Leitfaden zur Vorlesung Vertiefung Mathematik II für Naturwissenschaften und Technik

von Michael Baake

unter Einbezug einer Skriptvorlage von Dirk Frettlöh

ausgearbeitet von Andreas Klötgen

mit Ergänzungen von Michael Baake und Sebastian Herr

31. Januar 2024

Inhaltsverzeichnis

1 Motivation und Ursprünge

1.1 Motivation und Ziele

Die Fourier-Analysis findet heute in vielen Bereichen ihre Anwendung. Dazu gehören neben mathematischen Teilgebieten wie Zahlentheorie, Statistik oder Kombinatorik auch Gebiete der Signalverarbeitung oder Kryptographie. Ein in diesem Leitfaden beschriebenes Beispiel beschäftigt sich mit der Komprimierung von Bilddateien auf einem Computer zur effizienteren Speicherung. Der mathematische Hintergrund ist auf den französischen Mathematiker Jean Baptiste Joseph Fourier zurückzuführen, welcher im 19. Jahrhundert daran arbeitete. Dabei versuchte er in einer seiner Arbeiten, eine periodische Funktion durch eine trigonometrische Reihe darzustellen. In Abschnitt 1.2 wird ein erster Ansatz für dieses Problem erläutert, auf welchem die Fourier-Analysis aufbaut.

1.2 Ursprünge: Die 1-dimensionale Wellengleichung

Die Ursprünge der Fourier-Analysis lassen sich in der Wellengleichung finden. Betrachtet man die 1-dimensionale Wellengleichung, eine lineare partielle Differentialgleichung 2. Ordnung für die Darstellung von Wellen,

$$\frac{\partial^2 u(x,t)}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 u(x,t)}{\partial t^2} = 0, \tag{1}$$

benötigt man zur (eindeutigen) Lösung folgende Anfangswerte:

$$\begin{aligned} u(x,0) &= g(x) \\ \frac{\partial u(x,0)}{\partial t} &= h(x). \end{aligned}$$

Die allgemeine Lösung wird an dieser Stelle nur kurz beschrieben, da es danach mehr um den Separationsansatz für die Anwendung auf die eingespannte Saite geht. Generell gilt für zwei beliebige, zweifach stetig ableitbare Funktionen Φ und Ψ einer Veränderlichen, dass die Summe $\Phi(x-ct)+\Psi(x+ct)$ die 1-dimensionale Wellengleichung löst (Beweis siehe Übungsaufgaben und Weiterführung in [1]). Dabei beschreibt der Term $\pm ct$ im Argument jeweils die Verschiebung des Wellenprofils $\Phi(x)$ bzw. $\Psi(x)$ entlang der x-Koordinate.

Betrachtet man nun eine an zwei Punkten (wir wählen 0 und π) eingespannte Saite, liegt eine spezielle Situation vor, die sich folgendermaßen behandeln lässt. Abbildung 1 zeigt eine Skizze der Situation. Diese Saite kann nun in eine Schwingung gebracht werden, wobei die Auslenkung an der Stelle x ($0 \leq x \leq \pi$) der Saite zur Zeit t ($t \geq 0$) gegeben ist durch eine Funktion $u\colon [0,\pi] \times \mathbb{R}_{\geq 0} \to \mathbb{R}$, welche durch die Wellengleichung aus (1) beschrieben wird. Der Separationsansatz $u(x,t) = v(x) \cdot w(t)$ wird in die Wellengleichung (1) eingesetzt, und es ergibt sich

$$\frac{\ddot{w}(t)}{w(t)} = \alpha^2 \frac{v''(x)}{v(x)}, \tag{2}$$

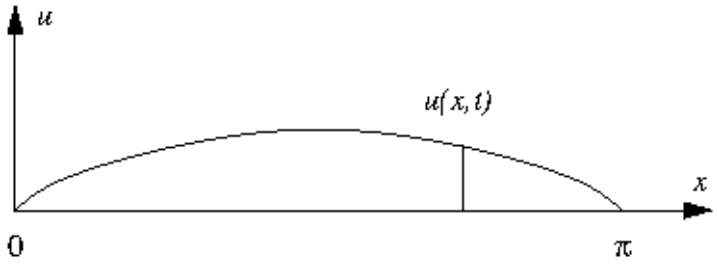

Abbildung 1: Eine in den Stellen 0 und π eingespannte Saite und ihre Auslenkung zum Zeitpunkt t, beschrieben durch die Funktion $u(x,t)$.

wobei $\alpha \in \mathbb{R}$ eine Geschwindigkeits-Konstante ist. Ein Punkt (Strich) steht für die Ableitung nach t (nach x).

Da die Variablen in (2) nun getrennt auf jeder Seite der Gleichung stehen, muss unabhängig von der Wahl der Variablen x und t die Gleichung dennoch gelten. Es folgt also, egal wie x oder t verändert werden, dass der Wert konstant bleibt. Zur Veranschaulichung und einfacheren Schreibweise sei im Folgenden $\frac{v''(x)}{v(x)} = -\lambda$. Durch Umstellen der beiden Gleichungen erhält man dann:

$$v''(x) + \lambda v(x) \;=\; 0 \quad \text{und} \quad \ddot{w}(t) + \alpha^2\, \lambda\, w(t) \;=\; 0\,. \tag{3}$$

Gilt Gl. (3), so erfüllt $u(x,t) = v(x) \cdot w(t)$ die Saitengleichung. An dieser Stelle wird das Problem der Randbedingung deutlich, denn die Saite ist an zwei Stellen eingespannt. An den Stellen $x = 0$ und $x = \pi$ bestehen die Randbedingungen

$$v(0) \;=\; 0 \quad \text{und} \quad v(\pi) \;=\; 0 \tag{4}$$

für die Gleichung $v''(x) + \lambda v(x) = 0$. Dieses sogenannte *Randwertproblem* kann nur für bestimmte Werte von λ eine Lösung besitzen, wobei wir hier unter einer Lösung eine Funktion verstehen wollen, die zweimal stetig differenzierbar ist und die sowohl die DGL als auch die Randbedingungen erfüllt.

Lemma 1.1. *Ist $v \not\equiv 0$ eine relle Lösung des Randwertproblems für v, so ist $\lambda > 0$.*

Beweis. Da v stetig und nicht identisch 0 ist, gilt $\int_0^\pi \big(v(x)\big)^2\,\mathrm{d}x > 0$ nach den üblichen Argumenten aus Analysis I. Nun können wir folgendermaßen rechnen:

$$\begin{aligned}
\lambda \int_0^\pi \big(v(x)\big)^2\,\mathrm{d}x \;&=\; -\int_0^\pi v(x)\, v''(x)\,\mathrm{d}x \\
&=\; -\underbrace{\big[v(x)\, v'(x)\big]_0^\pi}_{=0} + \int_0^\pi \big(v'(x)\big)^2\,\mathrm{d}x \;=\; \int_0^\pi \big(v'(x)\big)^2\,\mathrm{d}x \;>\; 0,
\end{aligned}$$

wobei der erste Schritt die Lösungseigenschaft der DGL benutzt, der zweite partielle Integration, und der dritte die Randbedingung. Da nach Annahme auch v' stetig ist, aber

nicht identisch verschwinden kann (weil dann v konstant und folglich 0 sein müsste), folgt die Ungleichung, und wir bekommen

$$\lambda = \frac{\int_0^\pi \big(v'(x)\big)^2 \,\mathrm{d}x}{\int_0^\pi \big(v(x)\big)^2 \,\mathrm{d}x} > 0$$

wie behauptet. □

Die allgemeine Lösung der Differentialgleichung aus (3) für v (vgl. Übungsaufgaben für Lösungsweg) lautet

$$v(x) = c_1 \cdot \cos(\sqrt{\lambda} \cdot x) + c_2 \cdot \sin(\sqrt{\lambda} \cdot x),$$

mit noch zu bestimmenden Konstanten c_1 und c_2. Aus den Randbedingungen ergibt sich

$$\begin{aligned} v(0) = 0 &\implies c_1 = 0, \\ v(\pi) = 0 &\implies c_2 \cdot \sin(\sqrt{\lambda} \cdot \pi) = 0, \end{aligned}$$

wobei die Lösung für c_2 nicht eindeutig ist. Da aber $c_2 \neq 0$ gefordert ist (ansonsten bekämen wir $v \equiv 0$), muss der zweite Term gleich 0 sein, also $\sin(\sqrt{\lambda} \cdot \pi) = 0$. Die Sinus-Funktion ist in allen ganzzahligen Vielfachen von π gleich 0, also geht $\sqrt{\lambda} = n$ für $n \in \mathbb{Z}$. Da $n = 0$ wiederum $v \equiv 0$ bedeutet, scheidet dieser Wert aus. Ausserdem bedeuten n und $-n$ letztlich dieselbe Funktion, da $\sin(-x) = -\sin(x)$ und ein Vorzeichen in die Konstante c_2 absorbiert werden kann. Also bleibt $\lambda = n^2$ mit $n \in \mathbb{N}$. Mit Hilfe dieser Vorüberlegung kann man nun auch die zweite Gleichung aus (3) durch Einsetzen lösen: $\ddot{w} + (\alpha n)^2 w = 0$. Man erhält nun für die ursprüngliche Auslenkung mit Anfangsbedinungen (4) einer eingespannten Saite

$$\begin{aligned} v_n(x) &= \sin(nx), \\ w_n(t) &= A_n \cos(\alpha n t) + B_n \sin(\alpha n t), \\ u_n(x,t) &= \sin(nx)\big(A_n \cos(\alpha n t) + B_n \sin(\alpha n t)\big). \end{aligned}$$

Das Ergebnis des Separationsansatzes mit $\lambda = n^2$ ist also eine Teillösung mit zwei Konstanten A_n und B_n, wobei die Teillösungen durch die Wahl von $n \in \mathbb{N}$ bestimmt werden.

Für zwei Lösungen $u_n(x,t)$ und $u_m(x,t)$ $(n, m \in \mathbb{N})$ ist (wegen der Linearität der Gleichung) auch die Summe $u_n(x,t) + u_m(x,t)$ eine Lösung des Problems. Es folgt als Ansatz eine Summierung über alle möglichen Einzellösungen des Problems,

$$u(x,t) = \sum_{n \geq 1} u_n(x,t), \tag{5}$$

wobei es eine nicht-triviale Aussage ist, dass man hierdurch tatsächlich alle möglichen Lösungen des Randwertproblems bekommen kann.

Die Startauslenkung zur Zeit $t = 0$ ist gegeben durch die Funktion $g(x)$. Das Geschwindigkeitsprofil $h(x)$, welches die Saite zum Zeitpunkt $t = 0$ besitzt, bestimmt dann

zusammen mit $g(x)$ die anschließende Bewegung. Hierbei gilt insbesondere, dass diese Funktionen sinnvoll sein müssen. Das heißt, dass g keine Sprungstellen haben darf, da die Saite dann an so einer Stelle gerissen ist. Es muss also gelten, dass g und h stetig sind. Mit Hilfe der Summe (5) lassen sich die Anfangsdaten wie folgt ausdrücken:

$$\begin{aligned} u(x,0) &= g(x) = \sum_{n\geq 1} A_n \sin(nx), \\ \frac{\partial u}{\partial t}(x,0) &= h(x) = \sum_{n\geq 1} \alpha n \cdot B_n \cdot \sin(nx). \end{aligned}$$

Außerdem muss die Reihe aus (5) konvergieren und zweimal stetig nach x und t ableitbar sein, damit diese eine (klassische) Lösung der ursprünglichen Gleichung (1) ist. Darauf kommen wir später noch einmal zurück.

Aus dem Beispiel der 1-dimensionalen Wellengleichung lässt sich nun die Fourier-Analysis begründen. Die Lösung bedeutet nämlich die Darstellung einer Funktion durch eine trigonometrische Reihe. Die Fourier-Analysis beschäftigt sich aber auch mit komplexen Funktionen, bei denen dann komplexe e-Funktionen zum Einsatz kommen (später mehr dazu). Zunächst stößt man aber auf die Fourier-Reihen, welche sich direkt aus der aufgestellten Reihe aus (5) herleiten lassen. Im nächsten Abschnitt 2 werden diese weiter untersucht, um die Lösungen für die neu eingeführten Größen A_n und B_n zu bestimmen.

2 Fourier-Reihen

Betrachtet man die in Abschnitt 1.2 aufgestellte Reihe etwas allgemeiner, erhält man die Definition einer trigonometrischen Reihe für das Intervall $[-\pi, \pi]$:

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nx) + b_n \sin(nx) \tag{6}$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx)\, \mathrm{d}x, \qquad (n \in \mathbb{N}_0) \tag{7}$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx)\, \mathrm{d}x, \qquad (n \in \mathbb{N}) \tag{8}$$

Definition 2.1. Gegeben sei eine Funktion $f\colon [-\pi, \pi] \longrightarrow \mathbb{R}$, mit $f(\pi) = f(-\pi)$. Die Reihe aus (6) heißt *Fourier-Reihe* (kurz: FR) von f auf $[-\pi, \pi]$. Die Koeffizienten a_n und b_n aus (7) und (8) heißen *Fourier-Koeffizienten* von f.

In diesem Zusammenhang stellen wir uns f auch auf ganz $\mathbb{R}$ fortgesetzt vor, indem wir fordern, dass $f(x+2\pi) = f(x)$ für alle $x \in \mathbb{R}$ gilt. Dann ist f eine 2π-periodische Funktion (eine allgemeine Definition hierzu folgt noch). Sofern die Summe auf der rechten Seite von (6) in geeigneter Weise konvergiert (auch dies soll später noch genauer diskutiert werden), ist auch diese Summe 2π-periodisch.

An dieser Stelle wollen wir zunächst gleichmäßige Konvergenz der Fourier-Reihe voraussetzen, da dies in der unten benutzten Berechnung der Fourier-Koeffizienten a_n und b_n verwendet wird. Näheres dazu wird es später geben, hier reicht es erst einmal als Voraussetzung. Für diese Berechnung ist zusätzlich ein Lemma hilfreich.

Lemma 2.2.

$$\frac{1}{\pi} \int_{-\pi}^{\pi} \cos(ns) \sin(ms)\, \mathrm{d}s = 0 \qquad (n, m \geq 0) \tag{9}$$

$$\frac{1}{\pi} \int_{-\pi}^{\pi} \cos(ns) \cos(ms)\, \mathrm{d}s = \delta_{n,m} \qquad (n, m \geq 1) \tag{10}$$

$$\frac{1}{\pi} \int_{-\pi}^{\pi} \sin(ns) \sin(ms)\, \mathrm{d}s = \delta_{n,m} \qquad (n, m \geq 1) \tag{11}$$

$$\frac{1}{\pi} \int_{-\pi}^{\pi} \mathrm{d}s = 2\,, \tag{12}$$

wobei $\delta_{n,m} = \begin{cases} 1, & \textit{falls } n = m\,, \\ 0, & \textit{sonst}\,. \end{cases}$

Beweis. Dazu hilft eine kurze Vorüberlegung, die für $n \in \mathbb{Z}$ stimmt:

$$\int_{-\pi}^{\pi} \cos(nx)\, \mathrm{d}x = \begin{cases} \left[\frac{\sin(nx)}{n}\right]_{-\pi}^{\pi} = 0\,, & n \neq 0\,, \\ 2\pi\,, & n = 0\,, \end{cases} \tag{13}$$

$$\int_{-\pi}^{\pi} \sin(nx)\, \mathrm{d}x = 0\,. \tag{14}$$

Zusätzlich werden noch die Additionstheoreme der trigonometrischen Funktionen verwendet:

$$\cos(\alpha+\beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta) \tag{15}$$
$$\sin(\alpha+\beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta) \tag{16}$$
$$\cos(\alpha-\beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta) \tag{17}$$
$$\sin(\alpha-\beta) = \sin(\alpha)\cos(\beta) - \cos(\alpha)\sin(\beta) \tag{18}$$

Daraus folgt: Beweis für (9), für $n, m \geq 0$:

$$\frac{1}{\pi}\int_{-\pi}^{\pi} \cos(ns)\sin(ms)\,\mathrm{d}s \overset{(16)-(18)}{=} \frac{1}{2\pi}\int_{-\pi}^{\pi} \sin((n+m)s) - \sin((n-m)s)\,\mathrm{d}s \overset{(14)}{=} 0.$$

Beweis für (10), für $n, m \geq 1$:

$$\begin{aligned}\frac{1}{\pi}\int_{-\pi}^{\pi} \cos(ns)\cos(ms)\,\mathrm{d}s &\overset{(15)+(17)}{=} \frac{1}{2\pi}\int_{-\pi}^{\pi} \cos((n+m)s) + \cos((n-m)s)\,\mathrm{d}s \\ &\overset{(13)}{=} \delta_{n,m}\,.\end{aligned}$$

