrm{Abb}\,([c,d],\mathbf{K})$ *ist stetig.*

(b) *Ist* $f_y \in \mathrm{Abb}\,([a,b]\times[c,d],\mathbf{K})$ *stetig, so ist die Funktion* $F(y)$ *auf dem Intervall* $[c,d]$ *stetig differenzierbar und die Ableitung* $F'(y)$ *kann durch Differentiation unter dem Integral gewonnen werden:*

$$\boxed{F'(y) = \frac{d}{dy}\int_a^b f(x,y)\,dx = \int_a^b \frac{\partial}{\partial y} f(x,y)\,dx.}$$

(c) *Die Funktion* $F(y)$ *ist auf dem Intervall* $[c,d]$ *integrierbar, und es gilt*

$$\boxed{\int_c^d F(y)\,dy = \int_c^d \Big(\int_a^b f(x,y)\,dx\Big)\,dy = \int_a^b \Big(\int_c^d f(x,y)\,dy\Big)\,dx.}$$

Begründungen: (a) Die Funktion f ist auf der kompakten Teilmenge $[a,b]\times[c,d] \subset \mathbf{R}^2$ sogar gleichmäßig stetig:

$$\forall\,\epsilon>0\ \exists\,\delta=\delta(\epsilon)\ :\ |f(x,y)-f(x,y_0)|<\epsilon\ \ \forall\,x\in[a,b]\ \ \forall\,y,\,y_0\in[c,d]\ \text{mit}\ |y-y_0|<\delta.$$

Hieraus erschließen wir

$$|F(y)-F(y_0)| \le \int_a^b |f(x,y)-f(x,y_0)|\,dx \le \epsilon(b-a),$$

also die behauptete Stetigkeit.

(b) Aus der Taylor–Formel erhalten wir

$$f(x,y_0+h) = f(x,y_0) + f_y(x,y_0+\theta h)\cdot h, \qquad 0<\theta<1.$$

Somit gilt für das in (a) zu $\epsilon>0$ existierende δ, wenn wir $|h|<\delta$ wählen:

$$\Big|\frac{1}{h}\,(F(y_0+h)-F(y_0)) - \int_a^b f_y(x,y_0)\,dx\Big| \le \int_a^b \Big|\frac{1}{h}\,(f(x,y_0+h)-f(x,y_0)) - f_y(x,y_0)\Big|\,dx$$

$$\le \int_a^b |f_y(x,y_0+\theta h)-f_y(x,y_0)|\,dx \overset{\text{(a)}}{\le} \epsilon(b-a).$$

Also existiert der Grenzwert

$$\lim_{h\to 0}\frac{1}{h}\,(F(y_0+h)-F(y_0)) = \int_a^b f_y(x,y_0)\,dx\ \ \forall\,y_0\in[c,d].$$

(c) Wir setzen

$$G(y) := \int_c^y F(s)\,ds = \int_c^y \Big(\int_a^b f(x,s)\,dx\Big)\,ds.$$

Die Funktion $F(y)$ ist gemäß (a) stetig und somit auf beschränkten Intervallen integrierbar, so dass $G(y)$ stetig differenzierbar ist:

$$G'(y) = F(y) = \int_a^b f(x,y)\,dx.$$

Die Funktion

$$H(y) := \int_a^b \Big(\int_c^y f(x,s)\,ds\Big)\,dx$$

ist gemäß (b) ebenfalls differenzierbar, und es gilt

$$H'(y) = \int_a^b \Big(\frac{\partial}{\partial y}\int_c^y f(x,s)\,ds\Big)\,dx = \int_a^b f(x,y)\,dx = F(y).$$

Wir erschließen daraus $G(y) = H(y) + const$, und wegen $G(c) = H(c) = 0$ muss $const = 0$ gelten. Mit $G(d) = H(d)$ folgt schließlich die Behauptung. □

Bemerkung 17.11 Die Stetigkeitsvoraussetzungen in Satz 17.17 sind unangemessen hoch. Zum Beispiel erfüllt die Funktion $f(x,y) := e^y \cdot \sin\frac{1}{x}$ **nicht** die Stetigkeitsvoraussetzungen, wenn wir $[a,b] := [-1,1]$ wählen, obwohl die Funktion

$$F(y) = e^y \int_{-1}^{1} \sin\frac{1}{x}\,dx$$

stetig und auch stetig differenzierbar ist. Die folgende Definition ist für den vorliegenden Fall angemessener: □

Definition 17.10 *Es seien $X \subset \mathbf{R}$, $Y \subset \mathbf{R}$ Teilmengen. Eine Funktion $f \in \mathrm{Abb}\,(X \times Y, \mathbf{K})$ heiße* **gleichstetig in der Variablen** *y, wenn gilt:*

$$\boxed{\forall\, \epsilon > 0\ \exists\, \delta = \delta(\epsilon)\ :\ |f(x,y) - f(x,y_0)| < \epsilon\ \ \forall\, x \in X\ \ \forall\, y,\, y_0 \in Y\ \ mit\ \ |y - y_0| < \delta.}$$

Ganz analog wird die Gleichstetigkeit in der Variablen x erklärt. Offenbar ist die Funktion $f(x,y) := e^y \cdot \sin\frac{1}{x}$ gleichstetig in der Variablen $y \in [c,d] =: Y$ für alle $x \in \mathbf{R}$. Die Voraussetzungen des Satzes 17.17 können jetzt in der folgenden Weise abgeschwächt werden:

- Zu (a): Die Funktion $f \in \mathrm{Abb}\,([a,b] \times [c,d], \mathbf{K})$ ist gleichstetig in der Variablen y und das Integral $\int_a^b f(x,y)\,dx$ existiert in jedem Punkt $y \in [c,d]$.
- Zu (b): Die Funktion $f_y \in \mathrm{Abb}\,([a,b] \times [c,d], \mathbf{K})$ ist gleichstetig in der Variablen y und das Integral $\int_a^b f_y(x,y)\,dx$ existiert in jedem Punkt $y \in [c,d]$.
- Zu (c): Die Funktion $f \in \mathrm{Abb}\,([a,b] \times [c,d], \mathbf{K})$ ist gleichstetig in jeder der Variablen x und y.

.

BSP. (17.8.2) Wir betrachten das Parameterintegral

$$F(y) := \int_0^1 \ln(x^2 + y^2)\,dx, \qquad y \neq 0.$$

Die Ableitung $F'(y)$ kann ohne explizite Kenntnis der Funktion $F(y)$ berechnet werden, denn gemäß Satz 17.17(b) gilt ja

$$F'(y) = \frac{d}{dy}\int_0^1 \ln(x^2+y^2)\,dx = \int_0^1 \frac{\partial}{\partial y}\ln(x^2+y^2)\,dx = \int_0^1 \frac{2y}{x^2+y^2}\,dx = 2\mathrm{arc\,tan}_H\,\frac{1}{y}, \qquad y \neq 0.$$

BSP. (17.8.3) Zu festen Zahlen $r < s$ betrachten wir das Parameterintegral

$$F(y) := \cos y \int_r^s (\sin y)^{x^2}\,dx.$$

Man berechne das Integral $I := \int_0^{\pi/2} F(y)\,dy$. Man beachte, dass das Integral $\int_r^s (\sin y)^{x^2}\,dx$, $0 \le y \le \frac{\pi}{2}$, nicht elementar darstellbar ist. Da der Integrand jedoch auf der Menge $[r,s] \times [0, \frac{\pi}{2}]$ stetig ist, dürfen wir Satz 17.17(c) verwenden:

$$\begin{aligned} I = \int_0^{\pi/2} F(y)\,dy &= \int_0^{\pi/2} \Big(\int_r^s \cos y (\sin y)^{x^2}\,dx\Big)\,dy = \int_r^s \Big(\int_0^{\pi/2} \cos y (\sin y)^{x^2}\,dy\Big)\,dx \\ &= \int_r^s \frac{1}{1+x^2}\,(\sin y)^{x^2+1}\Big|_0^{\pi/2}\,dx = \int_r^s \frac{dx}{1+x^2} = \arctan s - \arctan r. \end{aligned}$$

Wir übertragen nun die Aussagen des Satzes 17.17 auf solche Parameterintegrale, deren Integrationsgrenzen variabel sind. Zunächst gilt die folgende Verallgemeinerung der Regel Satz 17.17(b):

Satz 17.18 (Leibnizsche Differentiationsregel)
Gegeben seien eine stetige Funktion $f \in \mathrm{Abb}\,([a,b]\times[c,d],\mathbf{K})$ *mit stetiger partieller Ableitung* $f_y \in \mathrm{Abb}\,([a,b]\times[c,d],\mathbf{K})$ *sowie stetig differenzierbare Funktionen* $\varphi_j : [c,d] \to [a,b]$. *Dann ist das Parameterintegral*

$$F(y) := \int_{\varphi_1(y)}^{\varphi_2(y)} f(x,y)\,dx, \qquad y \in [c,d],$$

stetig differenzierbar, und es gilt

$$\boxed{F'(y) = \frac{d}{dy}\int_{\varphi_1(y)}^{\varphi_2(y)} f(x,y)\,dx = \int_{\varphi_1(y)}^{\varphi_2(y)} f_y(x,y)\,dx + f(\varphi_2(y),y)\cdot\varphi_2'(y) - f(\varphi_1(y),y)\cdot\varphi_1'(y).} \tag{8.7}$$

Begründung: Die Funktion

$$G(y,u,v) := \int_u^v f(x,y)\,dx$$

ist gemäß Satz 17.17(b) nach der Variablen y stetig partiell differenzierbar sowie stetig differenzierbar nach den Integrationsgrenzen u und v. Somit ist G differenzierbar, und die Differenzierbarkeit überträgt sich auf die Funktion $F(y) = G(y,\varphi_1(y),\varphi_2(y))$. Wir berechnen $F'(y)$ mit Hilfe der Kettenregel, Satz 13.16(d):

$$\begin{aligned}\frac{d}{dy}F(y) &= \langle \mathrm{grad}\,G(y), (1,\varphi_1'(y),\varphi_2'(y))^T\rangle = G_y + G_u\varphi_1' + G_v\varphi_2' \\ &= \int_{\varphi_1(y)}^{\varphi_2(y)} f_y(x,y)\,dx + f(\varphi_2(y),y)\cdot\varphi_2'(y) - f(\varphi_1(y),y)\cdot\varphi_1'(y).\end{aligned}$$

Das ist die behauptete Ableitungsregel. □

BSP. (17.8.4) Wir suchen die lokalen Extrema der Funktion

$$F(y) := \int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} (1+|x|)\,dx, \qquad -2 \le y \le 2.$$

Dazu zerlegen wir die Funktion $F(y)$ in die beiden Integrale

$$F(y) = \int_0^{\sqrt{4-y^2}} (1+x)\,dx + \int_{-\sqrt{4-y^2}}^{0} (1-x)\,dx, \qquad -2 < y < 2,$$

und wenden jeweils die Leibnizsche Differentiationsregel an:

$$F'(y) = \frac{1+\sqrt{4-y^2}}{\sqrt{4-y^2}}\,(-y) - \frac{1+\sqrt{4-y^2}}{-\sqrt{4-y^2}}\,(-y) = \frac{-2y(1+\sqrt{4-y^2})}{\sqrt{4-y^2}} \stackrel{!}{=} 0.$$

Dies gilt genau für $y_0 := 0$, und wir sehen, dass das Integrationsintervall bei $y_0 = 0$ am größten wird, so dass wegen der Positivität des Integranden ein **lokales Maximum** der Funktion $F(y)$ vorliegen muss. In den Punkten $y = \pm 2$ gilt $F(y) = 0$: Hier liegen Randminima der Funktion $F(y)$.

Sind $f \in \mathrm{Abb}\,([a,b] \times [c,d], \mathbf{K})$ und $\varphi_j \in \mathrm{Abb}\,([c,d],[a,b])$ stetige Funktionen, so ist auch die folgende Funktion stetig:

$$F(y) := \int\limits_{\varphi_1(y)}^{\varphi_2(y)} f(x,y)\,dx, \qquad y \in [c,d].$$

Somit existiert das Integral

$$\int\limits_c^d F(y)\,dy = \int\limits_c^d \Big(\int\limits_{\varphi_1(y)}^{\varphi_2(y)} f(x,y)\,dx \Big)\,dy. \tag{8.8}$$

Ganz analog können stetige Funktionen $\psi_j \in \mathrm{Abb}\,([a,b],[c,d])$ vorgegeben werden, und es kann die stetige Funktion

$$G(x) := \int\limits_{\psi_1(x)}^{\psi_2(x)} f(x,y)\,dy, \qquad x \in [a,b],$$

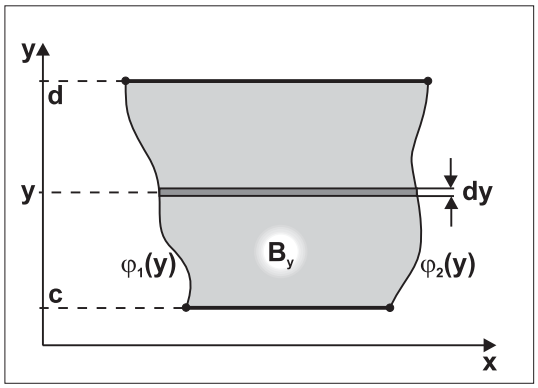
integriert werden. Es ist klar, dass das folgende Integral existiert:

$$\int\limits_a^b G(x)\,dx = \int\limits_a^b \Big(\int\limits_{\psi_1(x)}^{\psi_2(x)} f(x,y)\,dy \Big)\,dx. \tag{8.9}$$

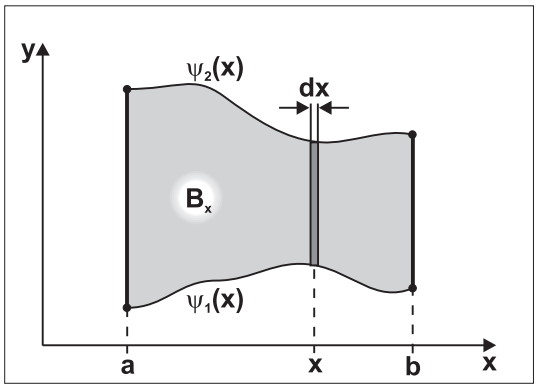
Definition 17.11 *Die Integrale* (8.8) *bzw.* (8.9) *heißen* **Doppelintegrale** *oder* **iterierte Integrale** *von f über den* **Normalbereich**

$$\begin{aligned} B_y &:= \{(x,y) : \varphi_1(y) \le x \le \varphi_2(y),\ \ c \le y \le d\} \quad bzw. \\ B_x &:= \{(x,y) : \psi_1(x) \le y \le \psi_2(x),\ \ a \le x \le b\}. \end{aligned}$$

Dabei gehört zu einem Normalbereich die Voraussetzung $\varphi_1(y) \le \varphi_2(y)\ \forall\, y \in [c,d]$ bzw. $\psi_1(x) \le \psi_2(x)\ \forall\, x \in [a,b]$ sowie die Stetigkeit der Funktionen φ_j, ψ_j, $j = 1,2$.