Beweis für (11), ebenfalls für $n, m \geq 1$:

$$\begin{aligned}\frac{1}{\pi}\int_{-\pi}^{\pi} \sin(ns)\sin(ms)\,\mathrm{d}s &\overset{(17)-(15)}{=} \frac{1}{2\pi}\int_{-\pi}^{\pi} \cos((n-m)s) - \cos((n+m)s)\,\mathrm{d}s \\ &\overset{(13)}{=} \delta_{n,m}\,.\end{aligned}$$

□

Mit Hilfe von Lemma 2.2 kann man nun die Koeffizienten der FR bestimmen. Zunächst für die a_n, indem die FR aus (6) mit dem Faktor $\frac{1}{\pi}\cos(ns)$ multipliziert wird und anschließend von $-\pi$ bis π integriert wird, für $n \geq 0$:

$$\begin{aligned}&\frac{1}{\pi}\int_{-\pi}^{\pi} f(s)\cos(ns)\,\mathrm{d}s \\ &= \frac{1}{\pi}\int_{-\pi}^{\pi}\left(\frac{a_0}{2} + \sum_{m=1}^{\infty} a_m\cos(ms) + b_m\sin(ms)\right)\cos(ns)\,\mathrm{d}s \\ &= \frac{a_0}{2\pi}\underbrace{\int_{-\pi}^{\pi}\cos(ns)\,\mathrm{d}s}_{=2\pi\delta_{0,n}} + \frac{1}{\pi}\int_{-\pi}^{\pi}\left(\sum_{m=1}^{\infty} a_m\cos(ms) + b_m\sin(ms)\right)\cos(ns)\,\mathrm{d}s \\ &\quad \text{Tauschen von Integral und Summe ist hier möglich,} \\ &\quad \text{da gleichmäßige Konvergenz vorausgesetzt wurde} \\ &= a_0\delta_{0,n} + \sum_{m=1}^{\infty} a_m \underbrace{\frac{1}{\pi}\int_{-\pi}^{\pi}\cos(ms)\cos(ns)\,\mathrm{d}s}_{=\delta_{m,n}\ (\text{nach (10)})} + \sum_{m=1}^{\infty} b_m \underbrace{\frac{1}{\pi}\int_{-\pi}^{\pi}\sin(ms)\cos(ns)\,\mathrm{d}s}_{=0\ (\text{nach (9)})} \\ &= \sum_{m=0}^{\infty} a_m\delta_{m,n} = a_n\,.\end{aligned}$$

Die Rechnung für die b_n läuft analog, diesmal für $n \geq 1$, nur dass jetzt zunächst mit dem Faktor $\frac{1}{\pi}\sin(ns)$ multipliziert wird und die Regel (11) verwendet wird. So erhält man die geschlossene Form der Fourier-Koeffizenten aus (7) und (8).

Definition 2.3. Gegeben sei eine Funktion $f\colon \mathbb{R} \to \mathbb{R}$, von der wir annehmen wollen, dass sie nicht konstant ist.
(a) f heißt *T-periodisch*, falls für alle $x \in \mathbb{R}$ gilt:

$$f(x+T) = f(x)\,.$$

Das kleinste positive T mit dieser Eigenschaft heißt *Periode* von f.
(b) Die Funktion f heißt *gerade* (bzw. *ungerade*), falls für alle $x \in \mathbb{R}$ gilt:

$$f(-x) = f(x) \quad \big(\text{bzw. } f(-x) = -f(x)\big).$$

Noch ist nicht klar, ob (6) wirklich gilt, also ob die Fourier-Reihe wirklich gegen f konvergiert (z.B. punktweise oder sogar gleichmäßig). Falls dies allerdings stimmt, ist die Funktion auf ganz $\mathbb{R}$ definiert. Zusätzlich gilt durch die Periodizität der Sinus- und Cosinus-Funktion, also $\sin(x) = \sin(x+2n\pi)$ und $\cos(x) = \cos(x+2n\pi)$ für alle $x \in \mathbb{R}$, $n \in \mathbb{Z}$, dass auch die Funktion f periodisch ist: $f(x) = f(x+2n\pi)$ gilt dann für alle $x \in \mathbb{R}, n \in \mathbb{Z}$. Die Funktion f ist also 2π-periodisch.

Beispiel 2.4. Wir setzen

$$f(x) = \begin{cases} 1\,, & x \in \bigcup_{n\in\mathbb{Z}}\big(2n\pi, (2n+1)\pi\big), \\ 0\,, & x = n\pi \text{ mit } n \in \mathbb{Z}, \\ -1\,, & x \in \bigcup_{n\in\mathbb{Z}}\big((2n-1)\pi, 2n\pi\big). \end{cases}$$

und betrachten nun das Rechtecksignal mit 2π-periodischer Fortsetzung in beiden Richtungen. Abbildung 2 zeigt diese Situation schematisch.

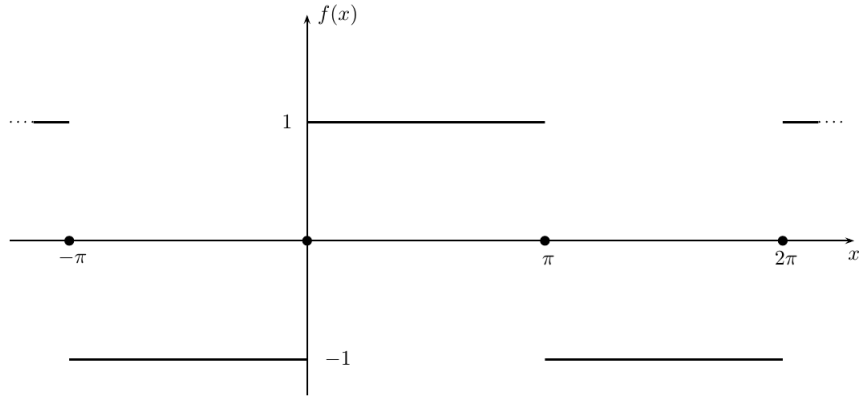

Abbildung 2: Der Verlauf des Rechtecksignals in einem kleinen Ausschnitt.

Die Funktion f aus Beispiel 2.4 ist eine ungerade Funktion, es gilt also $f(-x) = -f(x)$ für alle x. Zunächst werden nun die Fourier-Koeffizienten a_n und b_n berechnet.

$$\begin{aligned} a_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)\,\mathrm{d}x \overset{y=-x}{=} \frac{1}{\pi}\int_{\pi}^{-\pi} f(-y)\cos(-ny)(-1)\,\mathrm{d}y \\ &= \frac{1}{\pi}\int_{-\pi}^{\pi} -f(y)\cos(ny)\,\mathrm{d}y \;=\; -a_n \end{aligned}$$

Dabei wurde die Substitutionsregel eingesetzt und verwendet, dass die Cosinus-Funktion gerade ist. Man erhält, dass $a_n = -a_n$ gilt und damit $a_n = 0$, für alle $n \in \mathbb{N}_0$. Dies gilt allgemeiner für jede ungerade Funktion f. Einen Beweis für diese Aussage und das Verhalten einer geraden Funktion f bei der Berechnung der Fourier-Koeffizienten wird in den Übungsaufgaben weiter untersucht. Daher genügt es, an dieser Stelle nur noch die b_n weiter zu betrachten.

$$\begin{aligned} b_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)\,\mathrm{d}x \;=\; \frac{1}{\pi}\int_{-\pi}^{0} f(x)\sin(nx)\,\mathrm{d}x + \frac{1}{\pi}\int_{0}^{\pi} f(x)\sin(nx)\,\mathrm{d}x \\ &\overset{y=-x}{=} \frac{1}{\pi}\int_{\pi}^{0} f(-y)\sin(-ny)(-1)\,\mathrm{d}y + \frac{1}{\pi}\int_{0}^{\pi} f(x)\sin(nx)\,\mathrm{d}x \\ &= \frac{1}{\pi}\int_{0}^{\pi} (-f(y))\cdot(-\sin(ny))\,\mathrm{d}y + \frac{1}{\pi}\int_{0}^{\pi} f(x)\sin(nx)\,\mathrm{d}x \\ &= \frac{2}{\pi}\int_{0}^{\pi} f(x)\,\sin(nx)\,\mathrm{d}x \;=\; \frac{2}{\pi}\int_{0}^{\pi}\sin(nx)\,\mathrm{d}x \\ &= \frac{2}{\pi}\left[\frac{-\cos(nx)}{n}\right]_0^{\pi} \;=\; \frac{2}{\pi}\underbrace{(\frac{1}{n} - \frac{(-1)^n}{n})}_{=0 \text{ für gerade } n} \;=\; \begin{cases} 0\,, & n = 2k\,, \\ \frac{4}{n\pi}\,, & n = 2k+1\,. \end{cases} \end{aligned}$$

In dieser Rechnung wurde verwendet, dass der Wert der Funktion f im Intervall $(0, \pi)$ gleich 1 ist, und die beiden Randpunkte zum Wert des Integrals nicht beitragen. Für gerade n gilt weiter, dass das Ergebnis immer 0 ist. So ergibt sich die Fallunterscheidung für das Ergebnis der b_n. Damit lässt sich nun eine sukzessive Approximation FR_N auf der Basis der Fourier-Reihe des Rechtecksignals folgendermaßen durch endliche Summation aufstellen:

$$\mathrm{FR}_N := \frac{4}{\pi}\left(\sin(x) + \frac{1}{3}\sin(3x) + ... + \frac{\sin(2N+1)x}{2N+1}\right) = \frac{4}{\pi}\sum_{n=0}^{N}\frac{\sin((2n+1)x)}{2n+1}$$

Die berechnete Fourier-Reihe approximiert nun die Funktion $f(x)$ mit einer Folge von stetigen Funktionen in Abhängigkeit von N. Um dies zu veranschaulichen, zeigt Abbildung 3 einige Beispiele für verschiedene Werte von N, wobei zu erkennen ist, dass die sukzessive Approximation mit steigendem N besser wird.

Der erste Term der Fourier-Reihe ist ein Vielfaches der Sinus-Funktion, durch den zweiten Term erhält die Reihe allerdings schon eine Korrektur der Kurve in Richtung der ursprünglichen Funktion.

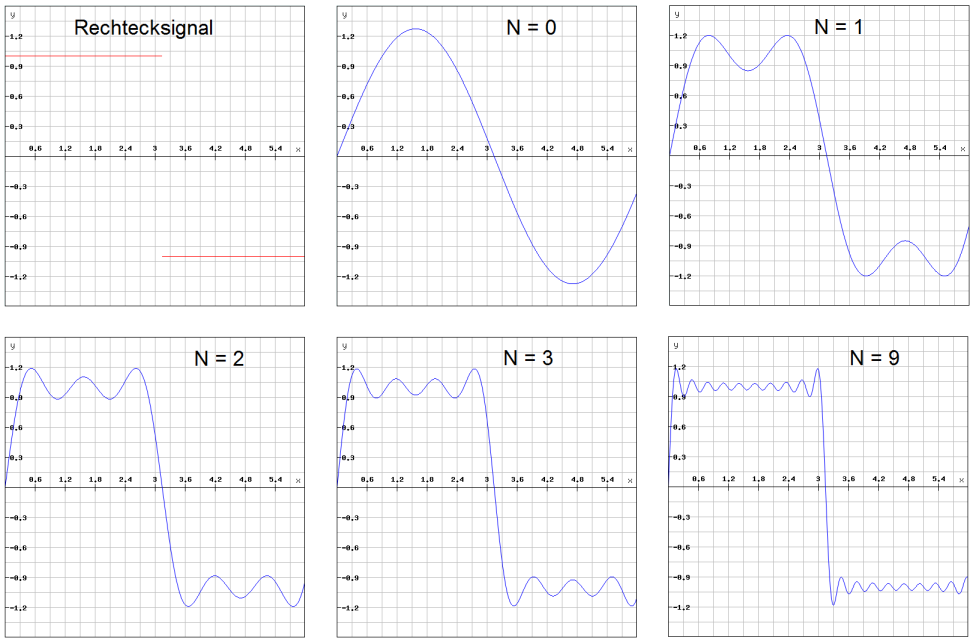

Abbildung 3: Rechtecksignal und die sukzessive Approximation durch die abgeschnittenen Fourier-Reihen mit $N = 0$, $N = 1$, $N = 2$, $N = 3$ und $N = 9$. An den Sprungstellen der ursprünglichen Funktion bei $n\pi$ sind die approximierenden Funktionen stetig, die Unstetigkeiten ergeben sich erst im Limes.

Es gibt an den Sprungstellen, also den ganzzahligen Vielfachen von π, in den Approximationen eine stärkere Ungenauigkeit. Die Fourier-Reihe zittert sozusagen stärker in diesem Bereich. Dies wird auch Gibbs'sches Phänomen genannt, wobei sich das Überschwingverhalten quantifizieren lässt. Die Untersuchung des Konvergenzverhaltens einer Fourier-Reihe ist also wichtig, und wird weiter unten noch genauer betrachtet.

2.1 Konvergenzverhalten der Fourier-Reihe

Definition 2.5. Gegeben sei eine Folge $(f_n)_{n\in\mathbb{N}}$ von Funktionen$f_n\colon \mathbb{D} \to \mathbb{R}$ oder auch $f_n\colon \mathbb{D} \to \mathbb{C}$, wobei $\mathbb{D}$ ein Intervall von $\mathbb{R}$ sei. Dann definieren wir wie folgt:

(a) $(f_n)_{n\in\mathbb{N}}$ heißt *punktweise konvergent* gegen die Funktion f, falls

$$\forall x \in \mathbb{D},\, \varepsilon > 0 :\ \exists N \in \mathbb{N} :\ \forall n \geq N :\ \ \big|f_n(x) - f(x)\big| < \varepsilon\,;$$

(b) $(f_n)_{n\in\mathbb{N}}$ heißt *gleichmäßig konvergent* gegen die Funktion f, falls

$$\forall \varepsilon > 0 :\ \exists N \in \mathbb{N} :\ \forall x \in \mathbb{D},\, n \geq N :\ \ \big|f_n(x) - f(x)\big| < \varepsilon\,;$$

(c) $(f_n)_{n\in\mathbb{N}}$ heißt *konvergent im quadratischen Mittel* gegen die Funktion f, falls

$$\big\|f_n - f\big\|_2 \to 0 \text{ für } n \to \infty$$

(vgl. für $\|\cdot\|_2$ auch Definition 3.1).

In einfachen Fällen lässt sich an der Fourier-Reihe direkt feststellen, ob gleichmäßige Konvergenz gegen die ursprüngliche Funktion f für $N \to \infty$ vorliegt.

Bemerkung 2.6. Ein hinreichendes Kriterium für gleichmäßige Konvergenz ist, dass die Koeffizienten der Fourier-Reihe schnell genug gegen 0 fallen. So z.B. für $|a_n|, |b_n| \leq \frac{c}{n^2}$ für eine Konstante $c > 0$, denn es gilt

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

Allerdings gibt es auch Fälle, bei denen dies nicht anwendbar ist, wie bei $\frac{1}{n}$, denn die harmonische Reihe divergiert bekanntlich. Ist jedoch f auf ganz $\mathbb{R}$ stetig und gilt zudem $|a_n|, |b_n| \leq \frac{c}{n}$ für eine Konstante $c > 0$, dann liegt wieder gleichmäßige Konvergenz vor.

Beispiel 2.7. Gegeben sei eine Funktion $f(x) = |x|$ für $[-\pi, \pi]$ mit 2π-periodischer Fortsetzung. Dies definiert eine stetige (!) Funktion, die oft als Dreiecks- oder Zick-Zack-Funktion bezeichnet wird. Die Fourier-Reihe dazu lautet:

$$f(x) = \frac{\pi}{2} - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\cos((2n-1)x)}{(2n-1)^2}$$

Es folgt, dass $\frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2}$ eine Majorante der Fourier-Reihe ist und es gilt nach obigem Kriterium gleichmäßige Konvergenz.

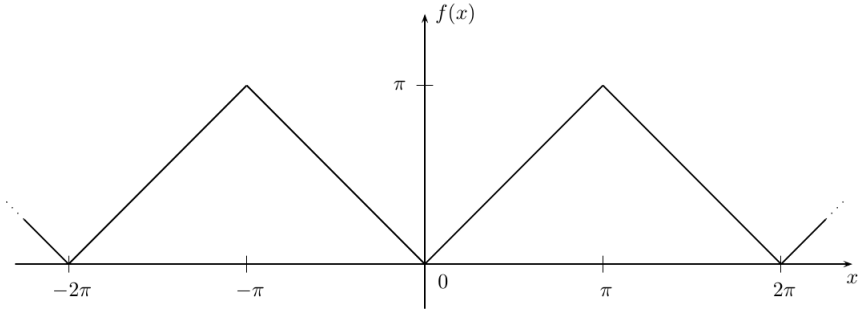

Abbildung 4: Verlauf des Dreieckssignals in einem kleinen Ausschnitt.