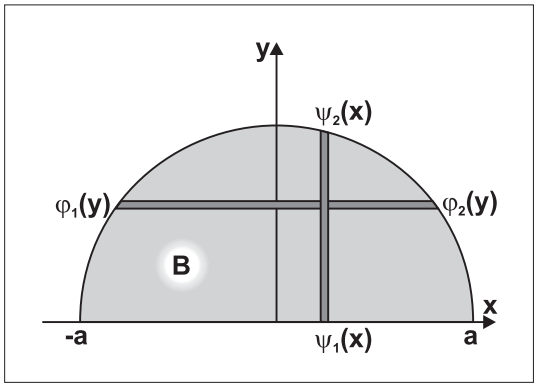


Normalbereich B_y: Jede Parallele zur x–Achse (mit höchstens zwei Ausnahmen) schneidet den Rand ∂B_y in höchstens zwei Punkten.

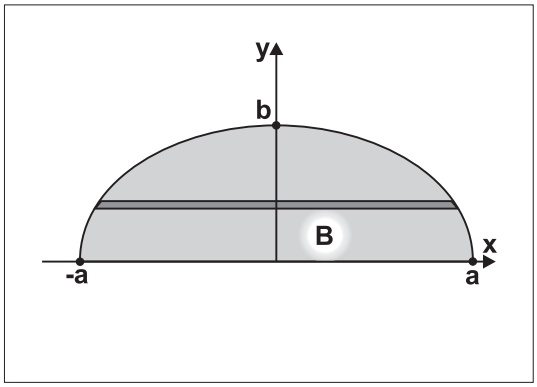


Normalbereich B_x: Jede Parallele zur y–Achse (mit höchstens zwei Ausnahmen) schneidet den Rand ∂B_x in höchstens zwei Punkten.

BSP. (17.8.5) Man berechne die beiden iterierten Integrale der Funktion $f(x,y) := x+y$ über den unten skizzierten Halbkreis B.



Integrationsbereich zum BSP. (17.8.5)



Integrationsbereich zum BSP. (17.8.6)

Lösung: Der Bereich B kann sowohl als Normalbereich in der Variablen y als auch als Normalbereich in der Variablen x aufgefasst werden: $B = B_y = B_x$ mit

$$\begin{aligned} B_y &= \{(x,y) : -\sqrt{a^2-y^2} \le x \le \sqrt{a^2-y^2},\ \ 0 \le y \le a\}, \\ B_x &= \{(x,y) : 0 \le y \le \sqrt{a^2-x^2},\ \ -a \le x \le a\}. \end{aligned}$$

Dementsprechend folgern wir:

$$\begin{aligned} I_1 &:= \int_0^a \Big(\int_{-\sqrt{a^2-y^2}}^{+\sqrt{a^2-y^2}} (x+y)\,dx \Big)\,dy = \int_0^a \big(\frac{x^2}{2} + xy\big)\Big|_{-\sqrt{a^2-y^2}}^{+\sqrt{a^2-y^2}}\,dy = \int_0^a 2y\sqrt{a^2-y^2}\,dy \\ &= -\frac{2}{3}\,(a^2-y^2)^{3/2}\Big|_0^a = \frac{2}{3}\,a^3. \end{aligned}$$

Andererseits gilt

$$\begin{aligned} I_2 &:= \int_{-a}^{+a} \Big(\int_0^{\sqrt{a^2-x^2}} (x+y)\,dy \Big)\,dx = \int_{-a}^{+a} \big(xy + \frac{y^2}{2}\big)\Big|_0^{\sqrt{a^2-x^2}}\,dx = \int_{-a}^{+a} \big(\underbrace{x\sqrt{a^2-x^2}}_{\text{ungerade}} + \frac{1}{2}\,\underbrace{(a^2-x^2)}_{\text{gerade}}\big)\,dx \\ &= \int_0^a (a^2-x^2)\,dx = a^3 - \frac{1}{3}\,a^3 = \frac{2}{3}\,a^3. \end{aligned}$$

Es gilt hier also $I_1 = I_2$, und dass dies kein Zufall ist, besagt die folgende Verallgemeinerung des Satzes 17.17(c).

Satz 17.19 *Der Bereich $B \subset \mathbf{R}^2$ sei sowohl bezüglich der Variablen x als auch der Variablen y ein* **Normalbereich**. *Ist die Funktion $f \in \mathrm{Abb}\,(B, \mathbf{K})$ stetig, so existieren die beiden iterierten Integrale* (8.8) *und* (8.9), *und es gilt Gleichheit:*

$$\boxed{\int_c^d \Big(\int_{\varphi_1(y)}^{\varphi_2(y)} f(x,y)\,dx \Big)\,dy = \int_a^b \Big(\int_{\psi_1(x)}^{\psi_2(x)} f(x,y)\,dy \Big)\,dx.} \tag{8.10}$$

BSP. (17.8.6) Zu berechnen ist das Doppelintegral der Funktion $f(x,y) := 1$ über die oben skizzierte Halbellipse B.

Lösung: Der Bereich B erfüllt die Voraussetzungen des Satzes 17.19, so dass zur Lösung der Aufgabe eines der beiden iterierten Integral in (8.10) berechnet werden muss, zum Beispiel das Doppelintegral über den Normalbereich

$$B_y := \{(x,y) \,:\, -a\sqrt{1-(y/b)^2} \le x \le a\sqrt{1-(y/b)^2},\ \ 0 \le y \le b\}.$$

Wir erhalten, indem wir in dem folgenden Integral die Substitution $y = b\sin\varphi,\ dy = b\cos\varphi\, d\varphi$ durchführen:

$$I \;=\; \int_0^b \Big(\int_{-a\sqrt{1-(y/b)^2}}^{a\sqrt{1-(y/b)^2}} 1\, dx \Big)\, dy = \int_0^b 2a\sqrt{1-(\frac{y}{b})^2}\, dy = 2ab \int_0^{\pi/2} \cos^2\varphi\, d\varphi = ab \int_0^{\pi/2} (1+\cos 2\varphi)\, d\varphi$$

$$= \frac{\pi ab}{2}.$$

Der Integralwert ist genau der **Flächeninhalt** der Halbellipse. Dies ist kein Zufall, wie wir in Kapitel 18 zeigen werden.

Wir betrachten abschließend **uneigentliche Parameterintegrale**, wie sie bereits in der Einleitung zu diesem Abschnitt aufgetreten waren. Das Verhalten solcher Integrale wird weitestgehend von dem Begriff der **gleichmäßigen Konvergenz** geprägt, ähnlich dem Verhalten von Funktionenfolgen und –reihen. Wir diskutieren hier nur solche uneigentlichen Integrale, die sich über **unbeschränkte** Integrationsintervalle erstrecken.

Definition 17.12 *Zu gegebener Funktion $f \in \mathrm{Abb}\,([a,\infty)\times[c,d],\mathbf{K})$ existiere für jedes feste $y \in [c,d]$ und für jede Zahl $b > a$ das Parameterintegral $\int_a^b f(x,y)\,dx$. Dann heiße*

$$\int_a^\infty f(x,y)\,dx := \lim_{b\to\infty} \int_a^b f(x,y)\,dx \tag{8.11}$$

ein **uneigentliches Parameterintegral** *mit dem Parameter y. Existiert der obige Grenzwert, so heiße das uneigentliche Parameterintegral* (8.11) konvergent*, und man setzt*

$$F(y) = \int_a^\infty f(x,y)\,dx, \qquad y \in [c,d].$$

Andernfalls heiße das Intgral (8.11) **divergent**. *Das uneigentliche Parameterintegral* (8.11) *heiße auf dem Intervall $[c,d]$* **gleichmäßig konvergent** *gegen $F(y)$, wenn es für alle $\epsilon > 0$ eine Zahl $B_0 = B_0(\epsilon)$ gibt, so dass gilt:*

$$\boxed{\Big|F(y) - \int_a^b f(x,y)\,dx\Big| = \Big|\int_b^\infty f(x,y)\,dx\Big| < \epsilon \ \ \forall\, b > B_0 \ \ \forall\, y \in [c,d].} \tag{8.12}$$

BSP. (17.8.7) Wir betrachten für festes $\alpha > 0$ das uneigentliche Parameterintegral

$$F(y) := \int_0^\infty e^{-\alpha x} \frac{\sin xy}{x}\, dx, \qquad y \in \mathbf{R}.$$

Für $x > 1$ haben wir $|\frac{\sin xy}{x}| < 1$, und somit

$$\Big|F(y) - \int_0^b e^{-\alpha x}\frac{\sin xy}{x}\, dx\Big| \le \int_b^\infty e^{-\alpha x}\, dx = \frac{1}{\alpha}\, e^{-b\alpha} < \epsilon,$$

für alle $b > B_0$, wenn B_0 gemäß $e^{-B_0\alpha} := \epsilon\alpha$ fixiert wird. Das uneigentliche Parameterintegral ist also gleichmäßig konvergent.

Ganz ähnlich wie bei Funktionenreihen gilt auch bei uneigentlichen Parameterintegralen ein Weierstrass–Kriterium zum Nachweis der gleichmäßigen Konvergenz:

Satz 17.20 (WEIERSTRASS–**Kriterium**)
Für gegebene Funktionen $f \in \mathrm{Abb}\,([a,\infty)\times[c,d],\mathbf{K})$ *und* $\varphi \in \mathrm{Abb}\,([a,\infty),[0,\infty))$ *gelte*

$$\text{(i)}\quad |f(x,y)| \le \varphi(x)\ \ \forall\, y\in[c,d], \qquad \text{(ii)}\quad \int_a^\infty \varphi(x)\,dx \text{ existiert.}$$

Dann ist das uneigentliche Parameterintegral $\int\limits_a^\infty f(x,y)\,dx$ **absolut** *und* **gleichmäßig** *konvergent.*

Begründung: Wegen (ii) existiert zu $\epsilon>0$ eine Zahl $B_0(\epsilon)$ mit $\int\limits_b^\infty \varphi(x)\,dx<\epsilon$ für alle $b>B_0$. Und hieraus folgern wir

$$\Big|F(y)-\int_a^b f(x,y)\,dx\Big| \le \int_b^\infty |f(x,y)|\,dx \overset{\text{(i)}}{\le} \int_b^\infty \varphi(x)\,dx<\epsilon\ \ \forall\, b>B_0\ \ \forall\, y\in[c,d].$$

Das war zu zeigen. □

BSP. (17.8.8) Zu festem $\alpha>0$ betrachten wir das uneigentliche Parameterintegral

$$F(y) := \int_0^\infty \frac{1-\cos xy}{x}\,e^{-\alpha x}\,dx.$$

Das Integral ist wegen $\frac{1-\cos xy}{x} = \frac{1}{2}xy^2+\mathcal{O}(x^2y^3)$ $(x\to 0)$ nur an der oberen Integrationsgrenze uneigentlich. Wir zerlegen $F(y)$ in die zwei Teilintegrale

$$F(y)=\int_0^1 \frac{1-\cos xy}{x}\,e^{-\alpha x}\,dx+\int_1^\infty \frac{1-\cos xy}{x}\,e^{-\alpha x}\,dx.$$

Für $x\ge 1$ gilt nun

$$\Big|\frac{1-\cos xy}{x}e^{-\alpha x}\Big| \le 2e^{-\alpha x} =: \varphi(x), \qquad \int_1^\infty \varphi(x)\,dx=\frac{2}{\alpha}\,e^{-\alpha}<\infty.$$

Also ist das uneigentliche Parameterintegral gemäß Satz 17.20 für alle $y\in\mathbf{R}$ absolut und gleichmäßig konvergent.

Satz 17.21 *Für die gegebene stetige Funktion* $f\in\mathrm{Abb}\,([a,\infty)\times[c,d],\mathbf{K})$ *konvergiere das uneigentliche Parameterintegral* $F(y):=\int\limits_a^\infty f(x,y)\,dx$ *auf dem Intervall* $[c,d]$ *gleichmäßig. Dann gelten die folgenden Aussagen:*

(a) *Die Funktion* $F\in\mathrm{Abb}\,([c,d],\mathbf{K})$ *ist stetig.*

(b) *Ist* $f_y\in\mathrm{Abb}\,([a,\infty)\times[c,d],\mathbf{K})$ *stetig und konvergiert das uneigentliche Parameterintegral* $\int\limits_a^\infty f_y(x,y)\,dx$ *gleichmäßig auf dem Intervall* $[c,d]$*, so folgt bei nur punktweiser Konvergenz des uneigentlichen Integrals* $F(y)=\int\limits_a^\infty f(x,y)\,dx$ *für jeden Punkt* $y\in[c,d]$ *die Ableitungsregel*

$$\boxed{F'(y)=\frac{d}{dy}\int_a^\infty f(x,y)\,dx=\int_a^\infty \frac{\partial}{\partial y}f(x,y)\,dx.}$$

(c) *Die Funktion* $F(y)$ *ist auf dem Intervall* $[c,d]$ *integrierbar, und es gilt*

$$\boxed{\int_c^d F(y)\,dy=\int_c^d\Big(\int_a^\infty f(x,y)\,dx\Big)\,dy=\int_a^\infty\Big(\int_c^d f(x,y)\,dy\Big)\,dx.}$$

Begründungen: Wir zeigen nur (a); der Rest folgt mit ganz analogen Schlüssen. Es sei nun $a =: a_0 < a_1 < a_2 < \cdots < a_n < \cdots$ eine monoton wachsende Folge mit $\lim\limits_{n\to\infty} a_n = +\infty$. Setzt man

$$u_n(y) := \int_{a_n}^{a_{n+1}} f(x,y)\,dx, \qquad n \in \mathbf{N}_0,$$

so ist $u_n \in \mathrm{Abb}\,([c,d],\mathbf{K})$ gemäß Satz 17.17(a) eine stetige Funktion. Die Funktionenreihe $F(y) = \sum\limits_{n=0}^{\infty} u_n(y)$ konvergiert nach Voraussetzung gleichmäßig, so dass die Stetigkeit der Grenzfunktion F aus Satz 9.4(b) folgt. □

BSP. (17.8.9) Wir betrachten das uneigentliche Parameterintegral

$$I_1(x,y) := \int_0^{\infty} e^{-\xi x}\cos\xi y\,d\xi, \qquad 0 < \epsilon \le x, \quad y \in \mathbf{R}.$$

Wegen

$$|e^{-\xi x}\cos\xi y| \le e^{-\epsilon\xi} \;\; \forall\, x \ge \epsilon \;\; \forall y \in \mathbf{R}, \qquad \int_0^{\infty} e^{-\epsilon\xi}\,d\xi = \frac{1}{\epsilon} < \infty,$$

konvergiert $I_1(x,y)$ gleichmäßig bezüglich $(x,y) \in G_\epsilon := [\epsilon,\infty)\times\mathbf{R}$. Somit ist die Funktion $I_1(x,y)$ auf der Menge G_ϵ stetig, und wir dürfen bezüglich y integrieren:

$$\int_0^y I_1(x,\eta)\,d\eta \overset{\text{Satz 17.21(c)}}{=} \int_0^{\infty} e^{-\xi x}\Big(\int_0^y \cos\xi\eta\,d\eta\Big)\,d\xi = \int_0^{\infty} e^{-\xi x}\frac{\sin\xi y}{\xi}\,d\xi =: I_2(x,y).$$

Andererseits folgt wegen $\cos\xi y = \frac{1}{2}(e^{i\xi y} + e^{-i\xi y})$ durch direkte Integration:

$$I_1(x,y) = \frac{1}{2}\int_0^{\infty}(e^{-\xi(x+iy)} + e^{-\xi(x-iy)})\,d\xi = \frac{x}{x^2+y^2}, \qquad (x,y) \in G_\epsilon.$$