Im Allgemeinen benötigt man jedoch noch weitere Kriterien, von denen wir hier eines vorstellen wollen.

Definition 2.8. Eine Funktion f ist von *beschränkter Variation* auf einem Intervall $[a, b]$, wenn ein $M > 0$ existiert, so dass für jede Zerlegung

$$Z = \{x_0 = a < x_1 < x_2 < \ldots < x_n = b\}$$

des Intervalls gilt:

$$V(f, z) = \sum_{k=1}^{n} \left| f(x_k) - f(x_{k-1}) \right| < M\,.$$

Jede monotone Funktion $f:\ [a,b] \to \mathbb{R}$ ist von beschränkter Variation, aber stetige Funktionen sind i.A. nicht von beschränkter Variation. Allgemein gilt, dass f genau dann von beschränkter Variation ist, wenn f als Differenz zweier monoton wachsender Funktionen geschrieben werden kann. Mehr dazu findet man in vielen Lehrbüchern zur Analysis. Die Bedeutung ergibt sich aus folgendem Resultat.

Satz 2.9. *Sei f eine 2π-periodische Funktion auf $\mathbb{R}$. Ist f auf $[-\pi,\pi]$ von beschränkter Variation, so konvergiert die FR für jedes $x \in \mathbb{R}$ gegen den Wert*

$$s(x) = \frac{1}{2}\big(f(x^+) + f(x^-)\big).$$

wobei $f(x^+)$ bzw. $f(x^-)$ den rechtsseitigen bzw. linksseitigen Limes an der Stelle x bezeichnet. Ist f in x stetig, dann liegt punktweise Konvergenz gegen $f(x)$ vor. Ist f auf $\mathbb{R}$ zusätzlich stetig, so konvergiert die FR gleichmäßig.

Auf diese Weise kann oft die starke Form der gleichmäßigen Konvergenz geprüft werden. Allerdings gibt es viele Fourier-Reihen, bei denen dies so nicht möglich ist bzw. keine gleichmäßige Konvergenz vorliegt. Um dennoch eine Aussage über das Verhalten der Fourier-Reihe für $N \to \infty$ bzw. der Approximation an die Funktion f zu treffen, gibt es schwächere Kriterien, z.B. ob die Fourier-Reihe (ggf. fast überall) punktweise konvergiert, oder ob sie im quadratischen Mittel konvergiert. Dazu später mehr in Kapitel 3.2.

2.2 Weiterentwicklung der Theorie der Fourier-Reihen

Für eine Weiterentwicklung der Theorie werden zunächst die Fourier-Koeffizienten etwas umgerechnet. Setze $a_{-n} = a_n$ und $b_{-n} = -b_n$ (somit ist $b_0 = 0$) und beachte, dass $\mathrm{e}^{\mathrm{i}x} = \cos(x) + i\sin(x)$ gilt. Wir nehmen an, dass f eine gleichmäßig konvergente FR besitzt. Dann folgt:

$$\begin{aligned} f(x) &= \frac{1}{2}\sum_{n\in\mathbb{Z}}\Big(\frac{a_n}{2}\big(\mathrm{e}^{\mathrm{i}nx} + \mathrm{e}^{-\mathrm{i}nx}\big) + \frac{b_n}{2\mathrm{i}}\big(\mathrm{e}^{\mathrm{i}nx} - \mathrm{e}^{-\mathrm{i}nx}\big)\Big) \\ &= \frac{1}{2}\sum_{n\in\mathbb{Z}}\Big(\frac{a_n + a_{-n}}{2} + \frac{b_n - b_{-n}}{2i}\Big)\mathrm{e}^{\mathrm{i}nx} \\ &= \sum_{n\in\mathbb{Z}}\underbrace{\frac{1}{2}(a_n - \mathrm{i}b_n)}_{=:\,c_n}\mathrm{e}^{\mathrm{i}nx} \;=\; \sum_{n\in\mathbb{Z}} c_n \mathrm{e}^{\mathrm{i}nx}. \end{aligned} \tag{19}$$

wobei

$$c_n \;=\; \frac{a_n - \mathrm{i}b_n}{2} \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi} f(s)\underbrace{\big(\cos(ns) - \mathrm{i}\sin(ns)\big)}_{=\mathrm{e}^{-\mathrm{i}ns}}\,\mathrm{d}s \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi} f(s)\mathrm{e}^{-\mathrm{i}ns}\,\mathrm{d}s\,.$$

Weiter gilt

$$\begin{aligned}
\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}ns}\,\mathrm{e}^{\mathrm{i}ms}\,\mathrm{d}s &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}(m-n)s}\,\mathrm{d}s \\
&= \begin{cases} m \neq n : \ \frac{\mathrm{e}^{\mathrm{i}(m-n)s}}{2\pi\mathrm{i}\,(m-n)}\Big|_{-\pi}^{\pi} = \frac{1}{2\pi\mathrm{i}\,(m-n)}\underbrace{\left(\mathrm{e}^{\mathrm{i}(m-n)\pi} - \mathrm{e}^{-\mathrm{i}(m-n)\pi}\right)}_{=0} \\ m = n : \ 1 \end{cases} \\
&= \delta_{m,n}\,.
\end{aligned}$$

Dieses Ergebnis sieht wie ein Skalarprodukt aus. Um die neuen Koeffizienten c_n besser darzustellen, findet an dieser Stelle zunächst ein Einschub zum Dirac-Formalismus statt, um eine vorteilhafte Notation für Skalarprodukte zu erhalten.

Für das Beispiel des reellen Vektorraums $\mathbb{R}^2$ geht das so:

$$\begin{aligned}
& & \text{Basis:} \quad & e_1 = \left(\begin{smallmatrix}1\\0\end{smallmatrix}\right) = |1\rangle,\ e_2 = \left(\begin{smallmatrix}0\\1\end{smallmatrix}\right) = |2\rangle & \\
\langle x| = |x\rangle^t & & \text{also} \quad & \langle 1| = (1,0),\ \langle 2| = (0,1) & \\
\langle 1|1\rangle & = & & (1,0)\left(\begin{smallmatrix}1\\0\end{smallmatrix}\right) = 1 & (20)\\
\langle 2|2\rangle & = & & (0,1)\left(\begin{smallmatrix}0\\1\end{smallmatrix}\right) = 1 & (21)\\
\langle 1|2\rangle = \langle 2|1\rangle & = & & 0\,. & (22)
\end{aligned}$$

Dabei geben die Gleichungen (20) und (21) jeweils das Quadrat der Länge des Vektors an und (22) besagt, dass die Vektoren senkrecht aufeinander stehen. Durch diese Regeln lässt sich die Vektor- und Matrixmultiplikation in folgende Form bringen:

$$\begin{aligned}
|1\rangle\langle 1| &= \left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)(1,0) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \\
|2\rangle\langle 2| &= \left(\begin{smallmatrix}0\\1\end{smallmatrix}\right)(0,1) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \\
(|1\rangle\langle 1|)^2 &= |1\rangle\langle 1\underbrace{|\,|}_{=|}1\rangle\langle 1| \;=\; |1\rangle\langle 1| \\
|1\rangle\langle 1| + |2\rangle\langle 2| &= \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \;=\; \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbb{1}\,.
\end{aligned}$$

Dabei ist $\mathbb{1}$ immer die Einheitsmatrix des entsprechenden Vektorraums. Es folgt eine Basisdarstellung für einen Vektor $v \in \mathbb{R}^2$:

$$\begin{aligned}
v &= |v\rangle = \mathbb{1}|v\rangle \;=\; \big(|1\rangle\langle 1| + |2\rangle\langle 2|\big)|v\rangle \;=\; |1\rangle\underbrace{\langle 1|v\rangle}_{=v_1} + |2\rangle\underbrace{\langle 2|v\rangle}_{=v_2} \\
&= e_1 \cdot v_1 + e_2 \cdot v_2 \;=\; \left(\begin{smallmatrix}v_1\\v_2\end{smallmatrix}\right).
\end{aligned}$$

Die komplexe Form der Fourier-Reihe lässt sich nun folgendermaßen mit Hilfe des

neuen Formalismus darstellen. Dabei gilt für $m, n \in \mathbb{Z}$:

$$\begin{aligned} |m\rangle &= |\mathrm{e}^{\mathrm{i}mx}\rangle \\ \langle m|n\rangle &= \delta_{m,n} \\ \big(|n\rangle\langle n|\big)^2 &= |n\rangle \underbrace{\langle n|n\rangle}_{=1}\langle n| \;=\; |n\rangle\langle n|\,. \end{aligned}$$

Nehmen wir nun an, dass die Funktion f als Vektor wie folgt aussieht:

$$|f\rangle \;=\; \sum_n c_n|n\rangle.$$

Daraus lässt sich nun die zuvor entwickelte Erweiterung herleiten:

$$\langle m|f\rangle \;=\; \langle m|\sum_n c_n|n\rangle \;=\; \sum_n c_n \underbrace{\langle m|n\rangle}_{=\delta_{m,n}} \;=\; c_m\,,$$

wobei wir die Rechenregeln für Skalarprodukte benutzt haben (s. nächstes Kapitel für eine kurze Wiederholung). Es bleibt aus der Summe nur der Term für $n = m$ bestehen und es ergibt sich für das Skalarprodukt $\langle m|f\rangle$ der Koeffizient c_m. Wird weiter angenommen, dass sich *jede* Funktion f auf diese Weise als (unendliche) Linearkombination darstellen lässt, mit $\{|m\rangle \mid m \in \mathbb{Z}\}$ als Basis (wir kommen später noch darauf zurück), dann folgt

$$|f\rangle \;=\; \underbrace{\Big(\sum_n |n\rangle\langle n|\Big)}_{=\mathbb{1}}|f\rangle \;=\; \sum_n |n\rangle\langle n|f\rangle \;=\; \sum_n \langle n|f\rangle\,|n\rangle \;=\; \sum_n c_n|n\rangle,$$

wobei wir über die passende Konvergenz später mehr sagen werden. Mit dieser Notation wird bei der Einführung von Hilbert-Räumen in Kapitel 3 Einiges systematischer, und das Rechnen etwas einfacher.

2.3 Vollständigkeit und Banach'scher Fixpunktsatz

Definition 2.10. Sei X eine Menge. Eine Abbildung $d : X \times X \to \mathbb{R}$ heißt *Metrik*, falls für alle $x, y, z \in X$ gilt:

$$\begin{aligned} d(x,y) &\geq 0\,, \quad \text{mit } d(x,y) = 0 \Leftrightarrow x = y \\ d(x,y) &= d(y,x) \qquad \text{(Symmetrie)} \\ d(x,y) &\leq d(x,z) + d(z,y) \quad \text{(Dreiecksungleichung)}. \end{aligned}$$

(X, d) heißt dann ein *metrischer Raum*.

Definition 2.11. Ein metrischer Raum (X, d) heißt *vollständig*, wenn jede Cauchy-Folge in X konvergent ist. Dabei heißt eine Folge $(a_n)_{n\in\mathbb{N}}$ *Cauchy-Folge*, falls gilt:

$$\forall \epsilon > 0 :\ \exists N \in \mathbb{N} :\ \forall m, n \geq N :\ d(a_n, a_m) < \epsilon\,.$$

Insbesondere ist jede konvergente Folge eine Cauchy-Folge, aber die Umkehrung erfordert die Vollständigkeit.

Definition 2.12. Eine Funktion $f : X \to X$ heißt *Lipschitz-stetig*, falls ein $L \geq 0$ existiert, so dass

$$d(f(x), f(y)) \leq L \cdot d(x, y)$$

für alle $x, y \in X$. Eine *Kontraktion* ist eine Lipschitz-stetige Abbildung mit $L < 1$, und L heißt dabei *Kontraktionskonstante*. Im Folgenden verwenden wir das Symbol ρ für solche Kontraktionskonstanten.

Satz 2.13. *(Banach'scher Fixpunktsatz)*
Sei (X, d) ein vollständiger metrischer Raum und $f : X \to X$ sei eine Kontraktion, mit Kontraktionskonstante $0 \leq \rho < 1$. Dann besitzt die Gleichung $f(x) = x$ genau eine Lösung $x^* \in X$, und jede Iterationsfolge mit $x_0 \in X$ und $x_{n+1} = f(x_n)$ für $n \geq 0$ konvergiert exponentiell schnell gegen x^*.

Beweis. Sei $f(x) = x$ und $f(y) = y$. Dann gilt

$$d(x, y) = d(f(x), f(y)) \leq \rho \cdot d(x, y)$$

mit $\rho < 1$ nach Voraussetzung. Das geht nur für $d(x, y) = 0$, also $x = y$. Damit ist gezeigt, dass es höchstens einen Fixpunkt x^* geben kann. Nun muss noch gezeigt werden, dass dieser wirklich existiert.

Sei $x_0 \in X$, $x_{n+1} = f(x_n)$ für $n \geq 0$. Mit der Dreiecksungleichung bekommen wir dann für $m \geq 1$:

$$\begin{aligned} d(x_n, x_{n+m}) &\leq \sum_{\ell=0}^{m-1} d(x_{n+\ell}, x_{n+\ell+1}) \leq d(x_0, x_1) \cdot \rho^n \cdot \underbrace{\sum_{\ell=0}^{m-1} \rho^\ell}_{=\frac{1-\rho^m}{1-\rho}} \\ &= d(x_0, x_1) \cdot \frac{\rho^n}{1-\rho} \cdot \underbrace{(1-\rho^m)}_{\leq 1} \leq \underbrace{d(x_0, x_1) \cdot \frac{\rho^n}{1-\rho}}_{\xrightarrow{n\to\infty} 0} \\ &\Rightarrow (x_n) \text{ ist Cauchy-Folge} \end{aligned}$$

Nach Voraussetzung ist der metrische Raum vollständig. Also konvergiert die Iterationsfolge $(x_n)_{n\in\mathbb{N}_0}$ gegen einen bestimmten Grenzwert $x^* = \lim_{n\to\infty} x_n$, wobei gilt:

$$x^* = \lim_{n\to\infty} x_n = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} f(x_n) = f(\lim_{n\to\infty} x_n) = f(x^*)\,.$$

Dabei ist der vorletzte Schritt eine Konsequenz der Stetigkeit von f an der Stelle x^*.

Dass die Konvergenz exponentiell schnell erfolgt, ergibt sich sofort aus der Relation $d(x^*, x_n) = \rho^n \cdot d(x^*, x_0)$, die eine Folge der Kontraktionseigenschaft ist. □

Für das Kontraktionsprinzip gibt es sowohl eine a priori als auch eine a posteriori Fehlerabschätzung. An dieser Stelle soll nur der Beweis für die a priori Abschätzung durchgeführt werden. Mit ihrer Hilfe kann man anschließend für einen gegebenen maximalen Abstand zur Lösung die maximale Iterationszahl im Vorfeld bestimmen.

Ergänzung 2.14. Die a priori Abschätzung lautet:

$$d(x^*, x_n) \leq \frac{\rho^n}{1-\rho}\, d(x_0, x_1)$$

Beweis. Wir erhalten mit der Dreiecksungleichung:

$$d(x^*, x_n) \;\leq\; d(x^*, x_{n+m}) + d(x_{n+m}, x_n) \;\leq\; \underbrace{\rho^{n+m}\, d(x^*, x_0)}_{\xrightarrow{m\to\infty} 0} + \underbrace{\frac{\rho^n}{1-\rho}\, d(x_0, x_1)}_{\text{unabhängig von } m}\,.$$

Es gilt also, mittels geeigneter Wahl von m:

$$\begin{aligned} \forall \epsilon > 0\colon \quad d(x^*, x_n) &< \epsilon + \frac{\rho^n}{1-\rho}\, d(x_0, x_1) \\ \overset{\epsilon \searrow 0}{\Longrightarrow} \quad d(x^*, x_n) &\leq \frac{\rho^n}{1-\rho}\, d(x_0, x_1)\,. \end{aligned}$$

□

Die a posteriori Abschätzung lautet

$$d(x^*, x_n) \;\leq\; \frac{\rho}{1-\rho}\, d(x_n, x_{n-1})$$

und gestattet die Abschätzung des Abstandes vom Fixpunkt aus den beiden letzten Iterationsschritten. Sie folgt aus einer kleinen Modifikation im Beweis von Satz 2.13.

Wir wollen uns nun zwei Anwendungen ansehen.

Beispiel 2.15. Zu berechnen sei die Lösung der transzendenten Gleichung $\mathrm{e}^{-\frac{x}{2}} - x = 0$ auf drei Nachkommastellen genau. Weiter sei der Startwert $x_0 = 1$ gegeben.