Demzufolge gilt

$$I_2(x,y) = \int_0^y \frac{x\,d\eta}{x^2+\eta^2} = \operatorname{arc\,tan}_H\frac{y}{x} = \int_0^{\infty} e^{-\xi x}\frac{\sin\xi y}{\xi}\,d\xi, \qquad (x,y) \in G_\epsilon.$$

Wir haben bereits in BSP. (17.8.7) gezeigt, dass das uneigentliche Parameterintegral $I_2(x,y)$ auf der Menge G_ϵ gleichmäßig konvergiert. Somit können wir ein zweites Mal unter Verwendung von Satz 17.21(c) über y integrieren:

$$\begin{aligned}\int_0^y I_2(x,\eta)\,d\eta &= \int_0^{\infty} e^{-\xi x}\frac{1}{\xi}\Big(\int_0^y \sin\xi\eta\,d\eta\Big)\,d\xi = \int_0^{\infty} e^{-\xi x}\frac{1-\cos\xi y}{\xi^2}\,d\xi = \int_0^y \operatorname{arc\,tan}_H\frac{\eta}{x}\,d\eta \\ &= y\operatorname{arc\,tan}_H\frac{y}{x} - \frac{x}{2}\ln(x^2+y^2).\end{aligned}$$

Nun gilt

$$\int_0^{\infty} e^{-\xi x}\frac{1-\cos\xi y}{\xi^2}\,d\xi = \int_0^1 e^{-\xi x}\frac{1-\cos\xi y}{\xi^2}\,d\xi + \int_1^{\infty} e^{-\xi x}\frac{1-\cos\xi y}{\xi^2}\,d\xi.$$

Die Integranden sind jeweils stetige Funktionen; das erste Integral liefert eine stetige Funktion der Parameter $(x,y) \in \mathbf{R}^2$. Im zweiten Integral gilt

$$|e^{-\xi x}\frac{1-\cos\xi y}{\xi^2}| \le \frac{1}{\xi^2}, \qquad (x,y) \in G_0 := [0,\infty)\times\mathbf{R},$$

sowie $\int\limits_1^{\infty}\frac{d\xi}{\xi^2} = 1$. Also liegt gleichmässige Konvergenz auf der Menge G_0 vor: Das uneigentliche Parameterintegral

$$\int_0^{\infty} e^{-\xi x}\frac{1-\cos\xi y}{\xi^2}\,d\xi = y\operatorname{arc\,tan}_H\frac{y}{x} - \frac{x}{2}\ln(x^2+y^2)$$

ist stetig in $(x, y) \in G_0$. Wir bilden auf beiden Seiten den Limes $x \to 0+$ und erhalten:

$$\int_0^\infty \frac{1 - \cos \xi y}{\xi^2}\, d\xi = \frac{\pi y}{2} \cdot \operatorname{sign} y.$$

Durch partielle Integration gewinnt man

$$\int_0^\infty \frac{1 - \cos \xi y}{\xi^2}\, d\xi = \frac{\cos \xi y - 1}{\xi}\Big|_0^\infty + y \int_0^\infty \frac{\sin \xi y}{\xi}\, d\xi = y \int_0^\infty \frac{\sin \xi y}{\xi}\, d\xi.$$

Hieraus resultiert das uneigentliche Parameterintegral

$$\boxed{\int_0^\infty \frac{\sin \xi y}{\xi}\, d\xi = \frac{\pi}{2} \operatorname{sign} y, \qquad y \in \mathbf{R}.} \tag{8.13}$$

Bemerkung 17.12 Da die Funktion $F(y) := \frac{\pi}{2} \operatorname{sign} y$ an der Stelle $y = 0$ unstetig ist, kann das uneigentliche Parameterintegral $\int_0^\infty \frac{\sin \xi y}{\xi}\, d\xi$ auf Intervallen $[c, d]$ mit $0 \in [c, d]$ **nicht** gleichmässig konvergieren. □

Kapitel 18

Mehrfache Integrale

18.1 Messbare Punktmengen

Dem abgeschlossenen Intervall $I_1 := [a,b] \subset \mathbf{R}$ ordnet man sinnvollerweise eine **Länge** $m(I_1) := b-a$ zu. Aus elementargeometrischer Sicht ist es ebenso sinnvoll, dem Rechteck $I_2 := [a_1,b_1]\times[a_2,b_2] \subset \mathbf{R}^2$ einen **Flächeninhalt** $m(I_2) := (b_1-a_1)\,(b_2-a_2)$ zuzuordnen. Die Verallgemeinerung auf den n–dimensionalen Fall liegt auf der Hand:

Definition 18.1 *Für gegebene Vektoren $\vec{a}$, $\vec{b} \in \mathbf{R}^n$ sei das* **abgeschlossene n–dimensionale Intervall** I_n *durch*

$$I_n := [\vec{a},\vec{b}] = \{\vec{x} \in \mathbf{R}^n \,:\, a_j \le x_j \le b_j,\;\; j = 1,2,\ldots,n\}$$

definiert. Dann heiße die Zahl

$$\boxed{m(I_n) := \prod_{j=1}^{n}(b_j - a_j)}$$

der **Inhalt** *oder das* **Maß** *von I_n.*

Zu den abgeschlossenen n–dimensionalen Intervallen $I_n \subset \mathbf{R}^n$ zählen insbesondere alle Intervalle mit der Kantenlänge 1, deren Anfangspunkt $\vec{a}$ **ganzzahlige** Komponenten hat:

$$I_n^{(0)} := \{\vec{x} \in \mathbf{R}^n \,:\, a_j \le x_j \le a_j + 1,\;\; a_j \in \mathbf{Z},\;\; j = 1,2,\ldots,n\}.$$

Da die Menge $\mathbf{Z}$ der ganzen Zahlen abzählbar ist, bilden solche Intervalle offenbar eine abzählbare Menge. Bei festgehaltenem $\vec{a} \in \mathbf{Z}^n$ bezeichne $I_n^{(k)}$ das n–dimensionale Intervall mit der Kantenlänge $\frac{1}{2^k}$, welches durch k–malige Halbierung der Seitenlängen aus dem Intervall $I_n^{(0)}$ entsteht. Wir beschreiben nun, wie einer gegebenen Teilmenge $M \subset \mathbf{R}^n$ ein sinnvolles **Maß** zugeordnet werden kann. Wir setzen (vgl. die Skizze auf der folgenden Seite) für ein festes $k \in \mathbf{N}_0$:

$$A_k := \bigcup I_n^{(k)} \;\text{ mit }\; I_n^{(k)} \subset M, \qquad B_k := \bigcup I_n^{(k)} \;\text{ mit }\; I_n^{(k)} \cap M \neq \emptyset.$$

Nun gilt ganz offensichtlich $A_k \subset M \subset B_k$, und die Maße

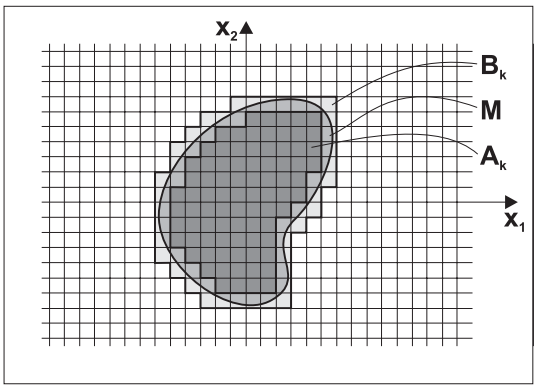
$$m(A_k) = \sum_{I_n^{(k)} \in A_k} m(I_n^{(k)}), \qquad m(B_k) = \sum_{I_n^{(k)} \in B_k} m(I_n^{(k)})$$

existieren. Durch Halbierung der Kantenlänge von $I_n^{(k)}$ entstehen Teilintervalle $I_n^{(k+1)}$, und somit Mengen A_{k+1}, B_{k+1}, für die gilt:

$$A_k \subset A_{k+1}, \quad B_{k+1} \subset B_k, \qquad m(A_k) \le m(A_{k+1}), \quad m(B_{k+1}) \le m(B_k).$$

Auf diese Weise erhält man monotone Folgen

$$(m(A_k))_{k\in\mathbf{N}} \text{ : monoton } \uparrow, \qquad (m(B_k))_{k\in\mathbf{N}} \text{ : monoton } \downarrow .$$



Zur Definition des Maßes einer Menge $M \subset \mathbf{R}^n$

Ist die Menge M **beschränkt**, so ist die Folge $m(A_k)$ nach oben beschränkt, während die Folge $m(B_k)$ nach unten beschränkt sein muss. Der Hauptsatz 3.3 über monotone Konvergenz liefert somit die Existenz der Grenzwerte $\lim\limits_{k\to\infty} m(A_k)$ sowie $\lim\limits_{k\to\infty} m(B_k)$.

Definition 18.2 *Es sei $M \subset \mathbf{R}^n$ eine beschränkte Menge, und es seien*

$$m_i(M) := \lim_{k\to\infty} m(A_k), \qquad m_a(M) := \lim_{k\to\infty} m(B_k)$$

die existierenden Grenzwerte. Dann heiße $m_i(M)$ bzw. $m_a(M)$ **inneres** *bzw.* **äußeres Maß** *der Menge M. Falls*

$$m_i(M) = m_a(M) =: m(M)$$

*gilt, so heiße $m(M)$ das n–***dimensionale R–(=**RIEMANN**) Maß** *von M. Die Menge M heiße dann* **R–messbar**. *Vereinbarungsgemäß gelte $m(\emptyset) = 0$.*

BSP. (18.1.1) Wir bestimmen das 2–dimensionale Maß der Strecke $M := \{(x_1, x_2) \in \mathbf{R}^2 \,:\, a \le x_1 \le b, \;\; x_2 = const\}$. Hier gilt offenkundig $A_k = \emptyset$ und somit $m(A_k) = 0$. Die Menge B_k enthält höchstens zwei Schichten von Teilintervallen $I_2^{(k)}$, und somit gilt $m(B_k) \le (b-a)\cdot 2\cdot \frac{1}{2^k} \to 0,\ k \to \infty$. Das heißt, wir haben $m_i(M) = m_a(M) = 0$.

Durch Verallgemeinerung dieses Beispiels gelangt man zu folgender Aussage:

Folgerung 18.1 *Jede rektifizierbare Parameterkurve in $\mathbf{R}^n$ (vgl. Definition 12.5) hat das n–dimensionale Maß Null. Jede beschränkte* **ebene** *Punktmenge hat das 3–dimensionale Maß Null.*

Es gibt auch Beispiele **nichtmessbarer** Mengen:

BSP. (18.1.2) Die Menge M sei in der folgenden Weise definiert:

$$M := \{(x_1, x_2) \in \mathbf{R}^2 \,:\, x_1 \in \mathbf{Q} \text{ und } 0 < x_1 < 1, \;\; x_2 \in \mathbf{R} \text{ und } 0 < x_2 < 1\}.$$

Da M keine inneren Punkte hat, gilt stets $m(A_k) = 0$, während wir $m(B_k) = 1$ haben. Hieraus folgern wir $0 = m_i(M) \neq m_a(M) = 1$, also Nichtmessbarkeit.

Wir geben hier ohne Beweis einige Eigenschaften messbarer Mengen an:

Satz 18.1 *Für gegebene Teilmengen M, M_1, $M_2 \subset \mathbf{R}^n$ gilt:*

(a) *Sind M_1 und M_2 messbar, und gilt $M_1^o \cap M_2^o = \emptyset$, so folgt*

$$\boxed{m(M_1 \cup M_2) = m(M_1) + m(M_2), \qquad \text{Additivität.}}$$

(b) *Sind M_1 und M_2 messbar, und gilt $M_2 \subset M_1$, so folgt*

$$\boxed{m(M_1 \setminus M_2) = m(M_1) - m(M_2).}$$

(c) *Die Menge M ist genau dann messbar, wenn $m(\partial M) = 0$ gilt.*

Aus Satz 18.1(c) ergibt sich unmittelbar:

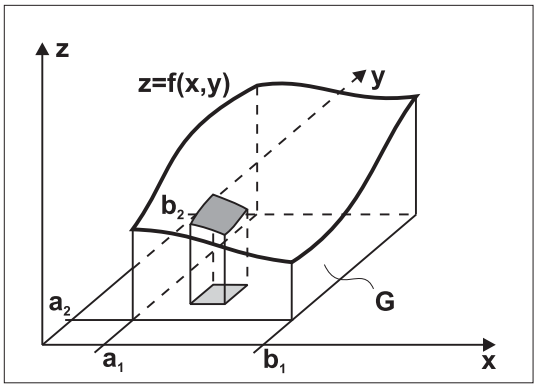
Folgerung 18.2 *Jede ebene Fläche, die von einer geschlossenen, rektifizierbaren doppelpunktfreien Parameterkurve berandet wird, ist messbar. Eine entsprechende Verallgemeinerung gilt in $\mathbf{R}^3$, usw.*

18.2 Ebene Bereichsintegrale

In Abschnitt 8.3 wurde das RIEMANN–Integral $\int_a^b f(x)\,dx$ einer Funktion $f \in \mathrm{Abb}\,(\mathbf{R}, \mathbf{R})$ geometrisch als Flächeninhalt derjenigen Fläche gedeutet, die zwischen dem Graphen $G(f)$ und der x–Achse liegt. Diese geometrische Deutung soll nun übertragen werden auf **Volumina prismatischer Körper**, deren Grundfläche einen Bereich G der (x, y)–Ebene ausfüllen und die "nach oben" durch eine Fläche $z = f(x, y)$, $(x, y) \in G$, begrenzt sind. Im einfachsten Fall sei G ein achsenparalleles ebenes abgeschlossenes Rechteck, das heißt, ein 2–dimensionales Intervall

$$G := \{(x, y) \in \mathbf{R}^2 \,:\, a_1 \le x \le b_1,\;\; a_2 \le y \le b_2\}$$

mit $m(G) = (b_1 - a_1)\,(b_2 - a_2)$.



Zweidimensionales Intervall als Grund–bereich eines prismatischen Körpers

Zu G definieren wir in Analogie zum eindimensionalen Fall eine **endliche Zerlegung**

$$Z_{nm} := \{G_{jk} \,:\, 1 \le j \le n,\;\; 1 \le k \le m\}$$

mit folgenden Eigenschaften

(Z1) $a_1 =: x_0 \le x_1 \le x_2 \le \cdots \le x_n := b_1, \qquad a_2 =: y_0 \le y_1 \le y_2 \le \cdots \le y_m := b_2.$

(Z2) $G_{jk} := I_j^x \times I_k^y$, wobei das Intervall I_j^x die Randpunkte x_{j-1} und x_j hat, während das Intervall I_k^y die Randpunkte y_{k-1} und y_k besitzt, wobei noch $I_j^x \neq \emptyset \neq I_k^y$ für $j = 1, 2, \ldots, n$ und $k = 1, 2, \ldots, m$ gelte.