Zunächst muss eine Fixpunktgleichung aufgestellt werden. Hier geht $\mathrm{e}^{-x/2} = x$. Damit lautet die Iteration $x_0 = 1$ mit

$$x_{n+1} \;=\; \mathrm{e}^{-\frac{x_n}{2}} \quad \text{für } n \geq 0\,.$$

Anschließend kann die Iteration beginnend mit dem Startwert durchgeführt werden, bis eine Genauigkeit auf 3 Nachkommastellen erreicht ist. Dies ist etwa ab der 8. Iteration

der Fall, wie durch die a priori Abschätzung bestimmt werden kann:

$$\begin{aligned} x_1 &= \mathrm{e}^{-\frac{x_0}{2}} \approx 0,606531 \\ x_2 &= \mathrm{e}^{-\frac{x_1}{2}} \approx 0,738403 \\ &\vdots \\ x_8 &\approx 0,703533 \\ x_9 &\approx 0,703444 \\ x_{10} &\approx 0,703475\,. \end{aligned}$$

Beispiel 2.16. (Cantor-Menge)
Wir betrachten den metrischer Raum (X, d), wobei

$$X = \{\text{Menge der nicht-leeren kompakten Teilmengen von } \mathbb{R}\}$$

und d die Distanz zwischen zwei Mengen sei. Gut geeignet ist dafür die Hausdorff-Metrik, da (X, d) dann vollständig ist.

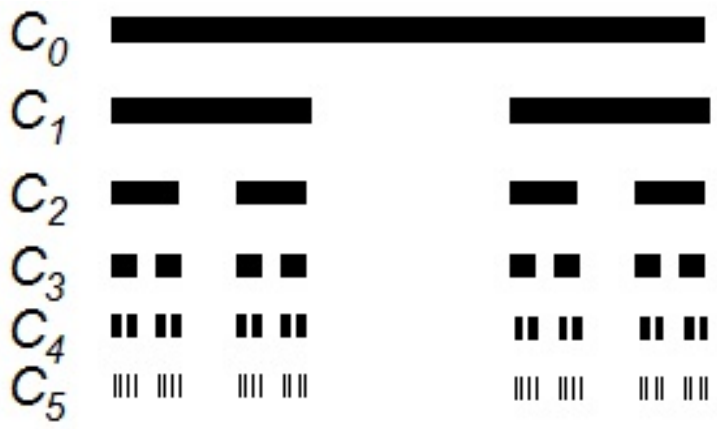

Abbildung 5: Skizze für das Intervall $C_0 = [0, 1]$ und anschließende Cantor-Iterationen.

Weiter seien zwei Funktionen

$$f_0(x) = \frac{1}{3}x \quad \text{und} \quad f_1(x) = \frac{1}{3}x + \frac{2}{3}$$

und eine kompakte Teilmenge $C_0 \in X$ als Startmenge gegeben. Die Menge $\mathcal{C}$, welche sich als Limes der Iterationsfolge

$$C_{n+1} = f_0(C_n) \cup f_1(C_n) \qquad \text{für } n \geq 0$$

ergibt, heißt *Cantor-Menge*. Die Iteration ist eine Kontraktion bzgl. der Hausdorff-Metrik, mit Kontraktionskonstante $\frac{1}{3}$. Daher ist $\mathcal{C}$ eine wohldefinierte kompakte Teilmenge von $\mathbb{R}$

(tatsächlich mit $\mathcal{C} \subset [0,1]$). Sie besitzt überabzählbar viele Elemente, hat aber dennoch Maß 0. Details hierzu findet man in den meisten Lehrbüchern der Analysis.

In ähnlicher Weise kann man auch andere Fraktale beschreiben, Abb. 6 zeigt dazu eine Illustration für das Beispiel des Sierpinski-Frakals in der Ebene.

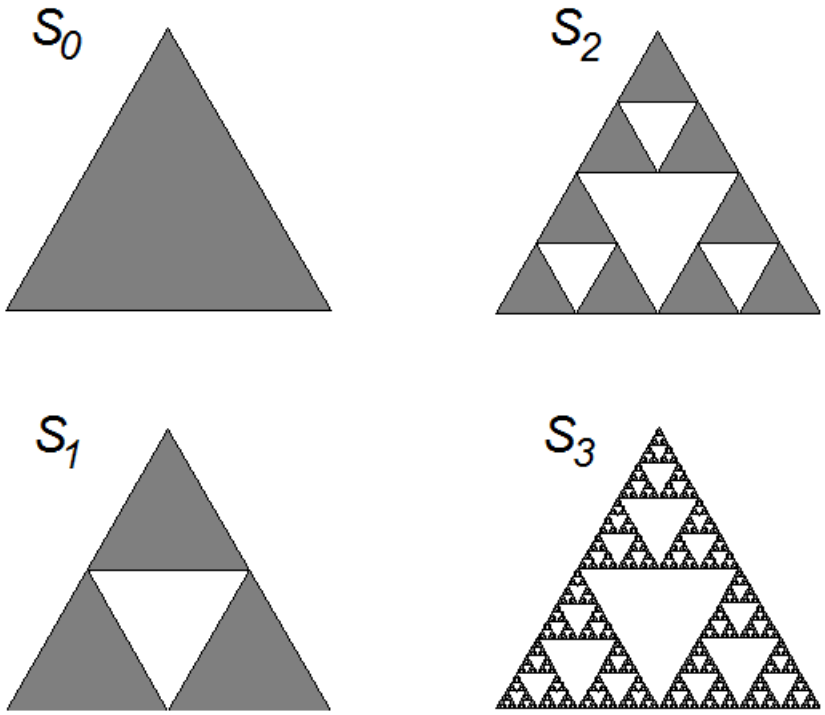

Abbildung 6: Skizze der ersten Iterationen zur Sierpinski-Menge. Sie beschreibt wie die Cantor-Menge ein Fraktal, welches im Grenzwert das Maß 0 hat, aber dennoch überabzählbar viele Punkte enthält.

3 Hilbert-Räume und Fourier-Reihen

3.1 Allgemeine Eigenschaften

Zunächst benötigt man die Definition einer *Norm*, da diese ein wichtiger Bestandteil der hier behandelten Vektorräume sein wird.

Definition 3.1. Sei V ein Vektorraum über $\mathbb{R}$ oder $\mathbb{C}$. Eine Abbildung $\|\cdot\|\colon V \longrightarrow \mathbb{R}$ heißt *Norm*, wenn für beliebige Vektoren u, v und Zahlen α aus $\mathbb{R}$ bzw. $\mathbb{C}$ folgende Eigenschaften gelten:

$$\begin{aligned} \|v\| &\geq 0\,, \quad \text{mit } \|v\| = 0 \Leftrightarrow v = 0 \\ \|\alpha v\| &= |\alpha|\|v\| \\ \|v+u\| &\leq \|v\| + \|u\| \end{aligned}$$

Mit $L^2([-\pi,\pi],\mathbb{C})$ wird der $\mathbb{C}$-Vektorraum der auf $[-\pi,\pi]$ quadratintegrablen Funktionen bezeichnet, d.h. $\int_{-\pi}^{\pi} \left|f(x)^2\right| \mathrm{d}x$ existiert im Sinne von Lebesgue. Dabei benutzen wir nun auch den Begriff des Skalarproduktes.

Definition 3.2. Sei V ein komplexer Vektorraum. Eine Abbildung $\langle\cdot|\cdot\rangle\colon V \times V \longrightarrow \mathbb{C}$ heißt *Skalarprodukt* auf V, wenn folgende Eigenschaften für beliebige Elemente $f, g, h \in V$ und Zahlen $\alpha \in \mathbb{C}$ erfüllt sind:

$$\begin{aligned} \langle f|g+h\rangle &= \langle f|g\rangle + \langle f|h\rangle \\ \langle f|\alpha g\rangle &= \alpha\langle f|g\rangle \\ \langle g|f\rangle &= \overline{\langle f|g\rangle} \quad \text{(komplex konjugiert)} \\ \langle f|f\rangle &\geq 0\,, \quad \text{mit } \langle f|f\rangle = 0 \Leftrightarrow f = 0\,. \end{aligned}$$

Man beachte, dass $\langle \alpha f|g\rangle = \overline{\alpha}\,\langle f|g\rangle$ eine Folge der geforderten Eigenschaften ist. Ist $\langle\cdot|\cdot\rangle$ ein Skalarprodukt, so definiert $\|f\| := \sqrt{\langle f|f\rangle}$ eine Norm auf V im Sinne von Definition 3.1. Dabei gilt dann die *Schwarz'sche Ungleichung*

$$|\langle f|g\rangle| \leq \|f\| \cdot \|g\|\,, \tag{23}$$

die eine der wichtigen Ungleichungen der Mathematik ist.

Mit $\langle f|g\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} \overline{f(x)}\, g(x)\,\mathrm{d}x$ ist auf dem Vektorraum $L^2([-\pi,\pi],\mathbb{C})$ ein Skalarprodukt definiert (siehe auch [3], Bsp. 18.5.). Bzgl. der zugehörigen Norm ist dieser Raum dann zudem vollständig gemäß Def. 2.11, also alle Cauchy-Folgen konvergieren. Vollständige $\mathbb{C}$-Vektorräume mit einem Skalarprodukt, das die Norm induziert, nennt man auch *Hilbert-Räume*.

Satz 3.3. *Jede 2π-periodische, stetige Funktion f lässt sich als konvergente Fourier-Reihe darstellen,*

$$f(x) = \sum_{n\in\mathbb{Z}} c_n\, \mathrm{e}^{\mathrm{i}nx} \quad \textit{mit} \quad c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(s)\, \mathrm{e}^{-\mathrm{i}ns}\,\mathrm{d}s\,,$$

wobei es sich um Konvergenz im Mittel handelt, also Konvergenz bezüglich $\|\cdot\| = \sqrt{\langle\cdot|\cdot\rangle}$.

Bemerkung 3.4. Dies ist ein sehr schöner Konvergenzbegriff, welcher systematischer als punktweise Konvergenz und effizienter als gleichmäßige Konvergenz ist. Zu beachten ist jedoch, dass nicht überall punktweise Konvergenz vorliegen muss. Wir werden später in diesem Kapitel noch weiter darauf eingehen, und etwas zu den Gründen für die Gültigkeit sagen.

Jetzt zeigt sich noch ein Schönheitsfehler: Bezüglich der Norm $\|.\|$ ist der Raum der stetigen Funktionen $C^0([0,2\pi],\langle\cdot|\cdot\rangle)$ nicht vollständig. Das heißt, dass der Limes einer im Mittel konvergenten Folge von Funktionen aus $C^0([0,2\pi],\langle\cdot|\cdot\rangle)$ selbst nicht wieder in diesem Raum liegen muss. Man betrachte z.B. eine Funktionen-Folge f_n, welche wie folgt aussieht:

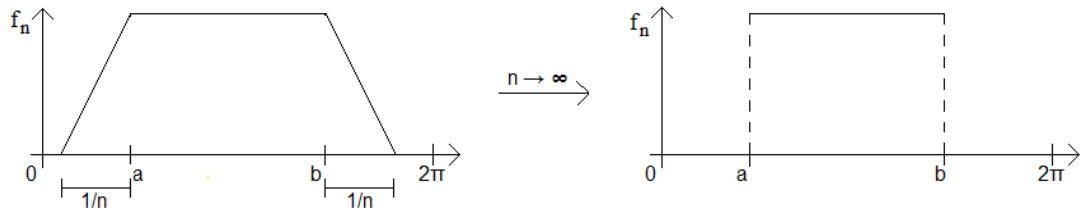

Abbildung 7: Skizze zum Schönheitsfehler der Unvollständigkeit.

Dann ergibt sich, dass der Grenzwert für $n \to \infty$ nicht stetig ist. Eine elegante Lösung dieses Problems, die zugleich eine wichtige Klasse vollständiger Räume liefert, basiert auf dem Integralbegriff von *Lebesgue*. Die übliche Betrachtungsweise eines Integrals nach

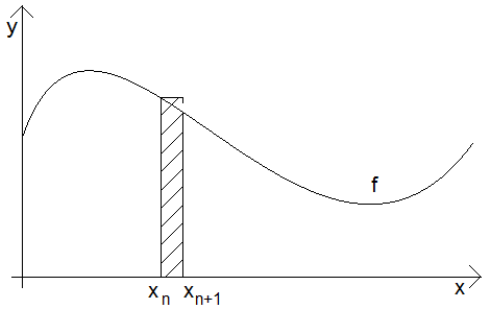

Abbildung 8: Skizze zum klassischen Riemann-Integral. Approximation von f durch Rechtecke entlang der x-Achse.

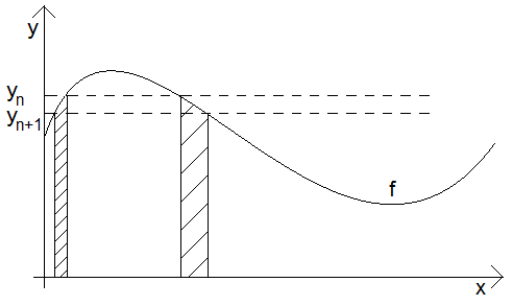

Abbildung 9: Skizze zum Integral nach Lebesgue. Approximation von f durch Rechtecke mit Auswahl entlang der y-Achse.

Riemann ist in Abbildung 8 zu sehen. Es werden sehr schmale Rechtecke zwischen zwei nah aneinander liegenden x-Werten mit Hilfe der entsprechenden Funktionswerte aufsummiert. Für verschwindend kleine Rechtecke erhält man das Integral der Funktion. Nach Lebesgue wird diese Technik zwar beibehalten, allerdings wird zwischen x- und y-Werten getauscht. Es wird also die Breite der Rechtecke anhand der Funktionswerte

vorgegeben und mittels aller zugehörigen x-Werte ggfs. mehrere Rechtecke pro Intervall berechnet. Damit ist der Raum

$$L^2([0,2\pi],\langle\cdot|\cdot\rangle) \;=\; \Big\{ f \Big| \int_0^{2\pi} |f(x)|^2 \,\mathrm{d}x \text{ ex. im Lebesgue-Sinne} \Big\}$$

vollständig. Aber auch hier gibt es noch eine (kleine) Einschränkung. Gegeben sei die Funktion f (dargestellt in Abbildung 10)

$$f(x) \;=\; \begin{cases} 1, & x \in [0,2\pi] \cap \mathbb{Q}, \\ 0, & \text{sonst auf } [0,2\pi]. \end{cases}$$

Da $\mathbb{Q}$ abzählbar ist, stimmt f fast überall (im Sinne der Lebesgue-Theorie) mit $g \equiv 0$ überein, und die beiden Funktionen werden als gleich angesehen.

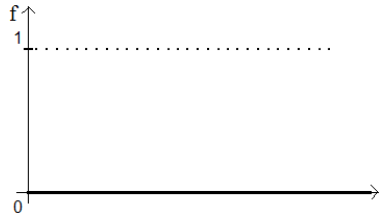

Abbildung 10: Skizze der Funktion f.

Dies spielt in der Praxis allerdings keine Rolle, da nie an einzelnen Punkten, sondern vielmehr auf sehr kleinen Intervallen mittelnd gemessen werden. Dies liegt an der endlichen Auflösung der Messgeräte.

Die Vektorräume, welche im Folgenden betrachtet werden, sehen folgendermaßen aus.

$$L^p([a,b],\mathbb{C}) \;:=\; \Big\{ f\colon [a,b] \longrightarrow \mathbb{C} \Big| \int_a^b |f(x)|^p \,\mathrm{d}x \text{ ex. im Lebesgue-Sinne} \Big\}$$

für $1 \le p < \infty$, und

$$L^\infty([a,b],\mathbb{C}) \;:=\; \big\{ f\colon [a,b] \longrightarrow \mathbb{C} \,\big|\, \sup_x |f(x)| < \infty \big\}.$$

Dies sind stets Banach-Räume. Einige in diesem Skript besprochene Eigenschaften gelten auch für $1 \le p < \infty$, dennoch ist nur für $p = 2$ der entsprechende Raum ein Hilbert-Raum. Hier ist die Norm durch das Skalarprodukt gegeben.

In einem Vektorraum ist eine Basis in der Regel sehr nützlich. In einem Hilbert-Raum ist es besonders hilfreich, eine Orthonormal-Basis zu haben.

Definition 3.5. Sei V ein Vektorraum über $\mathbb{R}$ oder $\mathbb{C}$, mit Skalarprodukt $\langle\cdot|\cdot\rangle$. Eine Menge an Vektoren $\{b_i \mid i \in I\}$ mit endlicher oder abzählbarer Indexmenge I heißt *Orthonormalsystem* (ONS), falls

$$\langle b_i \,|\, b_j \rangle \;=\; \delta_{ij}$$

für alle $i, j \in I$. Ein ONS heißt *maximal*, falls einzig das Nullelement orthogonal zu allen b_i des ONS ist.