(Z3) Für jedes Indexpaar $j \neq j'$ und $k \neq k'$ gelte

$$I_j^x \cap I_{j'}^x = \emptyset = I_k^y \cap I_{k'}^y, \qquad \bigcup_{j=1}^{n} I_j^x = [a_1, b_1], \quad \bigcup_{k=1}^{m} I_j^y = [a_2, b_2].$$

Es sei $m(G_{jk}) = (x_j - x_{j-1})\,(y_k - y_{k-1})$ das Maß der Menge G_{jk}, und es bezeichne schließlich

$$|Z_{nm}| := \max\{\sqrt{(x_j - x_{j-1})^2 + (y_k - y_{k-1})^2} \;:\; 1 \le j \le n, \;\; 1 \le k \le m\}$$

das **Feinheitsmaß** der endlichen Zerlegung Z_{nm}.

Sei nun $f \in \mathrm{Abb}\,(G, \mathbf{R})$ eine beschränkte Funktion. Bei gegebener endlicher Zerlegung Z_{nm} der Menge G sei $(\xi_j, \eta_k) \in G_{jk}$ ein beliebiger Zwischenpunkt. Dann heiße die Doppelsumme

$$S_{nm} := \sum_{j=1}^{n} \sum_{k=1}^{m} f(\xi_j, \eta_k)\, m(G_{jk}) \tag{2.1}$$

die der endlichen Zerlegung Z_{nm} zugeordnete RIEMANN**sche Näherungssumme**. Gilt $|Z_{nm}| \to 0$, so bleibt zu prüfen, ob die zugeordnete Folge (S_{nm}) der RIEMANNschen Näherungssummen für jede Wahl des Zwischenpunktes (ξ_j, η_k) gegen ein und denselben Grenzwert konvergiert.

Definition 18.3 *Die Funktion $f \in \mathrm{Abb}\,(G, \mathbf{R})$ sei auf dem Rechteckbereich $G := [a_1, b_1] \times [a_2, b_2]$ beschränkt. Konvergiert die Folge (2.1) der* RIEMANN*–Summen für jede Wahl von endlichen Zerlegungen $Z_{nm} := \{G_{jk} \;:\; 1 \le j \le n, \;\; 1 \le k \le m\}$ mit $|Z_{nm}| \to 0$ und für jede Wahl der Zwischenstellen $(\xi_j, \eta_k) \in G_{jk}$ gegen ein und denselben Grenzwert S, so heiße die Zahl S* **ebenes Bereichsintegral** *oder* **Flächenintegral** *der Funktion f über den Bereich G. In diesem Fall heiße f über dem Bereich G* RIEMANN**–integrierbar**, *kurz:* **R–integrierbar**, *und G heiße* **Integrationsbereich**. *Man schreibt dafür*

$$\boxed{S = \int\limits_G f\, dG \equiv \iint_G f(x, y)\, dx\, dy := \lim_{|Z_{nm}| \to 0} \sum_{j=1}^{n} \sum_{k=1}^{m} f(\xi_j, \eta_k)\, m(G_{jk}).} \tag{2.2}$$

Mit dieser Definition ist noch nichts darüber ausgesagt, für welche Funktionen f das Bereichsintegral existiert. Es sind jedoch alle stetigen Funktionen $f \in \mathrm{Abb}\,(G, \mathbf{R})$ R–integrierbar:

Satz 18.2 *Die Funktion $f \in \mathrm{Abb}\,(G, \mathbf{R})$ sei auf dem Rechteckbereich $G := [a_1, b_1] \times [a_2, b_2]$ stetig. Dann ist f über G R–integrierbar.*

Begründung: Die Funktion f ist auf der kompakten Teilmenge G sogar beschränkt und gleichmäßig stetig: Es existieren

$$\text{(i)} \quad \underline{M} := \min_{(x,y) \in G} f(x, y), \qquad \overline{M} := \max_{(x,y) \in G} f(x, y),$$

und es gilt

$$\text{(ii)} \quad \forall\, \epsilon > 0 \;\; \exists\, \delta = \delta(\epsilon) > 0 \;:\; |f(\vec{x}_1) - f(\vec{x}_2)| < \frac{\epsilon}{m(G)} \;\; \forall\, \vec{x}_1,\, \vec{x}_2 \in G \text{ mit } \|\vec{x}_1 - \vec{x}_2\| < \delta.$$

Sei nun $\epsilon > 0$ fest gewählt und dazu die Zahl $\delta > 0$ gemäß (ii) bestimmt. Sei $Z_{nm} = \{G_{jk} : 1 \le j \le n,\ \ 1 \le k \le m\}$ eine endliche Zerlegung von G mit $|Z_{nm}| < \delta$. Wir setzen

$$f(\vec{x}_{jk}) = \underline{M}_{jk} := \min_{(x,y)\in\overline{G}_{jk}} f(x,y), \qquad \overline{M}_{jk} := \max_{(x,y)\in\overline{G}_{jk}} f(x,y) = f(\vec{x}^{\,jk}),$$

und definieren dazu

$$\left.\begin{aligned} \underline{S}_{nm} &:= \sum_{j=1}^{n}\sum_{k=1}^{m} \underline{M}_{jk}\, m(G_{jk}) \ge \underline{M}\cdot m(G), \\ \overline{S}_{nm} &:= \sum_{j=1}^{n}\sum_{k=1}^{m} \overline{M}_{jk}\, m(G_{jk}) \le \overline{M}\cdot m(G). \end{aligned}\right\} \tag{2.3}$$

Dann gilt $\underline{S}_{nm} \le S_{nm} \le \overline{S}_{nm}$, und die Folge $\underline{S}_{nm}$ ist bei Verfeinerung der Zerlegung Z_{nm} monoton $\uparrow$, während die Folge $\overline{S}_{nm}$ monoton $\downarrow$. Wegen der Beschränktheit beider Folgen nach oben bzw. nach unten muss Konvergenz vorliegen. Wir haben somit

$$0 \le \overline{S}_{nm} - \underline{S}_{nm} = \sum_{j=1}^{n}\sum_{k=1}^{m} \left(f(\vec{x}^{\,jk}) - f(\vec{x}_{jk})\right) m(G_{jk}) \overset{\text{(ii)}}{\le} \frac{\epsilon}{m(G)} \sum_{j=1}^{n}\sum_{k=1}^{m} m(G_{jk}) = \epsilon,$$

und daraus folgt nach dem Einschließungskriterium

$$\lim_{|Z_{nm}|\to 0} \underline{S}_{nm} = \lim_{|Z_{nm}|\to 0} S_{nm} = \lim_{|Z_{nm}|\to 0} \overline{S}_{nm} =: S. \tag{2.4}$$

Dies ist die behauptete Integrierbarkeit. □

Die oben gegebene Definition ist offensichtlich zur direkten Berechnung eines ebenen Bereichsintegrals nicht sehr gut geeignet, da bereits in einfachen Fällen sehr komplizierte Rechnungen entstehen. Wir zeigen jetzt, dass die Berechnung eines ebenen Breichsintegrals $\int\limits_G f\, dG$ auf die Berechnung von **iterierten Integralen** zurückgeführt werden kann.

Dazu betrachten wir auf dem Rechteckbereich $G := [a_1, b_1] \times [a_2, b_2]$ eine stetige Funktion $f \in \mathrm{Abb}\,(G, \mathbf{R})$. Dann existiert das Parameterintegral

$$F(y) := \int_{a_1}^{b_1} f(x,y)\, dx,$$

und gemäß Satz 17.17(c) ist die Funktion $F(y)$ auf dem Intervall $[a_2, b_2]$ integrierbar. Sei nun Z_{nm} eine endliche Zerlegung des Bereichs G, $Z_{nm} = \{G_{jk} : 1 \le j \le n,\ \ 1 \le k \le m\}$. Wir setzen $m(G_{jk}) = |I_j^x| \cdot |I_k^y|$, worin $|I_j^x| := x_j - x_{j-1}$ und $|I_k^y| := y_k - y_{k-1}$ die Intervalllängen bezeichnen. Da die Funktion f integrierbar ist, konvergiert die Folge der RIEMANN–Summen S_{nm} aus (2.1). Wir schreiben die Relation (2.1) in der Form

$$S_{nm} = \sum_{k=1}^{m} \Big(\sum_{j=1}^{n} f(\xi_j, \eta_k)\, |I_j^x| \Big)\, |I_k^y|.$$

Wir führen die Summen $\underline{S}_{nm}$ und $\overline{S}_{nm}$ wieder gemäß (2.3) ein. Nun gilt offenbar

$$\underline{M}_{jk} \cdot |I_j^x| \le \int_{x_{j-1}}^{x_j} f(x, \eta_k)\, dx \le \overline{M}_{jk} \cdot |I_j^x|, \qquad j = 1, 2, \ldots, n,\ \ k = 1, 2, \ldots, m,$$

und daraus folgern wir

$$\underline{S}_{nm} \le \sum_{k=1}^{m} \Big(\sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} f(x, \eta_k)\, dx \Big)\, |I_k^y| = \sum_{k=1}^{m} F(\eta_k)\, |I_k^y| \le \overline{S}_{nm}.$$

Wegen (2.4) erhalten wir also im Limes $|Z_{nm}| \to 0$:

$$S = \int\limits_G f\, dG = \lim_{|Z_{nm}|\to 0} \sum_{k=1}^{m} F(\eta_k)\, |I_k^y| = \int_{a_2}^{b_2} \Big(\int_{a_1}^{b_1} f(x,y)\, dx \Big)\, dy.$$

Unter Verwendung von Satz 17.17(c) resultiert schließlich:

Satz 18.3 *Für eine auf dem Rechteckbereich* $G := [a_1, b_1] \times [a_2, b_2]$ *stetige Funktion* $f \in \mathrm{Abb}\,(G, \mathbf{R})$ *gilt*

$$\boxed{\int\limits_G f\,dG = \iint_G f(x,y)\,dx\,dy = \int_{a_2}^{b_2} \Big(\int_{a_1}^{b_1} f(x,y)\,dx \Big)\,dy = \int_{a_1}^{b_1} \Big(\int_{a_2}^{b_2} f(x,y)\,dy \Big)\,dx.}$$

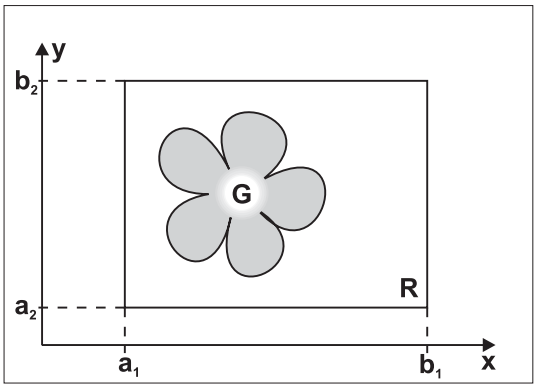
BSP. (18.2.1) Es seien $f(x,y) := x^y = e^{y \ln x}$ und $G := [0,1] \times [1,2]$ vorgegeben. Da die Funktion $f \in \mathrm{Abb}\,(G, \mathbf{R})$ stetig ist, folgern wir aus Satz 18.3:

$$\int\limits_G f\,dG = \int_1^2 \Big(\int_0^1 x^y\,dx \Big)\,dy = \int_1^2 \frac{x^{y+1}}{y+1}\Big|_0^1\,dy = \int_1^2 \frac{dy}{y+1} = \ln \frac{3}{2}.$$

Bemerkung 18.1 Ist $M \subset G := [a_1, b_1] \times [a_2, b_2]$ eine Menge vom 2–dimensionalen Maß Null: $m(M) = 0$, so erkennt man an der RIEMANN–Summe (2.1), dass die Funktionswerte $f(x,y)$ für $(x,y) \in M$ keinen Beitrag zum RIEMANN–Integral $\int\limits_G f\,dG$ liefern. Deshalb kann die Funktion f auf einer Menge vom 2–dimensionalen Maß Null ohne Änderung des Integralwertes abgeändert werden. Der Satz 18.2 kann deshalb in der folgenden Weise erweitert werden: □

Satz 18.4 *Die Funktion* $f \in \mathrm{Abb}\,(G, \mathbf{R})$ *sei auf dem Rechteckbereich* $G := [a_1, b_1] \times [a_2, b_2]$ *beschränkt und stetig mit Unstetigkeiten höchstens auf einer 2–dimensionalen Menge* M *vom Maße Null. Dann ist* f *über* G *R–integrierbar.*

Mit dem Ergebnis von Satz 18.4 sind wir jetzt in der Lage, ebene Bereichsintegrale $\int\limits_G f\,dG$ für solche Bereiche zu definieren, die keine achsenparallele Rechtecke sind.



Einschließung eines Bereichs G in ein achsenparalleles Rechteck R

Ist $G \subset \mathbf{R}^2$ eine messbare Menge, so folgt aus Satz 18.1 $m(\partial G) = 0$. Wir schließen G in ein achsenparalleles Rechteck R ein und setzen die gegebene Funktion $f \in \mathrm{Abb}\,(G, \mathbf{R})$ durch Null auf ganz R fort:

$$f_G(x,y) := \begin{cases} f(x,y) & : \ (x,y) \in G, \\ 0 & : \ (x,y) \in R \setminus G. \end{cases}$$

Ist die Funktion $f \in \mathrm{Abb}\,(G, \mathbf{R})$ stetig und beschränkt, so gilt dies auch für die Funktion $f_G \in \mathrm{Abb}\,(R, \mathbf{R})$ mit Ausnahme höchstens der Unstetigkeitsmenge ∂G vom Maße Null. Somit existiert gemäß Satz 18.4 das ebene Bereichsintegral $\int\limits_R f_G\,dR$, und wir definieren:

Definition 18.4 *Die Funktion $f \in \mathrm{Abb}\,(G, \mathbf{R})$ sei auf der messbaren Menge $G \subset \mathbf{R}^2$ beschränkt, und es sei $R \subset \mathbf{R}^2$ ein achsenparalleles Rechteck mit $G \subset R$. Sei*

$$f_G(x,y) := \begin{cases} f(x,y) & : \ (x,y) \in G, \\ 0 & : \ (x,y) \in R \setminus G. \end{cases}$$

Ist die Funktion f_G über dem Rechteck R integrierbar, so setzen wir $\int\limits_G f\,dG := \int\limits_R f_G\,dR$, und das Integral $\int\limits_G f\,dG$ heiße **ebenes Bereichsintegral** *von f über G.*

Nach dieser Definition können wir Satz 18.4 in der folgenden Form aufsschreiben:

Satz 18.5 *Die Funktion $f \in \mathrm{Abb}\,(G, \mathbf{R})$ habe auf dem beschränkten messbaren Bereich $G \subset \mathbf{R}^2$ höchstens auf einer Menge $M \subset G$ vom Maße Null Unstetigkeiten und sei sonst stetig. Dann existiert das ebene Bereichsintegral $\int\limits_G f\,dG$.*

Bemerkung 18.2 Gilt $f(x,y) := 1$ für alle $(x,y) \in G$, so ergibt sich aus der Konstruktion des ebenen Bereichsintegrals $\int\limits_G f\,dG$ sofort:

$$\boxed{A = \text{Fläche}\,(G) = m(G) = \int\limits_G 1\,dG = \int\limits_G dG.} \tag{2.5}$$

Wir nennen die Größe dG auch ein **Flächenelement**. □

Für ebene Bereichsintegrale gelten Rechenregeln ganz analog zu den Rechenregeln des eindimensionalen Riemann–Integrals:

Satz 18.6 *Auf einem beschränkten messbaren Bereich $G \subset \mathbf{R}^2$ seien integrierbare Funktionen $f, g \in \mathrm{Abb}\,(G, \mathbf{R})$ gegeben. Dann gilt:*

$$\boxed{\int\limits_G (\lambda\, f + \mu\, g)\,dG = \lambda \int\limits_G f\,dG + \mu \int\limits_G g\,dG \ \ \forall\ \lambda,\, \mu \in \mathbf{R}, \quad \text{Linearität},} \tag{2.6}$$

$$\boxed{f(x,y) \le g(x,y) \ \ \forall\ (x,y) \in G \quad \Rightarrow \quad \int\limits_G f\,dG \le \int\limits_G g\,dG,} \tag{2.7}$$

$$\boxed{\Big|\int\limits_G f\,dG\Big| \le \int\limits_G |f|\,dG,} \tag{2.8}$$

$$\boxed{G = G_1 \cup G_2 \ \ \textit{und} \ \ G_1^o \cap G_2^o = \emptyset \quad \Rightarrow \quad \int\limits_G f\,dG = \int\limits_{G_1} f\,dG + \int\limits_{G_2} f\,dG.} \tag{2.9}$$

Ist schließlich $f \ge 0$ und $\int\limits_G f\,dG = 0$, so kann $f \ne 0$ höchstens auf einer Menge $M \subset G$ vom Maße $m(M) = 0$ gelten.

Die Berechnung des ebenen Bereichsintegrals $\int\limits_G f\,dG$ wird an Hand seiner Definition in der Praxis kaum durchführbar sein. Aus diesem Grund soll die Berechnung von $\int\limits_G f\,dG$ wie bei der Vorgabe von Rechteckbereichen auf die Berechnung von **iterierten Integralen** zurückgeführt werden. Eine solche Rückführung gelingt stets dann, wenn G ein **Normalbereich** bezüglich einer der beiden Variablen x oder y ist.

Satz 18.7 *Gegeben sei eine stetige Funktion* $f \in \mathrm{Abb}\,(G, \mathbf{R})$ *auf einem beschränkten messbaren Bereich* $G \subset \mathbf{R}^2$.

(a) *Ist* $G = G_x$ *ein* **Normalbereich** *bezüglich der Variablen* x*:*

$$G_x = \{(x,y) \in \mathbf{R}^2 \,:\, g_1(x) \le y \le g_2(x),\ \ a \le x \le b\},$$

so gilt

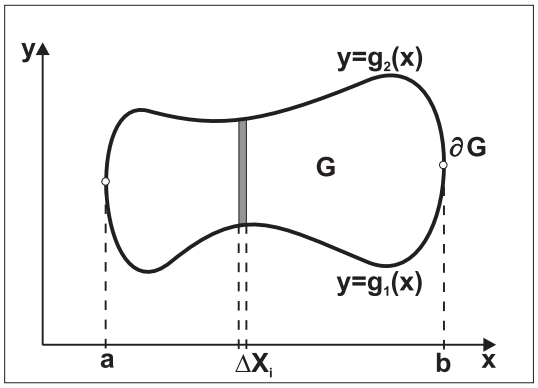
$$\boxed{\int\limits_G f\,dG = \int\limits_a^b \Big(\int\limits_{g_1(x)}^{g_2(x)} f(x,y)\,dy \Big)\,dx.} \tag{2.10}$$

(b) *Ist* $G = G_y$ *ein* **Normalbereich** *bezüglich der Variablen* y*:*

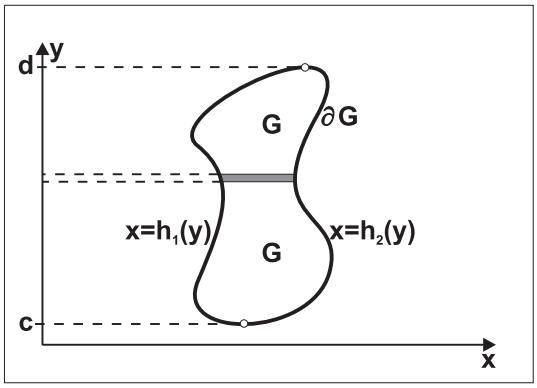
$$G_y = \{(x,y) \in \mathbf{R}^2 \,:\, h_1(y) \le x \le h_2(y),\ \ c \le y \le d\},$$

so gilt

$$\boxed{\int\limits_G f\,dG = \int\limits_c^d \Big(\int\limits_{h_1(y)}^{h_2(y)} f(x,y)\,dx \Big)\,dy.} \tag{2.11}$$



Berechnung von $\int\limits_G f\,dG$**, wenn** G **ein Normalbereich bezüglich** x **ist**



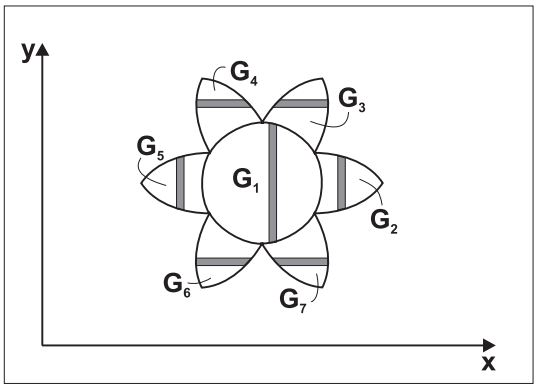
Berechnung von $\int\limits_G f\,dG$**, wenn** G **ein Normalbereich bezüglich** y **ist**

Man erhält eine *Begründung* ganz analog wie zum Satz 18.3. □

Bemerkung 18.3 Lässt sich die beschränkte messbare Menge $G \subset \mathbf{R}^2$ in **endlich viele** Normalbereiche $G_1, G_2, \ldots, G_m$ disjunkt zerlegen: $G_j^o \cap G_k^o = \emptyset$ für $j \ne k$, so folgt aus Satz 18.6:

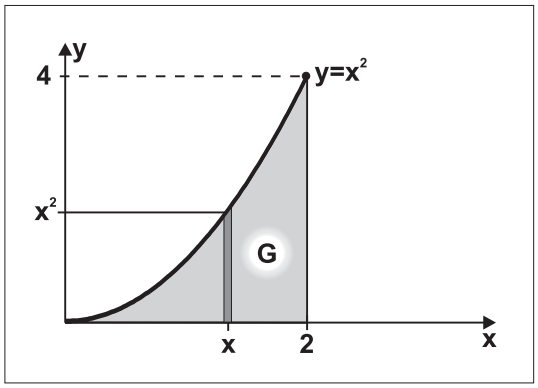
$$\int\limits_G f\,dG = \sum_{j=1}^{m} \int\limits_{G_j} f\,dG,$$

wobei jedes Integral auf der rechten Seite in der Form (2.10) oder (2.11) ausgewertet werden kann. □

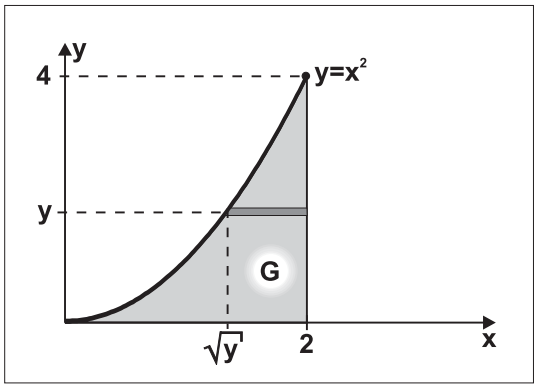


Zerlegung von G in Normalbereiche

BSP. (18.2.2) Gesucht ist das ebene Bereichsintegral der Funktion $f(x,y) := x^2 + y^2$ über dem Normalbereich $G := \{(x,y) \in \mathbf{R}^2 \,:\, 0 \le y \le x^2,\ \ 0 \le x \le 2\}$.



G als Normalbereich bezüglich x



G als Normalbereich bezüglich y

Lösung: (a) In der angegebenen Form ist $G = G_x$ ein Normalbereich bezüglich der Variablen x. Wir erhalten deshalb gemäß (2.10)

$$\int\limits_G f\,dG = \int\limits_0^2 \Big(\int\limits_0^{x^2} (x^2+y^2)\,dy \Big)\,dx = \int_0^2 (x^2 y + \frac{1}{3}\,y^3)\Big|_0^{x^2}\,dx = \int_0^2 (x^4 + \frac{1}{3}\,x^6)\,dx$$

$$= (\frac{1}{5}\,x^5 + \frac{1}{21}\,x^7)\Big|_0^2 = \frac{1312}{105}.$$

(b) Wir können G als Normalbereich $G_y = \{(x,y) \in \mathbf{R}^2 \,:\, \sqrt{y} \le x \le 2,\ \ 0 \le y \le 4\}$ bezüglich der Variablen y deuten und nun das ebene Bereichsintegral gemäß (2.11) auswerten:

$$\int\limits_G f\,dG = \int\limits_0^4 \Big(\int\limits_{\sqrt{y}}^2 (x^2+y^2)\,dx \Big)\,dy = \int_0^4 (\frac{1}{3}\,x^3 + y^2 x)\Big|_{\sqrt{y}}^2\,dy = \int_0^4 (\frac{8}{3} + 2y^2 - \frac{1}{3}\,y^{3/2} - y^{5/2})\,dy$$

$$= (\frac{8}{3}\,y + \frac{2}{3}\,y^3 - \frac{2}{15}\,y^{5/2} - \frac{2}{7}\,y^{7/2})\Big|_0^4 = \frac{1312}{105}.$$

18.3 Transformation von ebenen Bereichsintegralen. Die Greensche Formel

Ebene Bereichsintegrale wurden bisher ausschließlich in kartesischen Koordinaten berechnet. In vielen Fällen wird die Berechnung wesentlich vereinfacht, wenn geeignete Koordinatensy-

steme verwendet werden, die der Geometrie des Integrationsbereichs G angepasst sind. Ein wichtiger Vertreter solcher neuen Koordinaten sind ebene **Polarkoordinaten**

$$x = r\,\cos\varphi, \quad y = r\,\sin\varphi, \qquad 0 \le r, \quad 0 \le \varphi < 2\pi. \tag{3.1}$$

Polarkoordinaten lassen sich besonders erfolgreich bei *kreissymmetrischen* Integrationsproblemen einsetzen. Im allgemeinen Fall werden wir von Koordinaten

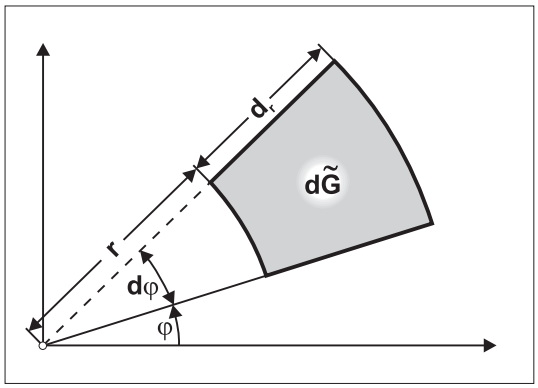
$$u = u(x,y), \quad v = v(x,y) \tag{3.2}$$

ausgehen. Vermöge der Transformation (3.2) wird der ebene Integrationsbereich $G \subset \mathbf{R}^2$ in einen Bereich $\tilde{G} := G_{(u,v)} \subset \mathbf{R}^2$ abgebildet. Um die Eindeutigkeit der Abbildung

$$G = G_{(x,y)} \xrightarrow{(u,v)} \tilde{G}_{(u,v)} = \tilde{G}$$

sicherzustellen, und um Messbarkeitseigenschaften zu erhalten, müssen wir verlangen, dass die Abbildung (3.2) **bijektiv** und **stetig** in beiden Richtungen ist. Diese Eigenschaften werden gemäß dem Satz 14.7 über die inverse Funktion gewährleistet, wenn gilt

$$u,\, v \in C^1(\overline{G}), \qquad \det \frac{\partial(u,v)}{\partial(x,y)} \equiv \begin{vmatrix} u_x & u_y \\ v_x & v_y \end{vmatrix} \neq 0 \;\; \forall\, (x,y) \in \overline{G}. \tag{3.3}$$



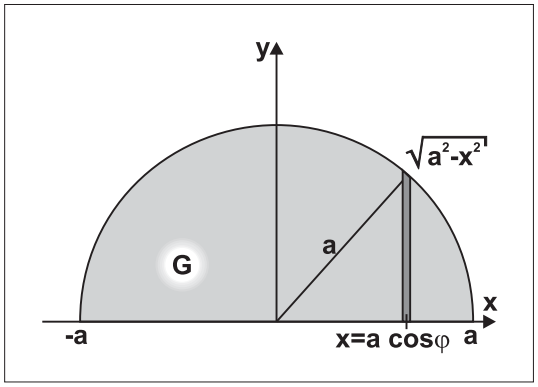
Das Flächenelement in Polarkoordinaten

Es bleibt die Frage zu klären, wie das Flächenelement dG in den neuen Koordinaten (3.2) ausgedrückt werden muss. In kartesischen Koordinaten gilt die elementargeometrische Beziehung $dG = dx\,dy$. Für Polarkoordinaten (3.1) deduziert man ebenfalls aus der naiven, elementargeometrischen Anschauung (siehe obige Skizze)

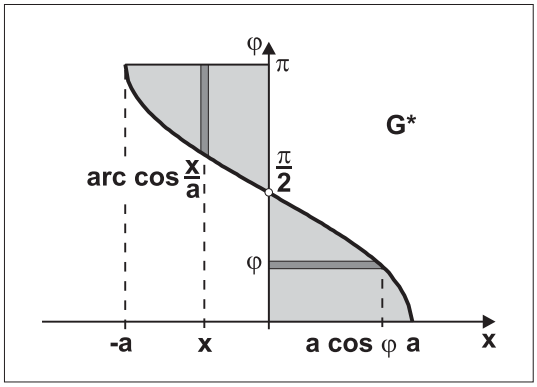
$$\boxed{d\tilde{G} = r\,dr\,d\varphi.}$$

Wir zeigen die Richtigkeit dieser Beziehung für den Halbkreis

$$G := \{(x,y) \in \mathbf{R}^2 \,:\, 0 \le y \le \sqrt{a^2 - x^2}, \;\; -a \le x \le a\}.$$



Der Halbkreis G in kartesischen Koordinaten



Transformation von G auf Polarkoordinaten

Es sei eine stetige Funktion $f \in \mathrm{Abb}\,(G, \mathbf{R})$ gegeben. Dann gilt für das ebene Bereichsintegral von f über G offenbar

$$\int\limits_G f\, dG = \int\limits_{-a}^{+a} \Big(\int\limits_0^{\sqrt{a^2-x^2}} f(x,y)\, dy \Big)\, dx.$$

1.Schritt: Wir fixieren $x \in [-a, a]$ und substituieren $y = h(\varphi)$, wobei gilt:

$$x = r\, \cos\varphi, \quad y = r\, \sin\varphi \quad \Rightarrow \quad y = x\, \tan\varphi =: h(\varphi), \quad dy = \frac{x\, d\varphi}{\cos^2\varphi}.$$

Der Bereich G wird auf den oben skizzierten Bereich G^* transformiert, und wir erhalten

$$\int\limits_G f\, dG = -\int\limits_{-a}^{0} \Big(\int\limits_{\arccos\frac{x}{a}}^{\pi} f(x, h(\varphi)) \frac{x\, d\varphi}{\cos^2\varphi} \Big)\, dx + \int\limits_0^a \Big(\int\limits_0^{\arccos\frac{x}{a}} f(x, h(\varphi)) \frac{x\, d\varphi}{\cos^2\varphi} \Big)\, dx \equiv \int\limits_{G^*} f\, dG^*.$$