Ist V ein Hilbert-Raum, so ist ein maximales ONS eine *Basis*, es kann also jedes Element von V als (konvergente) Linearkombination in den Elementen des ONS dargestellt werden. Insbesondere ist ein ONS eine Basis, wenn V endliche Dimension hat.

3.2 Die Gauß'sche Approximationsaufgabe

Es sei eine Funktion f gegeben, die Element eines Hilbert-Raums sei. Außerdem sei $\{e_1, ..., e_n\}$ ein ONS, aber nicht notwendig maximal. Insbesondere verlangen wir an dieser Stelle nicht, dass es sich um eine Basis handelt. Wir wollen nun $\alpha_1, ..., \alpha_n \in \mathbb{C}$ so bestimmen, dass

$$\left\| f - \sum_{i=1}^{n} \alpha_i e_i \right\| \overset{(!)}{=} \min.$$

Anders ausgedrückt: Wie kann man f also mit endlich vielen Termen am besten approximieren? Dies nennt man die *Gauß'sche Approximationsaufgabe*. Mit Hilfe der Vorarbeit zu Vektorräumen lässt sich diese Frage wie folgt beantworten.

Satz 3.6. *Die Gauß'sche Approximationsaufgabe wird eindeutig gelöst duch*

$$\alpha_\ell = \langle e_\ell \,|\, f \rangle, \quad \text{für } 1 \le \ell \le n,$$

und $f - \sum_{\ell=1}^{n} \alpha_\ell e_\ell$ *steht senkrecht auf* e_i *für alle* $1 \le i \le n$. *Dabei gilt die Bessel'sche Ungleichung*

$$\sum_{\ell=1}^{n} \left| \langle e_\ell \,|\, f \rangle \right|^2 \le \|f\|^2.$$

Beweis. Wir beginnen mit einer einfachen Rechnung:

$$\begin{aligned} 0 \quad &\le \quad \left\| f - \sum_{\ell=1}^{n} \alpha_\ell e_\ell \right\|^2 = \left\langle f - \sum_{\ell=1}^{n} \alpha_\ell e_\ell \,\middle|\, f - \sum_{\ell=1}^{n} \alpha_\ell e_\ell \right\rangle \\ &= \quad \langle f \,|\, f \rangle - \sum_{\ell=1}^{n} \alpha_\ell \langle f \,|\, e_\ell \rangle - \sum_{\ell=1}^{n} \overline{\alpha_\ell} \langle e_\ell \,|\, f \rangle + \underbrace{\sum_{\ell=1}^{n} \sum_{\ell'=1}^{n} \overline{\alpha_\ell} \alpha_{\ell'} \langle e_\ell \,|\, e_{\ell'} \rangle}_{= \sum_{\ell=1}^{n} |\alpha_\ell|^2} \\ &\overset{\text{quadr. Erg.}}{=} \langle f \,|\, f \rangle + \underbrace{\sum_{\ell=1}^{n} \left| \langle e_\ell \,|\, f \rangle - \alpha_\ell \right|^2}_{\ge 0, \text{ einziger variabler Term}} - \sum_{\ell=1}^{n} \left| \langle e_\ell \,|\, f \rangle \right|^2. \end{aligned}$$

Der Abstand wird also genau dann minimal, wenn $\alpha_\ell = \langle e_\ell \,|\, f \rangle$ für alle $1 \le \ell \le n$ gilt. Damit folgt dann auch sofort die Bessel'sche Ungleichung.
Ist $z = f - \sum_{\ell=1}^{n} \langle e_\ell \,|\, f \rangle \, e_\ell$, so gilt

$$\langle e_m \,|\, z \rangle = \langle e_m \,|\, f \rangle - \sum_{\ell=1}^{n} \langle e_\ell \,|\, f \rangle \underbrace{\langle e_m \,|\, e_\ell \rangle}_{= \delta_{m,\ell}} = \langle e_m \,|\, f \rangle - \langle e_m \,|\, f \rangle = 0.$$

Damit ist auch die zweite Behauptung gezeigt. □

Es zeigt sich also, dass die beste Approximation an die Funktion f gegeben ist durch die Teilsummen der Fourier-Reihe, wobei mit 'beste Approximation' der Abstand bezüglich der L^2-Norm gemeint ist.

Folgerung 3.7. *Es gilt die Bessel'sche Gleichung:*

$$\Big\| f - \sum_{\ell=1}^{n} \langle e_\ell \,|\, f \rangle \, e_\ell \Big\|^2 \;=\; \|f\|^2 \;-\; \sum_{\ell=1}^{n} \big| \langle e_\ell \,|\, f \rangle \big|^2$$

wie sich direkt aus der Rechnung im obigen Beweis ergibt.

Nun können wir eine Erweiterung auf ein abzählbares ONS wie folgt formulieren.

Satz 3.8. *Sei V ein unendlich-dimensionaler komplexer Vektorraum mit Skalarprodukt, und sei $\{e_i \mid i \in \mathbb{N}\}$ ein ONS in V. Dann gilt:*

1. $\sum_{\ell=1}^{\infty} \alpha_\ell e_\ell$ *ist Cauchy-Folge* $\;\Leftrightarrow\;$ $\sum_{\ell=1}^{\infty} |a_\ell|^2$ *konvergiert;*
2. $f = \sum_{\ell=1}^{\infty} \alpha_\ell e_\ell \;\Rightarrow\; \alpha_\ell = \langle e_\ell \,|\, f \rangle$ *für alle ℓ, und es liegt unbedingte Konvergenz vor;*
3. $\sum_{\ell=1}^{\infty} \langle e_\ell \,|\, f \rangle \, e_\ell = f \;\Leftrightarrow\; \sum_{\ell=1}^{\infty} \big| \langle e_\ell \,|\, f \rangle \big|^2 = \|f\|^2$ *(Parseval'sche Gleichung).*

Beweis. Die 1. Behauptung folgt aus einer Rechnung, mit $1 \le m \le n$:

$$\Big\| \sum_{\ell=m}^{n} \alpha_\ell e_\ell \Big\|^2 \;=\; \Big\langle \sum_{k=m}^{n} \alpha_k e_k \,\Big|\, \sum_{\ell=m}^{n} \alpha_\ell e_\ell \Big\rangle \;=\; \sum_{k,\ell=1}^{n} \overline{\alpha_k}\, \alpha_\ell \underbrace{\langle e_k | e_\ell \rangle}_{\delta_{k,\ell}} \;=\; \sum_{\ell=m}^{n} |\alpha_\ell|^2.$$

Für die 2. Behauptung: Sei $f_n = \sum_{\ell=1}^{n} \alpha_\ell e_\ell$ und $f_n \xrightarrow{n\to\infty} f$ bzgl. der Norm $\|.\|$. Dann ist

$$\big| \langle e_k \,|\, f_n \rangle - \langle e_k \,|\, f \rangle \big| \;=\; \big| \langle e_k \,|\, f_n - f \rangle \big| \;\le\; \underbrace{\|e_k\|}_{=1} \cdot \|f_n - f\| \;=\; \underbrace{\|f_n - f\|}_{\xrightarrow{n\to\infty} 0}$$

Da aber $\langle e_k | f_n \rangle = \alpha_k$ für alle $n \ge k$ nach Satz 3.6, folgt die Formel für den Koeffizienten aus der Stetigkeit der Linearform $g \mapsto \langle e_k | g \rangle$. Die unbedingte Konvergenz stellen wir einen Moment zurück.

Die 3. Behauptung folgt nun sofort aus der Bessel'schen Gleichung.

Somit bleibt die 'Umordnungsresistenz' von der 2. Behauptung zu zeigen. Wir haben bisher gezeigt:

$$f = \sum_{\ell=1}^{\infty} \alpha_\ell e_\ell \;\Rightarrow\; f = \sum_{\ell=1}^{\infty} \langle e_\ell \,|\, f \rangle \, e_\ell \;\Rightarrow\; \sum_{\ell=1}^{\infty} \underbrace{\big| \langle e_\ell \,|\, f \rangle \big|^2}_{\ge 0} \text{ konvergiert absolut.}$$

Daher können wir in der letzten Summe beliebig umordnen (was einer Umnummerierung der Elemente unseres ONS entspricht), ohne den Grenzwert zu ändern. Aus der 3. Behautpung, die wir ja schon bewiesen haben, folgt dann aber, dass $\sum_{\ell=1}^{\infty} \langle e_\ell \,|\, f \rangle \, e_\ell$ auch bei beliebiger Umordnung stets gegen f konvergiert. □

3.3 Fourier-Reihen: Ergänzungen

Im Hinblick auf die weitere Entwicklung wollen wir eine neue Notation einführen.

Definition 3.9. Betrachte eine Funktion $f \in L^1([-\pi,\pi],\mathbb{C})$ mit periodischer Fortsetzung. Die Koeffizienten

$$\widehat{f}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\,\mathrm{e}^{-\mathrm{i}nx}\,\mathrm{d}x \tag{24}$$

mit $n \in \mathbb{Z}$ heißen die *Fourier-Koeffizienten* von f.

Bemerkung 3.10. Wir halten kurz einige Eigenschaften fest:

- Die Fourier-Koeffizienten aus (24) sind wohl definiert wegen

  $$|\mathrm{e}^{-\mathrm{i}nx} f(x)| = |f(x)|,$$

  weswegen auch $x \mapsto \mathrm{e}^{-\mathrm{i}nx} f(x)$ eine L^1-Funktion auf dem Intervall $[-\pi,\pi]$ ist.

- Die Fourier-Koeffizienten von f sind beschränkt, denn

  $$|\widehat{f}(x)| \;\le\; \frac{1}{2\pi}\|f\|_1 \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi}|f(x)|dx \;<\; \infty.$$

- Die Abbildung $f \mapsto \widehat{f}$ ist linear, denn

  $$(\widehat{\alpha f + \beta g})(n) \;=\; \alpha\widehat{f}(n) + \beta\widehat{g}(n)$$

  gilt für alle $f, g \in L^1([-\pi,\pi],\mathbb{C})$ und alle $\alpha, \beta \in \mathbb{C}$.

Satz 3.11. *(Eindeutigkeitssatz)*
Gegeben seien zwei Funktionen $f, g \in L^1([-\pi,\pi],\mathbb{C})$ mit $\widehat{f}(n) = \widehat{g}(n)$ für alle $n \in \mathbb{Z}$. Dann folgt $f = g$ in L^1. Dies bedeutet, dass $f(x) = g(x)$ für fast alle $x \in [-\pi,\pi]$ gilt.

Für die in Gl. (24) definierten Fourier-Koeffizienten gibt es allerdings noch ein Problem. Die Fourier-Reihe zu den Koeffizienten $\widehat{f}$

$$\sum_{n\in\mathbb{Z}} \mathrm{e}^{\mathrm{i}nx}\widehat{f}(x)$$

muss nämlich nicht konvergieren! Nach Kolmogorov (1926) existiert eine Funktion $f \in L^1([-\pi,\pi],\mathbb{C})$ dessen Fourier-Reihe nirgends konvergiert. Das ändert sich, wenn wir $f \in L^p([-\pi,\pi],\mathbb{C})$ mit $p > 1$ betrachten. Zunächst gilt:

Satz 3.12. *(Carleson, 1966)*
Die Fourier-Reihe einer periodisch fortgesetzten L^2-Funktion ist fast überall punktweise konvergent.

Bemerkung 3.13. Der Satz von Carleson klärt eine Vermutung von Lusin (1915). Hunt (1968) zeigte dann, dass sich dies auf Funktionen $f \in L^p([-\pi,\pi],\mathbb{C})$ für $1 < p < \infty$ erweitern lässt.

Dabei ist der Begriff 'fast überall auf [a,b]' mathematisch präzise. Es bedeutet konkret, dass die Aussage überall außer auf einer Teilmenge vom Maß 0 gilt. Man beachte, dass nicht nur abzälbare Mengen vom Maß 0 sind, es gibt auch andere. Eine überabzählbare Ausnahmemenge ist z.B. die Cantor-Menge aus Beispiel 2.16.

Die Idee dieses Begriffs kann man sich auch mittels der Wahrscheinlichkeitstheorie näher bringen. Es gibt Ereignisse mit einer Wahrscheinlichkeit von 1, wie z.B. das Werfen von Kopf bei einem unendlich oft wiederholten Münzwurf. Man spricht davon, dass 'fast immer' irgendwann Kopf fallen wird.

Nun kommen wir zu einem sehr wichtigen Konzept.

Definition 3.14. Es seien $f, g \in L^1([-\pi,\pi],\mathbb{C})$. Die Funktion $f * g$, die durch

$$(f*g)(t) \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\tau)\, g(t-\tau)\,\mathrm{d}\tau$$

definiert ist, heißt Faltung von f und g (dabei sind f, g für die Definition des Integrals als periodisch fortgesetzt gedacht).

Lemma 3.15. *Es seien $f, g \in L^1([-\pi,\pi],\mathbb{C})$. Dann existiert $(f*g)(t)$ für fast alle $t \in [-\pi,\pi]$, man hat $f*g \in L^1([-\pi,\pi],\mathbb{C})$, und es gelten folgende Eigenschaften:*

$$\begin{aligned} f*g &= g*f && \textit{(kommutativ)}\\ (f*g)*h &= f*(g*h) && \textit{(assoziativ)}\\ f*(\alpha g+\beta h) &= \alpha(f*g)+\beta(f*h) && \textit{(linear)} \end{aligned}$$

für beliebige $f, g, h \in L^1([-\pi,\pi],\mathbb{C})$ und $\alpha, \beta \in \mathbb{C}$.

Beweis. siehe Übung. □

Die Faltung hat starke Konsequenzen:

Satz 3.16. *(Faltungssatz)*
Für beliebige $f, g \in L^1([-\pi,\pi],\mathbb{C})$ gilt $\widehat{(f*g)}(n) = \widehat{f}(n)\cdot\widehat{g}(n)$ für alle $n \in \mathbb{Z}$.

Beweis. Dies folgt wieder aus einer kleinen Rechnung:

$$\begin{aligned} 4\pi^2\widehat{(f*g)}(n) &= \int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}nt}\int_{-\pi}^{\pi} f(\tau)\, g(t-\tau)\,\mathrm{d}\tau\,\mathrm{d}t\\ &\overset{\text{Fubini}}{=} \int_{-\pi}^{\pi}\underbrace{\int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}n(t-\tau)}\, g(t-\tau)\,\mathrm{d}t}_{=\int_{-\pi}^{\pi}\mathrm{e}^{-\mathrm{i}nt'}g(t')\,\mathrm{d}t' = 2\pi\,\widehat{g}(n)}\,\mathrm{e}^{-\mathrm{i}n\tau} f(\tau)\,\mathrm{d}\tau\\ &= 2\pi\,\widehat{g}(n)\int_{-\pi}^{\pi}\mathrm{e}^{-\mathrm{i}n\tau} f(\tau)\,\mathrm{d}\tau \;=\; 4\pi^2\,\widehat{f}(n)\widehat{g}(n). \end{aligned}$$

Dabei wird im vorletzten Schritt die 2π-Periodizität von g verwendet, um nach einer Verschiebung um τ den Funktionswert an $t' \in [-\pi,\pi]$ zu berechnen. □

Beispiel 3.17. Ein einfaches Beispiel ist die Faltung der charakteristischen Funktion einer Menge mit sich selbst. Gegeben sei dazu

$$f(x) := \begin{cases} 1, & \text{falls } |x| \leq \frac{\pi}{2}, \\ 0, & \text{sonst.} \end{cases}$$

Die Faltung $f * f$ kann nun in mehrere Bereiche aufgeteilt werden, wobei der Wert zwischen $-\pi \leq t \leq 0$ zunimmt und im Bereich $0 \leq t \leq \pi$ wieder abfällt. Das Ergebnis lautet:

$$\big(f * f\big)(x) \;=\; \frac{1}{2\pi} \begin{cases} 0, & x \leq -\pi, \\ \pi + x, & -\pi < x \leq 0, \\ \pi - x, & 0 < x \leq \pi, \\ 0, & \pi < x. \end{cases}$$

Die zugehörige Rechnung sollte man selber einmal durchführen, um sich mit der Faltung vertraut zu machen. Der Vorfaktor $\frac{1}{2\pi}$ kommt von der oben gewählten Definition der Faltung.

4 Fourier-Transformation

4.1 Von der Fourier-Reihe zur Fourier-Transformation

Bislang wurden Funktionen f betrachtet, die aus reinen Wellen zusammengesetzt sind, wie die Sinus- oder die Cosinus-Funktion. Dabei traten wegen der gewählten Periode von 2π nur Argumente der Form $2\pi kx$ mit $k \in \mathbb{N}_0$ auf. Die jeweiligen Anteile der reinen Wellen an Funktion f können durch die Fourier-Koeffizienten dargestellt werden. Was passiert aber, wenn f aus ganz verschiedenen Wellen zusammengesetzt ist, die untereinander nicht in diesem Sinne vergleichbar sind?