Da beide iterierten Integrale über Normalbereiche erstreckt werden, können die Integrationsreihenfolgen vertauscht werden:

$$\int\limits_{G^*} f\, dG^* = \int\limits_0^{\pi/2} \Big(\int\limits_0^{a\,\cos\varphi} f(x, h(\varphi)) x\, dx \Big) \frac{d\varphi}{\cos^2\varphi} - \int\limits_{\pi/2}^{\pi} \Big(\int\limits_{a\,\cos\varphi}^{0} f(x, h(\varphi)) x\, dx \Big) \frac{d\varphi}{\cos^2\varphi}.$$

2.Schritt: Wir fixieren nun $\varphi \in [0, \pi]$ und substituieren $x = g(r)$, wobei gilt:

$$x = r\, \cos\varphi =: g(r), \qquad dx = \cos\varphi\, dr, \quad x\, dx = r\, \cos^2\varphi\, dr.$$

Der Bereich G^* wird auf den folgenden Bereich $\tilde{G}$ transformiert:

$$\tilde{G} := \{(r, \varphi) \in \mathbf{R}^2 \,:\, 0 \le r \le a, \;\; 0 \le \varphi \le \pi\}.$$

Es gilt somit

$$\begin{aligned}
\int\limits_G f\, dG \;&=\; \int\limits_{G^*} f\, dG^* \\
&= \int\limits_0^{\pi/2} \Big(\int\limits_0^a f(g(r), h(\varphi)) r\, \cos^2\varphi\, dr \Big) \frac{d\varphi}{\cos^2\varphi} - \int\limits_{\pi/2}^{\pi} \Big(\int\limits_a^0 f(g(r), h(\varphi)) r\, \cos^2\varphi\, dr \Big) \frac{d\varphi}{\cos^2\varphi} \\
&= \int\limits_{\varphi=0}^{\pi} \Big(\int\limits_{r=0}^{a} f(g(r), h(\varphi)) r\, dr \Big)\, d\varphi = \int\limits_{\tilde{G}} f(r, \varphi) r\, dr\, d\varphi = \int\limits_{\tilde{G}} f\, d\tilde{G}.
\end{aligned}$$

Berechnen wir andererseits die Determinante der JACOBI–Matrix der Abbildung (3.1), so finden wir

$$\det \frac{\partial(x,y)}{\partial(r,\varphi)} = \begin{vmatrix} x_r & x_\varphi \\ y_r & y_\varphi \end{vmatrix} = \begin{vmatrix} \cos\varphi & -r\, \sin\varphi \\ \sin\varphi & r\, \cos\varphi \end{vmatrix} = r.$$

Das heißt, wir haben die Beziehung

$$d\tilde{G} = \Big| \det \frac{\partial(x,y)}{\partial(r,\varphi)} \Big|\, dr\, d\varphi.$$

Diese Beziehung charakterisiert bereits das allgemeine Transformationsverhalten des Flächenelements dG. Wir geben hier ohne Beweis den folgenden Satz an.

Satz 18.8 *Gegeben seien ein beschränkter messbarer Bereich $G \subset \mathbf{R}^2$ in der (x,y)–Ebene sowie eine stetige Funktion $f \in \mathrm{Abb}\,(G,\mathbf{R})$. Es seien ferner u,v ebene krummlinige Koordinaten*

$$x = x(u,v), \quad y = y(u,v), \tag{3.4}$$

und es sei $\tilde{G}$ ein beschränkter messbarer Bereich der (u,v)–Ebene mit den Eigenschaften

$$x, y \in C^1(\tilde{G}), \tag{3.5}$$

$$\det \frac{\partial(x,y)}{\partial(u,v)} \neq 0 \ \ \forall\, (u,v) \in \tilde{G}. \tag{3.6}$$

Wird $\tilde{G}$ durch die Abbildung (3.4) eineindeutig auf den Bereich G abgebildet (eventuell bis auf Randpunkte von $\tilde{G}$), so gilt:

$$\boxed{\begin{aligned} \int\limits_G f(x,y)\,dG &= \iint_{G_{(x,y)}} f(x,y)\,dx\,dy = \iint_{\tilde{G}_{(u,v)}} f(x(u,v),y(u,v))\Big|\det \frac{\partial(x,y)}{\partial(u,v)}\Big|\,du\,dv \\ &= \int\limits_{\tilde{G}} f(u,v)\,d\tilde{G}. \end{aligned}} \tag{3.7}$$

Bemerkung 18.4 Die Bedingung (3.6) kann ohne Änderung der Aussage von Satz 18.8 wie folgt abgeändert werden:

$$\det \frac{\partial(x,y)}{\partial(u,v)} \neq 0 \ \ \forall\, (u,v) \in \tilde{G} \setminus M, \tag{3.8}$$

worin M eine messbare Menge in der (u,v)–Ebene vom Maße Null sei. □

BSP. (18.3.1) Wir betrachten hier ebene Bereichsintegrale auf **unbeschränkten** Bereichen $G \subset \mathbf{R}^2$. In diesem Fall braucht das uneigentliche Integral $\int\limits_G f\,dG$ selbst bei stetigem Integranden $f \in \mathrm{Abb}\,(G,\mathbf{R})$ nicht zu existieren, wie man sich am Beispiel $f = 1$ vergegenwärtigt. Bezeichne $B_r(\vec{0}) \in \mathbf{R}^2$ die offene Kreisscheibe mit Mittelpunkt $\vec{0}$ und Radius $r > 0$, so existiert in vielen Fällen der Grenzwert

$$\lim_{r\to\infty} \int\limits_{G\cap B_r(\vec{0})} f\,dG.$$

In diesem Fall definiert man das **uneigentliche ebene Bereichsintegral** $\int\limits_G f\,dG$ gemäß

$$\int\limits_G f\,dG := \lim_{r\to\infty} \int\limits_{G\cap B_r(\vec{0})} f\,dG.$$

Der folgende Satz liefert ein hinreichendes Kriterium für die Existenz uneigentlicher Bereichsintegrale:

Satz 18.9 *Es sei $G \subset \mathbf{R}^2$ ein unbeschränkter Bereich, und für jedes $r > 0$ sei $G_r := G \cap B_r(\vec{0})$ messbar. Die stetige Funktion $f \in \mathrm{Abb}\,(G,\mathbf{R})$ erfülle für eine Zahl $\alpha > 2$ die Ungleichung*

$$|f(x,y)| \leq \frac{C}{r^\alpha} \ \ \forall\, r := \sqrt{x^2+y^2} \ : \ (x,y) \in G, \quad C = const.$$

Dann konvergiert das uneigentliche Bereichsintegral $\int\limits_G f\,dG$.

Begründung: Wir setzen $r_n := n$ für $n \in \mathbf{N}$. Die Bereiche $G_n := G \cap B_{r_n}(\vec{0})$ sind beschränkt und messbar, und die Funktion $f \in \mathrm{Abb}\,(G_n, \mathbf{R})$ ist stetig. Somit existieren die Bereichsintegrale

$$I_n := \int\limits_{G_n} f\, dG \;\; \forall\, n \in \mathbf{N}.$$

Die Folge $(I_n)_{n\in\mathbf{N}} \subset \mathbf{R}$ ist eine CAUCHY–Folge. Gelte nämlich $n > m \in \mathbf{N}$. Dann folgt unter Verwendung von Polarkoordinaten

$$\begin{aligned} |I_n - I_m| &= \Big|\int\limits_{G_n} f\, dG - \int\limits_{G_m} f\, dG\Big| = \Big|\int\limits_{G_n\setminus G_m} f\, dG\Big| \le C \int\limits_{G_n\setminus G_m} \frac{1}{r^\alpha}\, dG \\ &\le C \int\limits_0^{2\pi} \Big(\int\limits_m^n \frac{1}{r^\alpha}\, r\, dr\Big)\, d\varphi = \frac{2\pi C}{\alpha-2}\,(\frac{1}{m^{\alpha-2}} - \frac{1}{n^{\alpha-2}}) \to 0,\ n > m \to \infty. \end{aligned}$$

Dies ist die behauptete Konvergenz. □

BSP. (18.3.2) Auf dem Bereich $G := \mathbf{R}^2$ ist das uneigentliche Bereichsintegral der Funktion $f(x,y) := e^{-(x^2+y^2)}$ zu berechnen.

Lösung: Man verwendet Polarkoordinaten und erhält einerseits

$$\iint_{\mathbf{R}^2} e^{-(x^2+y^2)}\, dx\, dy = \lim_{R\to\infty} \int\limits_{\varphi=0}^{2\pi} \Big(\int\limits_{r=0}^{R} e^{-r^2}\, r\, dr\Big)\, d\varphi = \lim_{R\to\infty} \int\limits_0^{2\pi} -\frac{1}{2}\, e^{-r^2}\Big|_0^R\, d\varphi = \lim_{R\to\infty} \pi(1 - e^{-R^2}) = \pi.$$

Andererseits gilt

$$\iint_{\mathbf{R}^2} e^{-(x^2+y^2)}\, dx\, dy = \Big(\int_{-\infty}^{+\infty} e^{-x^2}\, dx\Big)\Big(\int_{-\infty}^{+\infty} e^{-y^2}\, dy\Big) = \Big(\int_{-\infty}^{+\infty} e^{-u^2}\, du\Big)^2,$$

und dieser Zusammenhang liefert eine neue Begründung für die Integralformel

$$\boxed{\int_{-\infty}^{+\infty} e^{-u^2}\, du = \sqrt{\pi}.}$$

Der Hauptsatz der Differential– und Integralrechnung einer reellen Veränderlichen besagt, dass für eine auf dem Intervall $I := [a,b]$ stetige und auf (a,b) stetig differenzierbare Funktion $h(x)$ die Beziehung

$$\int_a^b h'(x)\, dx = h(b) - h(a)$$

gilt. Eine Verallgemeinerung dieses Sachverhalts auf ebene Bereichsintegrale müsste zum Ausdruck bringen, dass das Integral $\int\limits_G f\, dG$ in einer Relation zu einem weiteren Integral $\int\limits_{\partial G} f\, ds$ über den Rand ∂G des Gebietes G steht, in welchem nur noch die Werte der zu integrierenden Funktion auf ∂G auftreten. Eine solche Beziehung gilt tatsächlich:

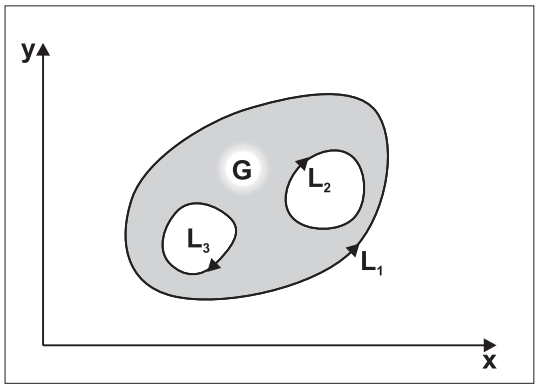
Satz 18.10 (von der GREENschen Formel)
Der beschränkte messbare Bereich $G \subset \mathbf{R}^2$ werde von einer stückweise glatten, geschlossenen, doppelpunktfreien ebenen Kurve C berandet. Diese sei so orientiert, dass das Innere von G stets links von C liegt, wenn C in positiver Richtung durchlaufen wird. Ferner seien $P, Q \in \mathrm{Abb}\,(\overline{G}, \mathbf{R})$ stetige Funktionen mit stetigen partiellen Ableitungen $P_y(x,y)$, $Q_x(x,y)$ im Inneren von G. Dann gilt

$$\boxed{\int\limits_G (Q_x(x,y) - P_y(x,y))\, dG = \oint\limits_C (P(x,y)\, dx + Q(x,y)\, dy).} \tag{3.9}$$

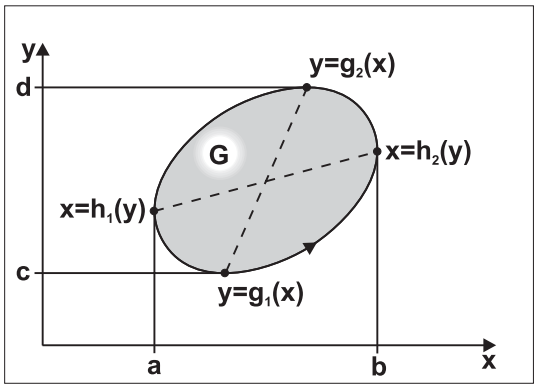
Bemerkung 18.5 Der Rand ∂G des Bereiches G kann sich aus endlich vielen, paarweise disjunkten einfach geschlossenen Kurven $L_1, L_2, \ldots, L_N$ zusammensetzen, wobei jeder Anteil stückweise glatt und im obigen Sinn positiv orientiert sei. Das heißt, jede Randkurve L_i besitzt eine Parameterdarstellung

$$\vec{x}_i(t) := (x_i(t), y_i(t))^T, \quad t \in [a_i, b_i],$$

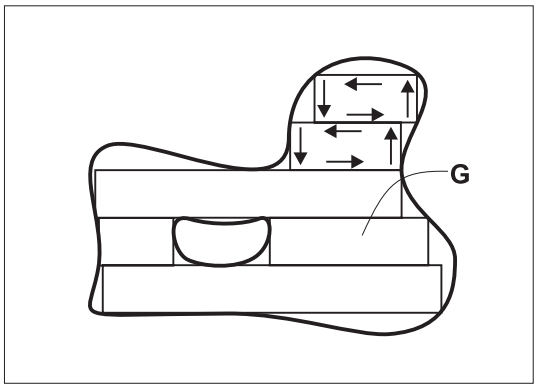
mit stetigen Funktionen x_i, y_i, $i = 1, 2, \ldots, N$, deren Ableitungen beschränkt und bis auf höchstens endlich viele Ausnahmestellen stetig sind. □



Zusammengesetzter Rand ∂G des Bereichs G



GREENsche Formel für einen Normalbereich



GREENsche Formel für einen Bereich, der in Normalbereiche zerlegt wird

Begründung von Satz 18.10 (Skizze): Es sei zunächst angenommen, dass G ein Normalbereich bezüglich der beiden Variablen x und y sei. Wir berechnen $\int\limits_G Q_x \, dG$ als iteriertes Integral:

$$\begin{aligned} \int\limits_G Q_x \, dG &= \int\limits_c^d \Big(\int\limits_{h_1(y)}^{h_2(y)} Q_x(x,y) \, dx \Big) \, dy = \int_c^d [Q(h_2(y), y) - Q(h_1(y), y)] \, dy \\ &= \int_c^d Q(h_2(y), y) \, dy + \int_d^c Q(h_1(y), y) \, dy = \oint\limits_C Q \, dy. \end{aligned}$$

Ganz entsprechend gilt

$$\begin{aligned} -\int\limits_G P_y \, dG &= -\int\limits_a^b \Big(\int\limits_{g_1(x)}^{g_2(x)} P_y(x,y) \, dy \Big) \, dx = -\int_a^b [P(x, g_2(x)) - P(x, g_1(x))] \, dx \\ &= \int_a^b P(x, g_1(x)) \, dx + \int_b^a P(x, g_2(x)) \, dx = \oint\limits_C P \, dx. \end{aligned}$$

Für diesen Fall gilt somit die GREENsche Formel. Ist nun G ein beliebiger beschränkter und messbarer Bereich, so kann eine Approximation an G durch eine genügend feine Rechteckeinteilung erreicht werden. Jedes Rechteck ist ein Normalbereich im geforderten Sinn, so dass die GREENsche Formel auf jedem Rechteck gilt. Beachtet man die Orientierung des Randintegrals, so erkennt man, dass sich die Integrale über die inneren Rechteckseiten wegen gegensätzlicher Orientierung gerade aufheben. Es bleiben nur die Randintegrale des approximierenden Polyeders übrig. Durch entsprechende Verfeinerung erhält man im Grenzübergang die Behauptung. □

Folgerung 18.3 *Wird in der* GREEN*schen Formel speziell* $P(x,y) := -\frac{1}{2}\,y$ *und* $Q(x,y) := \frac{1}{2}\,x$ *gesetzt, so resultiert eine Formel für die Flächenberechnung des Bereichs* G*:*

$$\boxed{A = \text{Fläche}\,(G) = m(G) = \frac{1}{2}\oint\limits_C (x\,dy - y\,dx).} \tag{3.10}$$

BSP. (18.3.3) Es soll mit der Formel (3.10) der Flächeninhalt einer Ellipse mit den Halbachsen a, b berechnet werden. Der Ellipsenbogen C hat die Parameterdarstellung

$$\vec{x}(t) = (a\cos t, b\sin t)^T, \quad t \in [0, 2\pi),$$

und daraus folgt

$$x\,dy - y\,dx = (x\dot{y} - y\dot{x})\,dt = ab\,dt, \qquad A_{\text{Ell}} = \frac{1}{2}\int_0^{2\pi} ab\,dt = \pi ab.$$

BSP. (18.3.4) Wird der Bereich G von einer Randkurve C in **Polarkoordinaten**–Darstellung

$$\vec{x}(\varphi) = (r(\varphi)\,\cos\varphi, r(\varphi)\,\sin\varphi)^T, \quad \varphi_0 \le \varphi \le \varphi_1,$$

berandet, so gilt mit dem Normalenvektor $\vec{n}(\varphi)$ an C:

$$\begin{aligned} x\,dy - y\,dx &= (x\dot{y} - y\dot{x})\,d\varphi = -\langle \vec{x}(\varphi), \vec{n}(\varphi)\rangle\,d\varphi = -\Big\langle r(\varphi)\begin{bmatrix}\cos\varphi\\ \sin\varphi\end{bmatrix}, \begin{bmatrix}-\dot{r}\sin\varphi - r\cos\varphi\\ \dot{r}\cos\varphi - r\sin\varphi\end{bmatrix}\Big\rangle\,d\varphi \\ &= r^2(\varphi)\,d\varphi, \end{aligned}$$

also

$$\boxed{A = \text{Fläche}\,(G) = m(G) = \frac{1}{2}\int_{\varphi_0}^{\varphi_1} r^2(\varphi)\,d\varphi,}$$

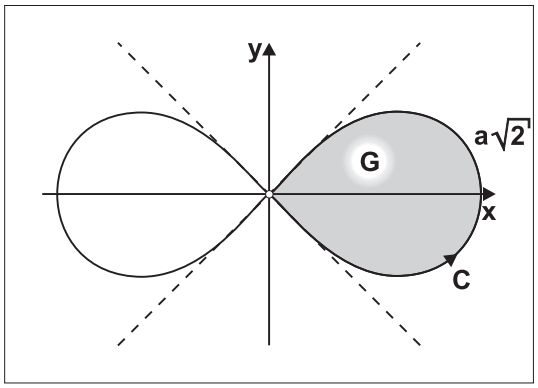
man vergleiche (4.3) in Abschnitt 8.4.

BSP. (18.3.5) Wir berechnen den Flächeninhalt eines **Lemniskatenblattes**: Der Lemniskatenbogen C hat die Polardarstellung

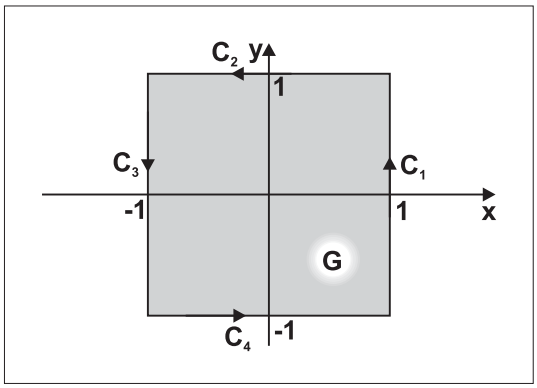
$$r(\varphi) = a\sqrt{2\cos 2\varphi}, \quad a > 0, \quad -\frac{\pi}{4} \le \varphi \le \frac{\pi}{4}, \quad \frac{3\pi}{4} \le \varphi \le \frac{5\pi}{4}.$$

Somit folgt

$$A = \frac{1}{2}\int\limits_{-\pi/4}^{+\pi/4} a^2 \cdot 2\cos 2\varphi\,d\varphi = \frac{a^2}{2}\sin 2\varphi\Big|_{-\pi/4}^{+\pi/4} = a^2.$$



Inhalt eines Lemniskatenblattes



Integrationsbereich im BSP. (18.3.6)

BSP. (18.3.6) Für den oben skizzierten Weg $C = C_1 \cup C_2 \cup C_3 \cup C_4$ berechne man das Wegintegral

$$I := \oint_C (e^{xy}\,dx + xy^2\,dy)$$

direkt und mit Hilfe der GREENschen Formel.

Lösung: Bei direkter Berechnung haben wir

- auf C_1: $x = 1, \quad y = t, \ -1 \le t \le 1, \quad dx = 0, \quad dy = dt,$
- auf C_2: $x = t, \ 1 \ge t \ge -1, \quad y = 1, \quad dx = dt, \quad dy = 0,$
- auf C_3: $x = -1, \quad y = t, \ 1 \ge t \ge -1, \quad dx = 0, \quad dy = dt,$
- auf C_4: $x = t, \ -1 \le t \le 1, \quad y = -1, \quad dx = dt, \quad dy = 0.$

Somit folgt

$$I = \int_{-1}^{1} t^2\,dt + \int_{1}^{-1} e^t\,dt + \int_{1}^{-1} (-t^2)\,dt + \int_{-1}^{1} e^{-t}\,dt = (\frac{1}{3}\,t^3 - e^t + \frac{1}{3}\,t^3 - e^{-t})\Big|_{-1}^{1} = \frac{4}{3}.$$

Formen wir andererseits das Wegintegral I mit Hilfe der GREENschen Formel in ein Bereichsintegral um, so resultiert:

$$I = \int_G (Q_x - P_y)\,dG = \int_G (y^2 - xe^{xy})\,dG = \int_{-1}^{1} \Big(\int_{-1}^{1} (y^2 - xe^{xy})\,dy \Big)\,dx = \int_{-1}^{1} (\frac{1}{3}\,y^3 - e^{xy})\Big|_{-1}^{1}\,dx$$

$$= \int_{-1}^{1} (\frac{2}{3} + e^{-x} - e^x)\,dx = (\frac{2}{3}\,x - e^{-x} - e^x)\Big|_{-1}^{1} = \frac{4}{3}.$$

18.4 Bereichsintegrale im $\mathbf{R}^n$

Die für ebene Bereichsintegrale angestellten Überlegungen können ganz analog auf den n–dimensionalen Fall, $n \ge 3$, übertragen werden. Für eine beschränkte messbare Menge $G \subset \mathbf{R}^n$ und eine stetige Funktion $f \in \mathrm{Abb}\,(G, \mathbf{R}^n)$ existiert das Bereichsintegral $\int_G f(\vec{x})\,dG$. Ist $G \subset \mathbf{R}^n$ ein n–dimensionales **Intervall**: $G := [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$, so kann das Integral $\int_G f(\vec{x})\,dG$ wiederum als iteriertes Integral berechnet werden:

$$\int_G f(\vec{x})\,dG = \int_{a_n}^{b_n} \Big(\int_{a_{n-1}}^{b_{n-1}} \Big(\cdots \Big(\int_{a_1}^{b_1} f(x_1, x_2, \ldots, x_n)\,dx_1 \Big)\,dx_2 \cdots \Big) dx_{n-1} \Big)\,dx_n.$$

Die Reihenfolge der Integration ist hier unerheblich. Für allgemeine Bereiche $G \subset \mathbf{R}^n$ gelten die bei den ebenen Bereichsintegralen getroffenen Einschränkungen hinsichtlich der Normalbereiche. Wir erläutern die Berechnungsmöglichkeiten an dem für die Anwendungen wichtigen Fall des **räumlichen Bereichsintegrals**. Man setzt hier für einen beschränkten messbaren Bereich $G \subset \mathbf{R}^3$ und für eine stetige Funktion $f \in \mathrm{Abb}\,(G, \mathbf{R})$:

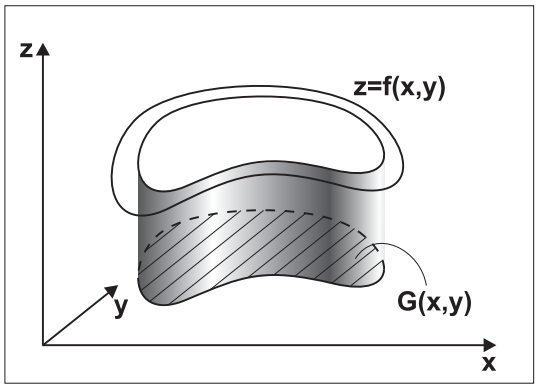
$$\int\limits_G f(\vec{x})\, dV = \iiint_G f(x,y,z)\, dx\, dy\, dz.$$

Im Fall der speziellen Funktion $f = 1$ resultiert

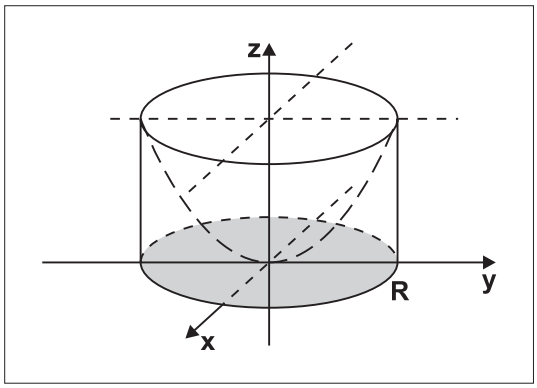
$$\boxed{V = \text{Volumen}\,(G) = m(G) = \int_G 1\, dV = \iiint_G dx\, dy\, dz.}$$

BSP. (18.4.1) Volumina **prismatischer Körper**. Bereiche $G \subset \mathbf{R}^3$ heißen **prismatisch**, wenn sie aus einem ebenen Bodenbereich $G_{(x,y)} \subset \mathbf{R}^2$ (ohne Einschränkung sei $G_{(x,y)}$ als Teilmenge der Ebene $z = 0$ angenommen), einer räumlichen Deckfläche $z = f(x,y) \geq 0$ und einer zylindrischen Mantelfläche (deren Mantellinien Parallelen zur z–Achse sind) bestehen. Ist $G_{(x,y)} \subset \mathbf{R}^2$ beschränkt und messbar und ist $f : G_{(x,y)} \to \mathbf{R}$ eine stetige Funktion, so erhält man

$$\boxed{V_{\text{Prism}} = \text{Volumen}\,(G) = \int_G dV = \iint_{G_{(x,y)}} \Big(\int_{z=0}^{f(x,y)} dz \Big)\, dx\, dy = \iint_{G_{(x,y)}} f(x,y)\, dx\, dy.}$$



Zum Volumen prismatischer Körper



Volumen des prismatischen Körpers von BSP. (18.4.2)

BSP. (18.4.2) Zu berechnen ist das Volumen des prismatischen Körpers

$$G := \{(x,y,z) \in \mathbf{R}^3 \,:\, 0 \leq x^2 + y^2 \leq R^2,\;\; 0 \leq z \leq x^2 + y^2\}.$$

Hier ist der Bodenbereich $G_{(x,y)} = \overline{B_R}(\vec{0})$ die abgeschlossene Kreisscheibe vom Radius $R > 0$ um den Mittelpunkt $\vec{0}$, und die Deckfläche ist das Paraboloid $z = x^2 + y^2$. Wegen der Rotationssymmetrie ist es angebracht, Polarkoordinaten zu verwenden. Man erhält mit dem Flächenelement $d\tilde{G} = r\, dr\, d\varphi$:

$$V = \iint_{0 \leq x^2+y^2 \leq R^2} (x^2 + y^2)\, dx\, dy = \int\limits_{\varphi=0}^{2\pi} \Big(\int\limits_{r=0}^{R} r^2 \cdot r\, dr \Big)\, d\varphi = \frac{1}{2}\, \pi R^4.$$

BSP. (18.4.3) Zwei Kreiszylinder mit Radien R_1, R_2 durchdringen sich zentrisch, so dass die Achsenrichtungen senkrecht zueinander sind. Man bestimme das Volumen des "Bohrloches" im dickeren Zylinder.