Um die Fourier-Reihe weiter zu verallgemeinern, müssen zunächst andere Perioden betrachtet werden. Bislang war die Fourier-Reihe für 2π-periodische Funktionen definiert, also für Funktionen, die wir z.B. auf $[-\pi, \pi]$ definiert und dann auf ganz $\mathbb{R}$ periodisch forgesetzt haben. Zunächst wird also ein allgemeineres Intervall $[-a, a]$ betrachtet, es soll also $f(x + 2a) = f(x)$ gelten. Aus Gl. (19) wird dann

$$f(x) = \sum_{n\in\mathbb{Z}} \tilde{c}_n \, \mathrm{e}^{\mathrm{i}nx\frac{\pi}{a}} \quad \text{mit} \quad \tilde{c}_n = \frac{1}{2a} \int_{-a}^{a} \mathrm{e}^{-\mathrm{i}nx\frac{\pi}{a}} f(x) \, \mathrm{d}x$$

und es gilt wieder Konvergenz im L^2-Sinne, hier für alle Funktionen aus $L^2([-a, a], \mathbb{C})$. Durch Umschreiben einiger Terme erhält man neue Funktionen:

$$f(x) = \sum_{n\in\mathbb{Z}} \underbrace{\Phi(n\tfrac{\pi}{a})}_{=\tilde{c}_n} \mathrm{e}^{\mathrm{i}nx\frac{\pi}{a}} \quad \text{mit} \quad \Phi(n\tfrac{\pi}{a}) = \frac{1}{2a} \int_{-a}^{a} \mathrm{e}^{-\mathrm{i}nx\frac{\pi}{a}} f(x) \, \mathrm{d}x$$

Dies kann man nun folgendermaßen für eine Heuristik verwenden: Betrachtet man nun $a \to \infty$, wird aus $\Phi(n\frac{\pi}{a})$ fast eine kontinuierliche Funktion. Außerdem bleibt $\frac{1}{2a} f(x)$ bezüglich der Integration auf $[-a, a]$ etwa gleich groß. Die soll folgende Korrespondenz motivieren:

$$\begin{aligned} \tfrac{1}{2a} f(x) &\curvearrowright g(x), \quad \text{mit } g \in L^1(\mathbb{R}, \mathbb{C}), \\ \Phi(n\tfrac{\pi}{a}) &\curvearrowright h(y), \end{aligned}$$

und man betrachtet

$$h(y) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}xy} g(x) \, \mathrm{d}x \quad \text{und} \quad g(x) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}yx} h(y) \, \mathrm{d}y. \tag{25}$$

Proposition 4.1. *Ist $g \in L^1(\mathbb{R}, \mathbb{C}) \widehat{=} L^1(\mathbb{R})$, so existiert die Abbildung h für alle $y \in \mathbb{R}$ und ist stetig. Dabei gilt:*

$$\|h\|_\infty = \sup_{y\in\mathbb{R}} |h(y)| \le \|g\|_1 := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} |g(x)| \, \mathrm{d}x.$$

Außerdem gilt $\lim_{|y|\to\infty} h(y) = 0$.

Beweis. Für die erste Behauptung machen wir eine kleine Rechnung:

$$\|h\|_\infty \;=\; \sup_y \frac{1}{\sqrt{2\pi}} \underbrace{\left| \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}xy} g(x)\,\mathrm{d}x \right|}_{\leq \int_{\mathbb{R}} \underbrace{|\mathrm{e}^{-\mathrm{i}xy}|}_{=1} |g(x)|\,\mathrm{d}x} \;\leq\; \|g\|_1 .$$

Für den Beweis der zweiten Behauptung, die auch als *Riemann–Lebesgue-Lemma* bekannt ist, verweisen wir auf [5] oder [7]. □

4.2 Fourier-Transformation

Nun wollen wir obige Heuristik präzise fassen. Wir tun dies gleich im $\mathbb{R}^n$, wobei dann $tx = \sum_i t_i x_i$ das Skalarprodukt zweier Vektoren bezeichnen soll.

Definition 4.2. Sei $f \in L^1(\mathbb{R}^n)$, also $\int_{\mathbb{R}^n} |f(x)|\,\mathrm{d}^n x < \infty$. Dann heißt

$$\widehat{f}(t) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}tx} f(x)\,\mathrm{d}^n x$$

die *Fourier-Transformation* von f.

Zu beachten ist, dass diese Definition nicht einheitlich so verwendet wird. Es gibt verschiedene Anwendungen aus Mechanik, Mathematik etc. in denen der Vorfaktor vor dem Integral und der Faktor im Exponenten aus praktischen Erwägungen anders gewählt werden. Für unsere Formulierung gilt folgendes Resultat.

Satz 4.3. *Ist $f \in L^1(\mathbb{R}^n)$, so existiert die Fourier-Transformierte $\widehat{f}$ gemäß Definition* 4.2. *Für beliebige $f, g, h \in L^1(\mathbb{R}^n)$ und $\alpha \in \mathbb{R}^n$ gelten dann die folgenden Eigenschaften:*

1. $g(x) = f(x)\,\mathrm{e}^{\mathrm{i}\alpha x} \quad\Rightarrow\quad \widehat{g}(t) = \widehat{f}(t-\alpha)$
2. $g(x) = f(x-\alpha) \quad\Rightarrow\quad \widehat{g}(t) = \mathrm{e}^{-\mathrm{i}\alpha t}\widehat{f}(t)$
3. *Ist $h = f * g$, so gilt der* Faltungssatz, *also $\widehat{h} = \widehat{f*g} = \widehat{f}\cdot\widehat{g}$, wobei die Faltung hier durch*
$$(f*g)(t) := \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} f(\tau)\,g(t-\tau)\,\mathrm{d}^n\tau$$
definiert ist.

Beweis. Die Existenzaussage ist eine Folgerung von Proposition 4.1, die sich unmittelbar auf den n-dimensionalen Fall erweitert.

Für die erste Eigenschaft machen wir eine kleine Rechnung:

$$\begin{aligned} \widehat{g}(t) &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}tx} f(x)\,\mathrm{e}^{\mathrm{i}\alpha x}\,\mathrm{d}^n x \\ &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}x(t-\alpha)} f(x)\,\mathrm{d}^n x \;=\; \widehat{f}(t-\alpha) \end{aligned}$$

Der Beweis für die zweite Eigenschaft verläuft analog.

Der Faltungssatz folgt aus folgender Rechnung:

$$\begin{aligned}
\widehat{h}(t) &= \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}tx} \int_{\mathbb{R}^n} f(x-y)\, g(y)\, \frac{\mathrm{d}^n y\, \mathrm{d}^n x}{(2\pi)^n} \\
&= \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}t(x-y)} f(x-y)\, \frac{\mathrm{d}^n x}{(2\pi)^{n/2}}\, \mathrm{e}^{-\mathrm{i}ty} g(y)\, \frac{\mathrm{d}^n y}{(2\pi)^{n/2}} \\
&= \int_{\mathbb{R}^n} \underbrace{\int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}tz} f(z)\, \frac{\mathrm{d}^n z}{(2\pi)^{n/2}}}_{=\widehat{f}(t)}\, \mathrm{e}^{-\mathrm{i}ty} g(y)\, \frac{\mathrm{d}^n y}{(2\pi)^{n/2}} \\
&= \widehat{f}(t) \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}ty} g(y)\, \frac{\mathrm{d}^n y}{(2\pi)^{n/2}} \;=\; \widehat{f}(t) \cdot \widehat{g}(t)\,.
\end{aligned}$$

Dabei folgt die zweite Zeile aus dem Satz von Fubini, und die dritte Zeile ergibt sich mittels der Substitution $z = x - y$. □

Beispiel 4.4. (Rechtecksfunktion)
Gegeben sei die (nicht-periodische) Rechtecksfunktion f durch

$$f(x) \;=\; \begin{cases} 1, & \text{für } x \in [-a, a], \\ 0, & \text{sonst.} \end{cases}$$

Dann bekommen wir

$$\begin{aligned}
\widehat{f}(t) &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}tx} f(x)\, \mathrm{d}x \;=\; \frac{1}{\sqrt{2\pi}} \int_{-a}^{a} \mathrm{e}^{-\mathrm{i}tx}\, \mathrm{d}x \;=\; \frac{1}{\sqrt{2\pi}} \Big[\tfrac{-1}{\mathrm{i}t}\, \mathrm{e}^{-\mathrm{i}tx} \Big]_{-a}^{a} \\
&= \frac{-1}{\mathrm{i}t\sqrt{2\pi}} \left(\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{\mathrm{i}ta} \right) \;=\; \frac{2}{t\sqrt{2\pi}} \sin(ta)
\end{aligned}$$

für die Fourier-Transformierte von f. Insbesondere ist $\widehat{f}(0) = 2a/\sqrt{2\pi}$, wie man entweder direkt oder mit der l'Hospital'schen Regel ausrechnen kann.

Nun wollen wir einige Eigenschaften der Fourier-Transformation festhalten. Natürlich ist sie wieder linear, aber wir bekommen auch folgende Resultate.

Satz 4.5. *Seien $f, g \in L^1(\mathbb{R}^n)$, und sei $\lambda > 0$. Dann gilt:*

1. $g(x) = \overline{f(-x)} \quad \Rightarrow \quad \widehat{g}(t) = \overline{\widehat{f}(t)}$;
2. $g(x) = f(\frac{x}{\lambda}) \quad \Rightarrow \quad \widehat{g}(t) = \lambda \widehat{f}(\lambda t)$.

Weiter gilt, falls g durch $g(x) = -\mathrm{i}\, x_k f(x)$ definiert ist und ebenfalls eine L^1-Funktion ist, dass $\widehat{f}$ partiell nach x_k differenzierbar ist mit der Ableitung

$$\frac{\partial}{\partial t_k} \widehat{f}(t) \;=\; \widehat{g}(t)$$

Beweis. Die ersten beiden Behauptungen ergeben sich analog zu früheren Aussagen durch direktes Nachrechnen (Übung !).

Mit $tx = t_1x_1 + \cdots + t_nx_n$ bekommen wir für die partielle Ableitung

$$\begin{aligned}\frac{\partial}{\partial t_k}\frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}tx} f(x)\,\mathrm{d}^n x &= \frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n} \underbrace{\Big(\frac{\partial}{\partial t_k}\mathrm{e}^{-\mathrm{i}tx}\Big)}_{=-i\,x_k\,\mathrm{e}^{-\mathrm{i}\,tx}} f(x)\,\mathrm{d}^n x \\ &= \frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n} \underbrace{(-\mathrm{i}\,x_k) f(x)}_{=g(x)}\,\mathrm{e}^{-\mathrm{i}tx}\,\mathrm{d}^n x \;=\; \widehat{g}(t)\,,\end{aligned}$$

woraus sich die letzte Behauptung ableiten lässt (man beachte noch die Bedingung an die Funktion g). □

Als wichtiges Beispiel betrachten wir jetzt etwas ausführlicher die durch

$$f(x) \;=\; \mathrm{e}^{-x^2/2}$$

definierte Funktion $f \in L^1(\mathbb{R})$. Dann ist die Fourier-Transformierte $\widehat{f}$ durch

$$\widehat{f}(t) \;=\; \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}t\xi} f(\xi)\,\mathrm{d}\xi \;=\; \mathrm{e}^{-t^2/2}$$

gegeben. Die Funktion f ist also eine Eigenfunktion zur Fourier-Transformierten zum Eigenwert 1, also ist $\widehat{f} = f$. Dies beweist man in der Regel mit Methoden der komplexen Analysis. Da uns das nicht zur Verfügung steht, setzen wir zum Nachweis den Satz von Picard–Lindelöf ein (siehe auch Kapitel 6.2).

Wir wollen also zeigen, dass f und $\widehat{f}$ dasselbe Anfangswertproblem lösen, nämlich

$$\frac{\mathrm{d}}{\mathrm{d}x}\,y(x) \;=\; -x\,y(x) \quad \text{mit} \quad y(0) \;=\; 1\,. \tag{26}$$

Klarerweise wird dies von f erfüllt, wie man direkt nachrechnen kann. Für $\widehat{f}$ bekommen wir mittels einer analogen Rechnung zum letzten Schritt im Beweis von Satz 4.5 und dem Einsatz von partieller Integration:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\,\widehat{f}\,(x) \;&=\; \frac{\mathrm{d}}{\mathrm{d}x}\,\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}x\xi} f(\xi)\,\mathrm{d}\xi \;=\; \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \Big(\frac{\partial}{\partial x}\,\mathrm{e}^{-\mathrm{i}x\xi}\Big) f(\xi)\,\mathrm{d}\xi \\ &=\; \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \big(-\mathrm{i}\,\xi\,\mathrm{e}^{-\mathrm{i}x\xi}\big)\, f(\xi)\,\mathrm{d}\xi \;=\; \frac{\mathrm{i}}{\sqrt{2\pi}}\int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}x\xi}\big(-\xi\, f(\xi)\big)\,\mathrm{d}\xi \\ &=\; \frac{\mathrm{i}}{\sqrt{2\pi}}\int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}x\xi}\Big(\frac{\mathrm{d}}{\mathrm{d}\xi}\,f(\xi)\Big)\,\mathrm{d}\xi \\ &=\; \frac{\mathrm{i}}{\sqrt{2\pi}}\big[\mathrm{e}^{-\mathrm{i}x\xi} f(\xi)\big]_{-\infty}^{\infty} \;-\; \frac{\mathrm{i}}{\sqrt{2\pi}}\int_{\mathbb{R}} \big(-\mathrm{i}x\,\mathrm{e}^{-\mathrm{i}x\xi}\big)\, f(\xi)\,\mathrm{d}\xi \\ &=\; 0 \;-\; \frac{x}{\sqrt{2\pi}}\int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}x\xi} f(\xi)\,\mathrm{d}\xi \;=\; -x\,\widehat{f}\,(x)\,,\end{aligned}$$

und damit erfüllt auch $\widehat{f}$ die DGL aus Gl. (26). Unsere Behauptung stimmt also, wenn wir noch die Anfangsbedingung für $\widehat{f}$ verifizieren können.

Dazu bemerken wir, dass

$$\widehat{f}(0) \;=\; \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \;=\; \frac{1}{\sqrt{\pi}} \int_{\mathbb{R}} \mathrm{e}^{-y^2}\,\mathrm{d}y,$$

mittels der Substitution $x = \sqrt{2}y$ im zweiten Schritt. Unsere Behauptung folgt nun aus folgendem Resultat.

Lemma 4.6. *Es gilt:*

$$I \;=\; \int_{-\infty}^{\infty} \mathrm{e}^{-y^2} dy \;=\; \sqrt{\pi}$$

Beweis. Da es keine einfache Stammfunktion von f gibt, behelfen wir uns mit einem weiteren Trick, indem wir das Quadrat von I berechnen:

$$\begin{aligned} I^2 \quad &= \quad \int_{-\infty}^{\infty} \mathrm{e}^{-x^2}\,\mathrm{d}x \int_{-\infty}^{\infty} \mathrm{e}^{-y^2}\,\mathrm{d}y \;=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-(x^2+y^2)}\,\mathrm{d}x\,\mathrm{d}y \\ &= \quad \int_0^{\infty}\int_0^{2\pi} \mathrm{e}^{-r^2} r\,\mathrm{d}\varphi\,\mathrm{d}r \;=\; 2\pi \int_0^{\infty} r\,\mathrm{e}^{-r^2}\,\mathrm{d}r \\ &\overset{s=r^2}{=} \quad \pi \int_0^{\infty} \mathrm{e}^{-s}\,\mathrm{d}s \;=\; \pi \end{aligned}$$

Dabei entsteht die zweite Zeile durch eine Transformation auf *Polarkoordinaten*, also durch die Substitution $x = r\cos(\varphi)$, $y = r\sin(\varphi)$. Dies führt auf

$$\mathrm{d}x\,\mathrm{d}y \;=\; \begin{vmatrix} \cos(\varphi) & -r\sin(\varphi) \\ \sin(\varphi) & r\cos(\varphi) \end{vmatrix} \mathrm{d}r\,\mathrm{d}\varphi \;=\; r\,\mathrm{d}r\,\mathrm{d}\varphi.$$

Da klarerweise $I > 0$ gelten muss, folgt aus $I^2 = \pi$ nun $I = \sqrt{\pi}$ wie behauptet. □

Nun wollen wir der Frage nachgehen, ob bzw. wann die Ausgangsfunktion f aus der Fourier-Transformierten $\widehat{f}$ berechenbar ist. Gesucht ist also eine inverse Abbildung zur Fourier-Transformation. Dabei wird zunächst ein formaler und nicht ganz korrekter Ansatz verfolgt. Falls auch $\widehat{f} \in L^1(\mathbb{R}^n)$, definieren wir eine Funktion g durch

$$g(x) \;=\; \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}tx}\,\widehat{f}(t)\,\mathrm{d}^n t.$$

Dies geschieht in Analogie zu unseren früheren Formeln. Dann sollte gelten (heuristisch):

$$\begin{aligned} g(x) \quad &= \quad \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}t(x-y)} f(y)\,\mathrm{d}^n y\,\mathrm{d}^n t \\ &\overset{(?)}{=} \quad \int_{\mathbb{R}^n} \underbrace{\left(\frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}t(x-y)}\,\mathrm{d}^n t\right)}_{=\delta(x-y)} f(y)\,\mathrm{d}^n y \;=\; \big(\delta * f\big)(x) \;\overset{(?)}{=}\; f(x), \end{aligned}$$

wobei $*$ hier für die Faltung *ohne* den in Satz 4.3 verwendeten Vorfaktor steht. Dabei sollte δ eine Funktion sein, die bzgl. der Faltung das neutrale Element ist. Das erste (?) deutet an, dass es so eine Funktion in $L^1(\mathbb{R}^n)$ nicht gibt, und wir also nicht wissen, ob und wie man die Rechnung noch 'retten' kann, um das zweite (?) zu rechtfertigen.