Lösung: Es sei $R_2 \geq R_1$ angenommen. Aus Symmetriegründen ist der Volumenanteil unterhalb der Ebene $z = 0$ (siehe folgende Skizze) genauso groß wie der Anteil oberhalb der Ebene $z = 0$. Der Bodenbereich des prismatischen Körpers oberhalb $z = 0$ ist die abgeschlossene Kreisscheibe $G_{(x,y)} := \overline{B_{R_1}}(\vec{0})$, und die Deckfläche ist der Halbzylindermantel $z = \sqrt{R_2^2 - y^2}$. Wir verwenden wegen der Kreissymmetrie wiederum Polarkoordinaten:

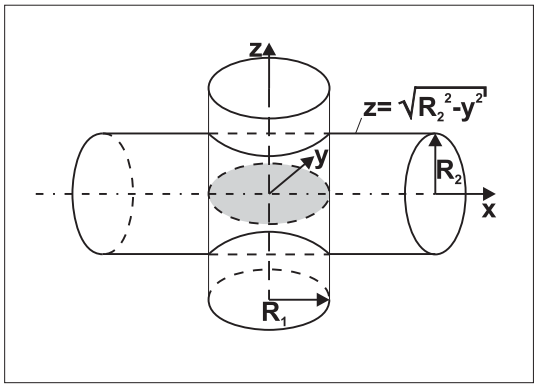
$$V = 2 \iint_{0 \leq x^2+y^2 \leq R_1^2} \sqrt{R_2^2 - y^2}\, dx\, dy = 2 \int_{\varphi=0}^{2\pi} \Big(\int_{r=0}^{R_1} \sqrt{R_2^2 - r^2 \sin^2\varphi}\, r\, dr \Big)\, d\varphi.$$

In dem letzten Integral führen wir eine Substitution $u := r^2 \sin^2\varphi$ durch und erhalten:

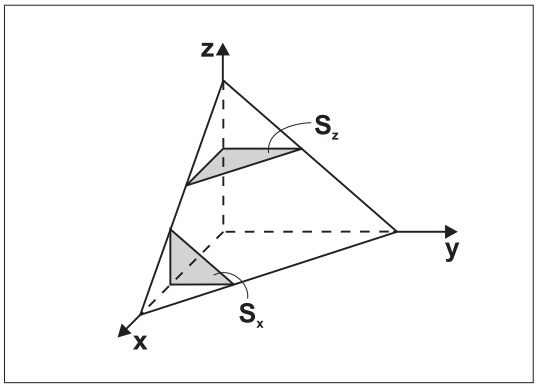
$$V = -\int_0^{2\pi} \frac{2}{3} \frac{(R_2^2 - R_1^2 \sin^2\varphi)^{3/2} - R_2^3}{\sin^2\varphi}\, d\varphi = \frac{8R_2^3}{3} \int_0^{\pi/2} \frac{1 - (1 - (\frac{R_1}{R_2})^2 \sin^2\varphi)^{3/2}}{\sin^2\varphi}\, d\varphi.$$

Für $R_2 > R_1$ kann dieses Integral nicht mehr elementar berechnet werden. Im Sonderfall $R_2 = R_1 =: R$ gilt jedoch:

$$\begin{aligned} V &= \frac{8R^3}{3} \int_0^{\pi/2} \frac{1 - \cos^3\varphi}{\sin^2\varphi}\, d\varphi = \frac{8R^3}{3} \int_0^{\pi/2} \Big(\cos\varphi + \frac{1 - \cos\varphi}{\sin^2\varphi} \Big)\, d\varphi \\ &= \frac{8R^3}{3} \int_0^{\pi/2} \Big(\cos\varphi + \frac{2\sin^2\frac{\varphi}{2}}{4\sin^2\frac{\varphi}{2}\cos^2\frac{\varphi}{2}} \Big)\, d\varphi = \frac{8R^3}{3} \big(\sin\varphi + \tan\frac{\varphi}{2} \big)\Big|_0^{\pi/2} = \frac{16R^3}{3}. \end{aligned}$$



Durchdringung zweier Kreiszylinder, BSP. (18.4.3)



Integration über ein Tetraeder BSP. (18.4.4)

BSP. (18.4.4) Wir zeigen hier zwei Möglichkeiten zur Berechnung des Bereichsintegrals $\int_G f(\vec{x})\, dV$ auf, wenn $G \subset \mathbf{R}^3$ der folgende Tetraederbereich ist:

$$G := \{(x, y, z) \in \mathbf{R}^3 \,:\, 0 \leq x,\ \ 0 \leq y,\ \ 0 \leq z,\ \ x + y + z = 1\}.$$

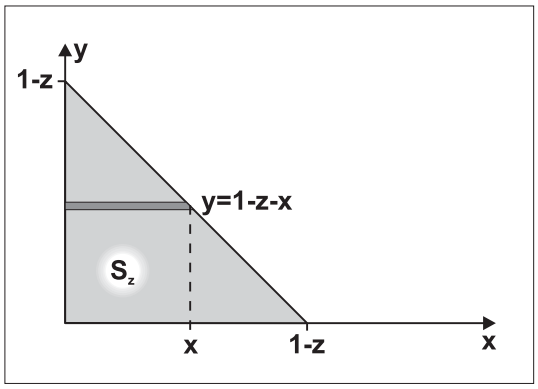
1. Weg:

(a) Man fixiert $z = const.$

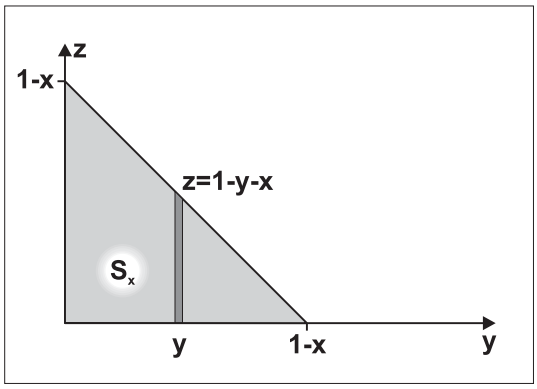
(b) Man bestimmt die Schnitte $S_z(x, y)$ des Bereichs G mit der Ebene $z = const.$

(c) Man bestimmt die Menge H aller z mit $S_z(x, y) \neq \emptyset$. Dann folgt

$$\int_G f(\vec{x})\, dV = \int_H \Big(\iint_{S_z(x,y)} f(x, y, z)\, dx\, dy \Big)\, dz.$$



1.Weg: Der Schnitt S_z des Tetraeders mit der Ebene $z = const$



2.Weg: Der Schnitt S_x des Tetraeders mit der Ebene $x = const$

Im vorliegenden Beispiel des Tetraeders G ist der Schnitt

$$S_z(x,y) := \{(x,y) \in \mathbf{R}^2 \,:\, 0 \le x,\;\; 0 \le y,\;\; 0 \le x+y \le 1-z\}, \quad S_z(x,y) \neq \emptyset \;\; \forall\, z \in [0,1],$$

ein **Normalbereich** bezüglich beider Variablen x und y. Deshalb kann das ebene Bereichsintegral über den Schnitt $S_z(x,y)$ auf zweierlei Arten als iteriertes Integral berechnet werden:

$$\int\limits_G f(\vec{x})\,dV = \int\limits_{z=0}^{1} \Big(\int\limits_{x=0}^{1-z} \Big(\int\limits_{y=0}^{1-z-x} f(x,y,z)\,dy\Big)\,dx\Big)\,dz = \int\limits_{z=0}^{1} \Big(\int\limits_{y=0}^{1-z} \Big(\int\limits_{x=0}^{1-z-y} f(x,y,z)\,dx\Big)\,dy\Big)\,dz.$$

2. Weg:

(a) Man fixiert $x = const$.

(b) Man bestimmt die Schnitte $S_x(y,z)$ des Bereichs G mit der Ebene $x = const$.

(c) Man bestimmt die Menge H aller x mit $S_x(y,z) \neq \emptyset$. Dann folgt

$$\int\limits_G f(\vec{x})\,dV = \int\limits_H \Big(\iint_{S_x(y,z)} f(x,y,z)\,dy\,dz\Big)\,dx.$$

Im vorliegenden Beispiel des Tetraeders G ist der Schnitt

$$S_x(y,z) := \{(y,z) \in \mathbf{R}^2 \,:\, 0 \le y,\;\; 0 \le z,\;\; 0 \le z+y \le 1-x\}, \quad S_x(y,z) \neq \emptyset \;\; \forall\, x \in [0,1],$$

wiederum ein **Normalbereich** bezüglich beider Variablen y und z. Deshalb kann das ebene Bereichsintegral über den Schnitt $S_x(y,z)$ ebenfalls auf zweierlei Arten als iteriertes Integral berechnet werden:

$$\int\limits_G f(\vec{x})\,dV = \int\limits_{x=0}^{1} \Big(\int\limits_{y=0}^{1-x} \Big(\int\limits_{z=0}^{1-y-x} f(x,y,z)\,dz\Big)\,dy\Big)\,dx = \int\limits_{x=0}^{1} \Big(\int\limits_{z=0}^{1-x} \Big(\int\limits_{y=0}^{1-z-x} f(x,y,z)\,dy\Big)\,dz\Big)\,dx.$$

Auch bei der Berechnung von räumlichen Bereichsintegralen kann die Einführung neuer Koordinaten häufig zu wesentlichen Vereinfachungen führen. Dem Satz 18.8 entspricht hier:

Satz 18.11 *Gegeben seien ein beschränkter messbarer Bereich $G \subset \mathbf{R}^3$ des (x,y,z)–Raumes und eine stetige Funktion $f \in \mathrm{Abb}\,(G,\mathbf{R})$. Ferner seien u,v,w krummlinige Koordinaten,*

$$x = x(u,v,w), \quad y = y(u,v,w), \quad z = z(u,v,w). \tag{4.1}$$

Es sei $\tilde{G}$ ein beschränkter messbarer Bereich des (u,v,w)–Raumes mit den Eigenschaften

$$x,\ y,\ z \in C^1(\tilde{G}), \tag{4.2}$$

$$\det \frac{\partial(x,y,z)}{\partial(u,v,w)} \neq 0 \ \ \forall\, (u,v,w) \in \tilde{G}. \tag{4.3}$$

Wird $\tilde{G}$ durch die Abbildung (4.1) *eineindeutig auf den Bereich G abgebildet* (*eventuell bis auf Randpunkte des Bereichs $\tilde{G}$*), *so gilt*

$$\boxed{\begin{aligned} \int\limits_G f(\vec{x})\, dV &= \iiint_{\tilde{G}_{(u,v,w)}} f(x(u,v,w), y(u,v,w), z(u,v,w)) \Big| \det \frac{\partial(x,y,z)}{\partial(u,v,w)} \Big| \, du\, dv\, dw \\ &= \int\limits_{\tilde{G}} f(\vec{u})\, d\tilde{V}. \end{aligned}} \tag{4.4}$$

Bemerkung 18.6 Die Bedingung (4.3) kann ohne Änderung der Aussage des Satzes 18.11 abgeschwächt werden zur Bedingung

$$\det \frac{\partial(x,y,z)}{\partial(u,v,w)} \neq 0 \ \ \forall\, (u,v,w) \in \tilde{G} \setminus M, \tag{4.5}$$

worin M eine messbare Teilmenge vom Maße Null des (u,v,w)–Raumes sei; zum Beispiel eine rektifizierbare Parameterkurve oder eine glatte Fläche (was zu erläutern wäre). □

BSP. (18.4.5) **Zylinderkoordinaten**: Anstelle von (u,v,w) wählt man das Koordinatentripel (r,φ,z) mit

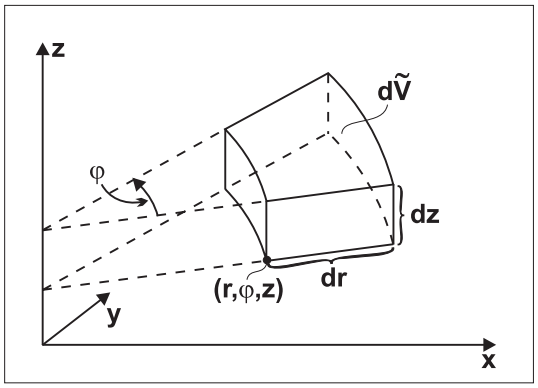
$$x = r\,\cos\varphi, \quad y = r\,\sin\varphi, \quad z = z, \qquad 0 \le r, \quad 0 \le \varphi < 2\pi, \quad z \in \mathbf{R}.$$

Man verifiziert sehr einfach

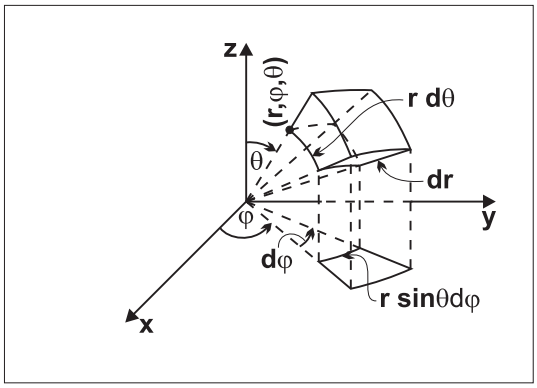
$$\Big| \det \frac{\partial(x,y,z)}{\partial(r,\varphi,z)} \Big| = r \neq 0 \ \ \forall\, r > 0,$$

und somit das folgende auch aus elementargeometrischer Sicht einleuchtende Ergebnis

$$\boxed{d\tilde{V} = r\, dr\, d\varphi\, dz.}$$



Volumenelement in räumlichen Zylinderkoordinaten



Volumenelement in räumlichen Kugelkoordinaten

BSP. (18.4.6) **Kugelkoordinaten**: Anstelle von (u, v, w) wählt man hier das Koordinatentripel (r, φ, ϑ) mit

$$x = r\,\cos\varphi\sin\vartheta, \quad y = r\,\sin\varphi\sin\vartheta, \quad z = r\,\cos\vartheta, \qquad 0 \le r, \quad 0 \le \varphi < 2\pi, \quad 0 \le \vartheta < \pi.$$

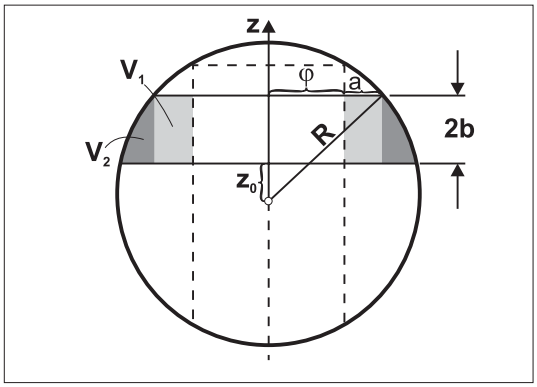
Man verifiziert wieder sehr einfach

$$\Big|\det\frac{\partial(x,y,z)}{\partial(r,\varphi,\vartheta)}\Big| = r^2\,\sin\vartheta \neq 0 \;\;\forall\, r > 0,\; \vartheta \neq 0,$$

und somit das folgende auch aus elementargeometrischer Sicht einleuchtende Ergebnis

$$\boxed{d\tilde{V} = r^2\,\sin\vartheta\,dr\,d\varphi\,d\vartheta.}$$

BSP. (18.4.7) Eine Kugel vom Radius R werde längs eines ihrer Durchmesser mit einer zylindrischen Bohrung vom Radius $\rho < R$ versehen. Es ist das Volumen des so entstehenden Ringes einer Breite $2b$, $b > R$, zu bestimmen.



Ein Ring der Breite $2b$, $b < R$, wird aus einer durchbohrten Kugel geschnitten

Lösung: Wir führen die obere Ringdicke a als Hilfsparameter ein. Dann gilt gemäß Skizze: $V = V_1 + V_2$ mit

$$V_1 := 2b\pi((\rho+a)^2 - \rho^2) = 2b\pi(a^2 + 2\rho a).$$

Wir berechnen das Volumen V_2 unter Verwendung von Zylinderkoordinaten:

$$V_2 = \int_{\varphi=0}^{2\pi}\Big(\int_{z=z_0}^{z_0+2b}\Big(\int_{r=\rho+a}^{\sqrt{R^2-z^2}} r\,dr\Big)\,dz\Big)\,d\varphi = 2\pi\int_{z_0}^{z_0+2b}\frac{1}{2}\,(R^2 - z^2 - (\rho+a)^2)\,dz$$

$$= \pi(R^2 - (\rho+a)^2)2b - \frac{\pi}{3}((z_0+2b)^3 - z_0^3).$$

Wir eliminieren nun die Hilfsgröße z_0, indem wir den Lehrsatz des PYTHAGORAS verwenden, nämlich $z_0 = \sqrt{R^2 - (\rho+a)^2} - 2b$. Aus dem obigen Resultat erhält man nach einigen elementaren Rechenschritten:

$$V_2 = 4\pi b^2(\sqrt{R^2 - (\rho+a)^2} - \frac{2}{3}\,b),$$

und somit das Gesamtvolumen

$$\boxed{V = V_1 + V_2 = 2\pi b(a^2 + 2\rho a) + 4\pi b^2(\sqrt{R^2 - (\rho+a)^2} - \frac{2}{3}\,b).}$$

Im **Sonderfall** $a = 0$ und $\rho = \sqrt{R^2 - b^2}$ wird ein symmetrischer Ring aus der Mittelschicht der Kugel geschnitten. Dessen Volumen beträgt

$$V = \frac{4}{3}\,\pi b^3,$$

und das ist genau das Volumen einer Vollkugel vom Radius b.

Index