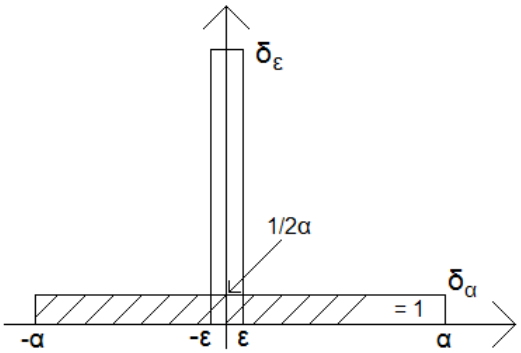

Abbildung 11: Illustration der Familie δ_ϵ mit Parameter $\epsilon > 0$, als Beispiel fuer eine approximierende Eins.

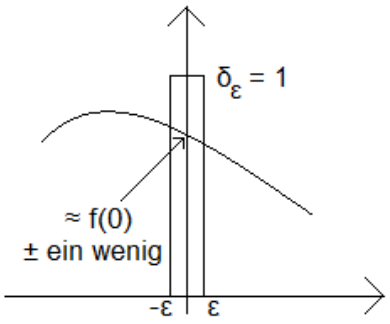

Abbildung 12: Approximation des Funktionswerts $f(0)$ durch das Integral.

Wir suchen also eine (symmetrische) 'Funktion' δ mit der Eigenschaft

$$\int_{\mathbb{R}} \delta(x)\, f(x)\, \mathrm{d}x \;=\; f(0)\,, \quad \text{für jede in 0 stetige Funktion } f.$$

Eine solche Funktion gibt es aber nicht. Man kann aber eine Familie von Funktionen finden, die von einem Parameter abhängt, die dann in einer kleinen ϵ-Umgebung von 0 den Funktionswert $f(0)$ immer besser 'herausholt'. Sei dazu $\delta_\epsilon = \frac{1}{2\epsilon} 1_{[-\epsilon,\epsilon]}$, vgl. Abb. 11 für eine Illustration. Dafür gelten folgende Eigenschaften ($\epsilon > 0$):

$$\begin{aligned} \int_{\mathbb{R}} \delta_\epsilon(x) dx &= 1\,, \\ \int_{\mathbb{R}} \delta_\epsilon(x)\, f(x)\, \mathrm{d}x &= \int_{-\epsilon}^{\epsilon} \frac{f(x)}{2\epsilon}\, \mathrm{d}x\,, \\ f \text{ stetig bei } 0 \;&\Rightarrow\; \lim_{\epsilon \searrow 0} \int_{\mathbb{R}} \delta_\epsilon(x)\, f(x)\, \mathrm{d}x \;=\; f(0)\,. \end{aligned}$$

Man beachte, dass man bei der letzten Eigenschaft Integral und Limes *nicht* vertauschen kann! Die Abbildungen 11 und 12 illustrieren die Eigenschaften. Man erhält durch sauberen Umgang mit solchen Familien $(\delta_\epsilon)_{\epsilon>0}$ (auch *approximierende Eins* genannt) folgenden Satz (hier ohne Beweis):

Satz 4.7. *Sei $f \in L^1(\mathbb{R}^n)$ mit der Eigenschaft dass auch $\widehat{f} \in L^1(\mathbb{R}^n)$ gilt, und sei g durch*

$$g(x) \;=\; \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}tx}\, \widehat{f}\,(t)\, \mathrm{d}^n t$$

definiert. Dann ist g stetig und es gilt $g(x) = f(x)$ fast überall. Also ist die Funktion g in L^1-Sinne die inverse Fourier-Transformierte von $\widehat{f}$.

Zu beachten ist hier, dass $\widehat{f}$ integrierbar eine starke Einschränkung ist, die *nicht* für jedes $f \in L^1(\mathbb{R}^n)$ erfüllt ist.

Folgerung 4.8. *(Eindeutigkeitssatz)*
Ist $f \in L^1(\mathbb{R}^n)$ und $\widehat{f} = 0$, also $\widehat{f}(t) = 0$ für fast alle $t \in \mathbb{R}^n$, dann ist auch $f = 0$, also $f(x) = 0$ fast überall.

Beweis. Die Funktion $\widehat{f}$ ist nach Voraussetzung integrierbar. Wir können also die Funktion g aus Satz 4.7 definieren, und bekommen $g(x) = 0$ für alle (!) x. Dann folgt $f = g$ in $L^1(\mathbb{R}^n)$, das heisst $f(x) = g(x) = 0$ fast überall. □

Man kann die Fourier-Transformation in verschiedenen Räumen betrachten, jeweils mit anderen Gegebenheiten. Weit verbreitet sind:

- Schwartz-Raum: $\mathcal{S}(\mathbb{R}^n)$ (Raum der unendlich schönen Funktionen)
  Für $n = 1$: C^∞, $f^{(n)}(x)P(x) \xrightarrow{x\pm\infty} 0$ für jedes $n \geq 0$ und jedes Polynom P (z.B. $f(x) = \mathrm{e}^{-x^2/2}$). Diese Funktionen sind also beliebig oft differenzierbar und verschwinden für sehr große bzw. kleine x-Werte 'sehr schnell'.

- Temperierte Distributionen: $\mathcal{S}'(\mathbb{R}^n)$
  $T\colon \mathcal{S}(\mathbb{R}^n) \longrightarrow \mathbb{C}$ linear und stetig (z.B. $\delta_x(\varphi) = \varphi(x)$). Die Fourier-Transformation ist dann mittels $\widehat{T}(\varphi) = T(\widehat{\varphi})$ definiert.

- Hilbert-Raum: $\mathcal{H} = L^2(\mathbb{R}^n) = \left\{ f\colon \mathbb{R}^n \longrightarrow \mathbb{C} \,\middle|\, \int_{\mathbb{R}^n} |f(x)|^2 \,\mathrm{d}^n x < \infty \right\}$
  Hier wird vieles wieder systematisch und besonders schön.

4.3 Fourier-Tranformation auf $L^2(\mathbb{R})$

Wir wollen jetzt die Fourier-Transformation auf dem Hilbert-Raum $\mathcal{H} = L^2(\mathbb{R})$ genauer studieren. Wir beschränken uns auf den eindimensionalen Fall, die Erweiterung auf den $\mathbb{R}^n$ kann dann analog behandelt werden. Wir wählen in $\mathcal{H}$ das Skalarprodukt

$$\langle f \,|\, g \rangle \;=\; \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \overline{f(x)}\, g(x)\, \mathrm{d}x$$

und die Norm entsprechend als $\|f\|_2 = \sqrt{\langle f \,|\, f \rangle}$. Als Hilbert-Raum ist $\mathcal{H}$ vollständig, also konvergieren alle Cauchy-Folgen in $L^2(\mathbb{R})$.

Der Raum $\mathcal{H}$ ist natürlich ∞-dimensional, und die Einheitskugel des Raums ist *nicht* kompakt. Zum Beispiel sieht man, dass $x \mapsto x^n \mathrm{e}^{-x^2/2}$ für jedes $n \in \mathbb{N}_0$ eine Funktion in $\mathcal{H}$ definiert. Diese sind aber linear unabhängig, denn für jede endliche Folge (α_n) komplexer Zahlen gilt:

$$\sum_{n=0}^{N} \alpha_n\, x^n\, \mathrm{e}^{-x^2/2} \;=\; 0 \quad \Leftrightarrow \quad \text{alle } \alpha_n = 0\,.$$

Für Funktionen $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ wissen wir bereits, wie wir $\widehat{f}$ zu definieren haben. Zum Glück liegt dieser Schnitt in $L^2(\mathbb{R})$ dicht, so dass wir eine Erweiterung auf den Hilbert-Raum durch eine eindeutige Fortsetzung bekommen.

Satz 4.9. *(Satz von Plancherel)*
Es gibt eine eindeutige Abbildung $\mathcal{F}\colon L^2(\mathbb{R}) \to L^2(\mathbb{R})$ mit folgenden Eigenschaften:

1. *$f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \;\Rightarrow\; \mathcal{F}(f) = \widehat{f}$;*
2. *$\mathcal{F}$ ist eine Isometrie, d.h. $\|\mathcal{F}(f)\|_2 = \|f\|_2$ (bzw. $\langle \mathcal{F}f | \mathcal{F}f\rangle = \langle f|f\rangle$);*
3. *$\mathcal{F}$ ist ein Hilbertraum-Isomorphismus, d.h. auf $\mathcal{H}$ existiert eine eindeutige Inverse; insbesondere ist $\mathcal{F}$ unitär;*
4. *Mit $\varphi_A(t) := \frac{1}{\sqrt{2\pi}} \int_{-A}^{A} \mathrm{e}^{-\mathrm{i}tx} f(x)\,\mathrm{d}x$ und $\psi_A(x) := \frac{1}{\sqrt{2\pi}} \int_{-A}^{A} \mathrm{e}^{\mathrm{i}tx} \widehat{f}(t)\,\mathrm{d}t$ gilt: $\big\|\varphi_A - \widehat{f}\big\|_2 \xrightarrow{A\to\infty} 0$ und $\big\|\psi_A - f\big\|_2 \xrightarrow{A\to\infty} 0$.*

Für einen Beweis dieses Satzes verweisen wir auf [7, Thm. 9.13].

An dieser Stelle wollen wir die beiden unter Punkt 2. genannten Bedingungen für die Isometrieeigenschaft etwas genauer betrachten.

Lemma 4.10. *Sei $\mathcal{L}$ ein linearer Operator auf dem komplexen Hilbert-Raum $\mathcal{H}$, wie z.B. $\mathcal{L} = \mathcal{F}$ auf $\mathcal{H} = L^2(\mathbb{R})$. Dann sind folgende Aussagen äquivalent:*

1. *$\|\mathcal{L}(f)\|_2 = \|f\|_2$ (Erhaltung der Norm) gilt für alle $f \in \mathcal{H}$;*
2. *$\big\langle \mathcal{L}(f) | \mathcal{L}(g)\big\rangle = \langle f\,|\,g\rangle$ (Erhaltung des Skalarprodukts) gilt für alle $f, g \in \mathcal{H}$.*

Beweis. Wegen $\|f\| = \sqrt{\langle f\,|\,f\rangle}$ folgt die erste Aussage einfach aus der zweiten indem wir $g = f$ wählen.

Für die umgekehrte Richtung benötigen wir die sogenannte (komplexe) *Polarisationsgleichung*, also

$$\begin{aligned}
&\|f+g\|^2 - \|f-g\|^2 + \mathrm{i}\,\|f - \mathrm{i}g\|^2 - \mathrm{i}\,\|f+ig\|^2 \\
&\quad= \|f\|^2 + \|g\|^2 + \langle f\,|\,g\rangle + \langle g\,|\,f\rangle - \|f\|^2 - \|g\|^2 + \langle f\,|\,g\rangle + \langle g|f\rangle + \\
&\qquad \mathrm{i}\big(\|f\|^2 - \|g\|^2 - \mathrm{i}\langle f\,|\,g\rangle + \mathrm{i}\langle g|f\rangle - \|f\|^2 + \|g\|^2 + \mathrm{i}\langle f|g\rangle - \mathrm{i}\langle g|f\rangle\big) \\
&\quad= 4\,\langle f\,|\,g\rangle.
\end{aligned}$$

Damit bekommen wir dann

$$\begin{aligned}
\big\langle \mathcal{L}(f)\,|\,\mathcal{L}(g)\big\rangle &= \frac{1}{4}\Big(\big\|\underbrace{\mathcal{L}(f)+\mathcal{L}(g)}_{=\mathcal{L}(f+g)}\big\|^2 - \ldots + \ldots - \mathrm{i}\,\big\|\underbrace{\mathcal{L}(f)+\mathrm{i}\,\mathcal{L}(g)}_{=\mathcal{L}(f+\mathrm{i}g)}\big\|^2\Big) \\
&= \frac{1}{4}\big(\|f+g\|^2 - \ldots + \ldots - \|f+\mathrm{i}g\|^2\big) \;=\; \langle f\,|\,g\rangle,
\end{aligned}$$

und das Lemma ist bewiesen. □

Bemerkung 4.11. In einem euklidischen Vektorraum V über $\mathbb{R}$ mit $\|x\| = \sqrt{\langle x|x\rangle}$ gilt $\langle x|y\rangle = \langle y|x\rangle$ und daher folgt hier einfach

$$\|x+y\|^2 - \|x-y\|^2 \;=\; \|x\|^2 + \|y\|^2 + \langle x|y\rangle + \underbrace{\langle y|x\rangle}_{=\langle x|y\rangle} - \|x\|^2 - \|y\|^2 + \langle x|y\rangle + \underbrace{\langle y|x\rangle}_{=\langle x|y\rangle} \;=\; 4\,\langle x|y\rangle,$$

was man auch die *reelle* Polarisationsgleichung nennt. Für eine allgemeine lineare Abbildung $\mathcal{L}\colon V \longrightarrow V$ sind dann auch in diesem Raum äquivalent:

1. $\|\mathcal{L}(x)\| = \|x\|$ gilt für alle $x \in V$;
2. $\langle \mathcal{L}(x)|\mathcal{L}(y)\rangle = \langle x|y\rangle$ gilt für alle $x, y \in V$.

Normerhaltung und Erhaltung des Skalarprodukts in Hilbert-Räumen sind also stets äquivalente Bedingungen.

Nun kommen wir zu einer überraschenden und wichtigen Eigenschaft der Fourier-Transformation $\mathcal{F}$ und ihrer Inversen, sowie einer wichtigen Konsequenz für $L^2(\mathbb{R})$. Sei $\widehat{g} = \mathcal{F}(g)$ mit $\widehat{g} \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. Dann ist

$$\big(\mathcal{F}^2 g\big)(x) \;=\; \big(\mathcal{F}(\mathcal{F}(g)\big)(x) \;=\; \big(\mathcal{F}(\widehat{g})\big)(x)) \;=\; \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}xy}\, \widehat{g}(y)\,\mathrm{d}y \;=\; g(-x)\,,$$

wobei man über den letzten Schritt ein wenig nachdenken sollte, bis er klar wird. Also ist $\big(\mathcal{F}^2(g)\big)(x) = g(-x)$ für alle g aus einer dichten Teilmenge von $L^2(\mathbb{R})$, wofür wir etwa alle Schwartz-Funktionen nehmen können, und nach dem Satz von Plancherel (also Satz 4.9) gilt diese Aussage dann auf ganz $L^2(\mathbb{R})$. Es folgt also

$$\mathcal{F}^4 \;=\; \mathbb{1}$$

auf ganz $L^2(\mathbb{R})$. Die Abbildung $\mathcal{F}$ ist also ein unitärer Operator mit $\mathcal{F}^4 = \mathbb{1}$.

Ist nun λ ein Eigenwert von $\mathcal{F}$, also $\mathcal{F}(g) = \lambda g$ für ein $0 \neq g \in L^2(\mathbb{R})$, so folgt

$$g \;=\; \mathcal{F}^4(g) \;=\; \lambda^4 g \quad \Rightarrow \quad \lambda^4 = 1\,.$$

Alle Eigenwerte von $\mathcal{F}$ müssen also 4. Einheitswurzeln sein, d.h. $\lambda \in \{1, \mathrm{i}, -1, -\mathrm{i}\}$. Damit haben wir folgendes Resultat gezeigt.

Satz 4.12. *Die Fourier-Transformation $\mathcal{F}\colon L^2(\mathbb{R}) \to L^2(\mathbb{R})$ ist eine unitäre Transformation der Ordnung 4 mit den Eigenwerten $1, -\mathrm{i}, -1, \mathrm{i}$.*

Beweisskizze. Wir wissen schon, dass $\mathcal{F}^4 = \mathbb{1}$, und dass keine anderen Eigenwerte vorkommen können. Wir müssen also noch zeigen, dass alle 4. Einheitswurzeln wirklich auftreten.

Die durch $x \mapsto \mathrm{e}^{-x^2/2}$ definierte Funktion ist eine Eigenfunktion zum Eigenwert 1, wie wir ebenfalls schon bewiesen haben. Nun gehen wir konstruktiv vor, indem wir induktiv weitere Eigenfunktionen angeben. Dazu setzen wir rekursiv

$$H_{n+1}(x) \;=\; 2x\,H_n(x) - 2n\,H_{n-1}(x)$$

für $n \in \mathbb{N}$, mit $H_0(x) = 1$ und $H_1(x) = 2x$. Dann ist $H(x)$ ein Polynom n-ter Ordnung, bekannt als n-tes *Hermite-Polynom*. Wir gehen einen Schritt weiter und betrachten die *Hermite-Funktionen*

$$h_n(x) \;=\; \alpha_n \, H_n(x) \, \mathrm{e}^{-x^2/2},$$

mit geeigneten Normierungskonstanten α_n. Nun rechnet man per Induktion nach, dass h_n eine Eigenfunktion von $\mathcal{F}$ zum Eigenwert $(-\mathrm{i})^n$ ist. □

Die Hermite-Funktionen haben noch eine ganze Reihe weiterer Eigenschaften, auf die wir hier nicht näher eingehen können. Sie treten z.B. beim harmonischen Oszillator in der Quantenmechanik auf. Für uns von Bedeutung ist aber das folgende Resultat, das wir ohne Beweis angeben (den man wieder mit einigen Tricks per Induktion führen kann).

Satz 4.13. *Die Hermite-Funktionen $\{h_n \mid n \in \mathbb{N}_0\}$ bilden eine Basis von $L^2(\mathbb{R})$, die aus Eigenfunktionen der Fourier-Transformation besteht. Dabei gilt*

$$\mathcal{F}(h_n) \;=\; (-\mathrm{i})^n \, h_n$$

für $n \in \mathbb{N}_0$. Mit $\alpha_n = (2^n \, n! \sqrt{n})^{-1/2}$ gilt außerdem

$$\int_{\mathbb{R}} \overline{h_m(x)} \cdot h_n(x) \, \mathrm{d}x \;=\; \delta_{m,n}$$

für alle $m, n \in \mathbb{N}_0$. Es handelt sich dann um eine ONB bzgl. dieses Skalarprodukts. □

Hier ist zu beachten, dass wir weiter oben einen Vorfaktor bei der Definition des Skalarproduktes gewählt hatten, der in dieser Formulierung nicht auftritt (man kann das natürlich ineinander umrechnen). Die Basiseigenschaft beinhaltet die früher schon behandelte Vollständigkeitsrelation

$$\sum_{n=0}^{\infty} |h_n\rangle\langle h_n| \;=\; \mathbb{1},$$

die wir jetzt noch elegant einsetzen können, um die Fourier-Transformation einer beliebigen L^2-Funktion zu berechnen.

Folgerung 4.14. *Bezüglich der Basis $\{h_n \mid n \in \mathbb{N}_0\}$, mit der Wahl der Konstanten α_n wie oben, ist $\mathcal{F}$ diagonal, und es gilt $\mathcal{F} = \mathrm{diag}(1, -\mathrm{i}, -1, \mathrm{i}, 1, -\mathrm{i}, -1, \mathrm{i}, \ldots)$, oder, in Dirac-Notation,*

$$\mathcal{F} \;=\; \sum_{n=0}^{\infty} |h_n\rangle (-\mathrm{i})^n \langle h_n|.$$

Entwickeln wir nun eine beliebige Funktion $g \in L^2(\mathbb{R})$ in unsere Basis aus normierten Hermite-Funktionen, also

$$|g\rangle \;=\; \sum_{\ell=0}^{\infty} c_\ell |h_\ell\rangle \quad \textit{mit} \quad c_\ell \;=\; \langle h_\ell | g\rangle$$

(was immer geht wegen der Vollständigkeitsrelation, mit den Koeffizienten gemäß der Gauß'schen Approximationsaufgabe), so bekommen wir

$$\mathcal{F}|g\rangle \;=\; \sum_{n,\ell} |h_n\rangle(-\mathrm{i})^n \langle h_n|c_\ell|h_\ell\rangle \;=\; \sum_{n,\ell} |h_n\rangle(-\mathrm{i})^n c_\ell \underbrace{\langle h_n|h_\ell\rangle}_{=\delta_{n,\ell}} \;=\; \sum_n c_n\,(-\mathrm{i})^n\,|h_n\rangle\,.$$

Hieraus ergibt sich auch noch eine zuweilen sehr nützliche Zerlegung des $L^2(\mathbb{R})$, nämlich

$$L^2(\mathbb{R}) \;=\; \bigoplus_{\ell=0}^{3} \big\langle h_{4m+\ell} \mid m \in \mathbb{N}_0 \big\rangle_{\mathbb{C}}\,.$$

Insbesondere lässt sich jedes $g \in L^2(\mathbb{R})$ eindeutig als Summe von 4 Termen schreiben,

$$g \;=\; g_0 + g_1 + g_2 + g_3 \quad \text{mit} \quad \mathcal{F}(g_\ell) \;=\; (-\mathrm{i})^\ell\, g_\ell\,.$$

Dies kann man als eine Verallgemeinerung der Zerlegung einer Funktion f in gerade und ungerade Anteile ansehen, also

$$f(x) \;=\; \underbrace{\frac{f(x)+f(-x)}{2}}_{=f_+(x)} + \underbrace{\frac{f(x)-f(-x)}{2}}_{=f_-(x)}$$

mit $f_\pm(-x) = \pm f_\pm(x)$.

5 Diskrete Fourier-Transformation

Ziel der diskreten Fourier-Transformation ist es nun, eine numerische Berechnung einer Fourier-Transformation oder auch vergleichbarer Integrale zu ermöglichen. Die Idee ist vergleichbar mit dem Riemann'schen Integralbegriff. Betrachte dazu Abbildung 13. Dabei wird die Funktion f in kleinere Abschnitte aufgeteilt, um diskret arbeiten zu können.

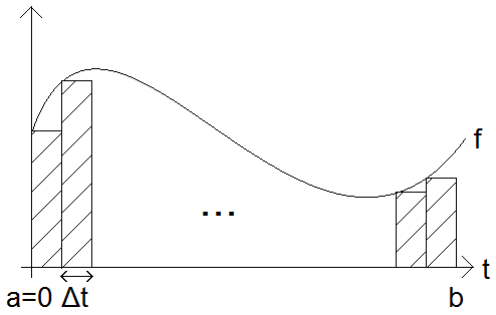

Abbildung 13: Idee der numerischen Berechnung der Fourier-Transformation einer Funktion. Die Idee folgt dabei dem Integralbegriff von Riemann.

Nun lässt sich die Funktion

$$\Phi(\omega) \;=\; \int_0^b \mathrm{e}^{-\mathrm{i}\omega t} f(t)\,\mathrm{d}t$$

durch die Summe der diskreten Berechnungen der einzelnen Abschnitte approximieren:

$$\Phi(\omega) \;\approx\; \sum_{k=0}^{N-1} f(t_k)\,\mathrm{e}^{-\mathrm{i}\omega t_k}\,\triangle t \quad \text{mit} \quad t_k = k\,\frac{b}{N} = k\,\triangle t\,.$$

Bei günstigen Voraussetzungen seitens der Funktion f wird der Fehler der Approximation klein, wenn $\triangle t \searrow 0$. Betrachtet man weiter nur Frequenzen $\omega_n = \frac{2\pi n}{b}$ für $n \in \mathbb{N}$, so gilt

$$\Phi(\omega_n) \;=\; \triangle t \underbrace{\sum_{k=0}^{N-1} f(t_k)\,\mathrm{e}^{-\mathrm{i}\,\frac{2\pi n k}{N}}}_{=d_n}\,. \tag{27}$$

Mit Hilfe der Approximation aus Gl. (27) für einzelne Frequenzen lässt sich nun die diskrete Fourier-Transformation folgendermaßen beschreiben. Für einen gegebenen Datensatz $\{f(t_0), f(t_1), ..., f(t_{N-1})\}$ erhält man den neuen Datensatz $\{d_0, d_1, ..., d_{N-1}\}$ durch eine lineare und invertierbare Transformation, welche später durch die sogenannte schnelle Fourier-Transformation effizient berechnet werden kann.

Dazu betrachten wir f nun als eine Funktion $f\colon C_N \longrightarrow \mathbb{C}$ mit der zyklischen Gruppe C_N. Letztere schreiben wir als $C_N = \{0, 1, \ldots, N-1\}$ mit Addition modulo N. Parallel betrachten wir f auch als N-Tupel komplexer Zahlen, also $f = \{f_0, f_1, ..., f_{N-1}\}$. Dabei

stellen wir uns vor, dass sich der Ursprung der Werte aus $f_i = f(t_i)$ mit $0 \le i \le N-1$ ableitet. Die beiden Darstellungsweisen sind äquivalent, aber es ist nützlich, beide parallel einzusetzen.

Auf der endlichen abelschen Gruppe C_N können wir eine 'Intregration' über eine beliebige Teilmenge $A \subseteq C_N$ definieren, nämlich einfach die Summation

$$\int_A f \;:=\; \sum_{i\in A} f(i).$$

Damit ist dann *jede* Funktion auf C_N integrabel, und wir haben auch

$$L^1(C_N) \;=\; L^2(C_N) \;\simeq\; \mathbb{C}^N.$$

Insbesondere können wir hier also immer im Hilbert-Raum arbeiten, wobei wir den $\mathbb{C}^N$ mit dem Standard-Skalarprodukt ausstatten, also

$$\langle f \,|\, g \rangle \;:=\; \sum_{k=0}^{N-1} \overline{f_k}\, g_k.$$

Definition 5.1. (diskrete Fourier-Transformation)
Die diskrete Fourier-Transformation (DFT) Df einer Funktion $f\colon C_N \longrightarrow \mathbb{C}$ ist gegeben durch

$$(Df)(n) \;=\; \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} f_k\, \mathrm{e}^{-2\pi \mathrm{i}\, kn/N}, \qquad \text{für } n \in C_N\,. \tag{28}$$

Die Transformation Df lässt sich auch in eine Matrix-Schreibweise bringen. Diese ist später wichtig für die effiziente Berechnung und wird daher an dieser Stelle eingeführt. Bezeichnet $\omega = \mathrm{e}^{2\pi\mathrm{i}/N}$ eine primitive N-te Einheitzwurzel, und $\overline{\omega} = \mathrm{e}^{-2\pi\mathrm{i}/N}$ ihre Inverse (bzgl. der Multiplikation), so definieren wir noch die Basis-Funktionen $\omega_j\colon C_N \longrightarrow \mathbb{C}$ durch $m \mapsto \omega_j(m) = \overline{\omega}^{jm}$ für $j, m \in C_N$.

Die Matrix-Schreibweise der DFT kann folgendermaßen notiert werden:

$$Df \;=\; \begin{pmatrix} (Df)(0) \\ \vdots \\ (Df)(N-1) \end{pmatrix} \;=\; M_N \begin{pmatrix} f_0 \\ \vdots \\ f_{N-1} \end{pmatrix}$$

mit

$$M_N \;:=\; \frac{1}{\sqrt{N}} \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \overline{\omega} & \overline{\omega}^2 & \cdots & \overline{\omega}^{N-1} \\ 1 & \overline{\omega}^2 & \overline{\omega}^4 & \cdots & \overline{\omega}^{2(N-1)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & \overline{\omega}^{N-1} & \overline{\omega}^{2(N-1)} & \cdots & \overline{\omega}^{(N-1)^2} \end{pmatrix}.$$

Lemma 5.2. *Sei $N \ge 2$ und $\xi = \mathrm{e}^{2\pi\mathrm{i}k/N}$ mit $1 \le k \le N-1$. Dann ist*

$$1 + \xi + \xi^2 + \ldots + \xi^{N-1} \;=\; 0.$$

Beweis. Unter den gewählten Voraussetzungen ist $\xi \neq 1$, und wir können die Summenformel fuer die endliche geometrische Reihe einsetzen, also

$$1 + \xi + \xi^2 + ... + \xi^{N-1} \;=\; \frac{1-\xi^N}{1-\xi} \;=\; \frac{1-\mathrm{e}^{2\pi \mathrm{i} k}}{1-\xi} \;=\; 0,$$

wobei wir im letzten Schritt benutzt haben, dass $\mathrm{e}^{2\pi \mathrm{i} k} = 1$ gilt. □

Auch wenn der $L^2(C_N) \simeq \mathbb{C}^N$ der DFT deutlich einfacher ist als unsere früheren Räume, ist es dennoch sinnvoll, eine ONB zu kennen. Diese ist durch die bisherige Vorarbeit leicht zu bestimmen:

Lemma 5.3. *Sei $N \in \mathbb{N}$ fest gewählt. Für die Funktionen ω_j mit $j \in C_N$ gilt dann*

$$\left\langle \frac{\omega_j}{\sqrt{N}} \middle| \frac{\omega_k}{\sqrt{N}} \right\rangle \;=\; \delta_{j,k}$$

für beliebige $j, k \in C_N$. Insbesondere ist dann $\left\{ \frac{\omega_j}{\sqrt{N}} \middle| j \in C_N \right\}$ eine ONB von $L^2(C_N)$.

Beweis. Dies können wir einfach nachrechnen:

$$\begin{aligned}
\left\langle \frac{\omega_j}{\sqrt{N}} \middle| \frac{\omega_k}{\sqrt{N}} \right\rangle \;&=\; \frac{1}{N} \sum_{\ell=0}^{N-1} \overline{\omega_j(\ell)}\, \omega_k(\ell) \;=\; \frac{1}{N} \sum_{\ell=0}^{N-1} \overline{\overline{\omega}^{j\ell}}\, \overline{\omega}^{k\ell} \\
&=\; \frac{1}{N} \sum_{\ell=0}^{N-1} \overline{\omega}^{(k-j)\ell} \;=\; \frac{1}{N} \sum_{\ell=0}^{N-1} \omega^{(j-k)\ell} \;=\; \begin{cases} 0, & \text{falls } j \neq k, \\ 1, & \text{falls } j = k, \end{cases}
\end{aligned}$$

wobei wir im letzten Schritt das Ergebnis von Lemma 5.2 eingesetzt haben. Die Basiseigenschaft ist nun klar, da wir N linear unabhängige Funktionen in $L^2(C_N)$ haben, der ein komplexer Vektorraum die Dimension N hat. □

Beispiel 5.4. Ein kurzes Beispiel soll die Idee der Matrix-Form verdeutlichen. Dabei ist die Matrix M_N für $N = 2$ wegen $\mathrm{e}^{\frac{-2\pi \mathrm{i}}{2}} = -1$ gerade gegeben durch

$$M_2 \;=\; \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Damit folgt für zwei beliebige Daten $f(0)$ und $f(1)$ die DFT

$$\begin{pmatrix} (Df)(0) \\ (Df)(1) \end{pmatrix} \;=\; \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} f(0) \\ f(1) \end{pmatrix} \;=\; \frac{1}{\sqrt{2}} \begin{pmatrix} f(0)+f(1) \\ f(0)-f(1) \end{pmatrix}.$$

Lemma 5.5. *Die Matrix M_N ist unitär, für jedes $N \in \mathbb{N}$.*

Beweis. Da M_N eine symmetrische Matrix ist, folgt Unitarität, wenn wir $M_N\, \overline{M_N} = \mathbb{1}$ nachrechnen. Wegen

$$M_N\, \overline{M_N} \;=\; \frac{1}{N} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & \overline{\omega} & \cdots & \overline{\omega}^{N-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & \overline{\omega}^{N-1} & \cdots & \overline{\omega}^{(N-1)^2} \end{pmatrix} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & \omega & \cdots & \omega^{N-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & \omega^{N-1} & \cdots & \omega^{(N-1)^2} \end{pmatrix}$$

bekommen wir offenbar für das Matrixelement auf Position (j,k) den Ausdruck

$$\frac{1}{N}\sum_{n=0}^{N-1}\overline{\omega}^{jn}\,\omega^{kn} \;=\; \frac{1}{N}\sum_{n=0}^{N-1}\mathrm{e}^{2\pi\mathrm{i}(k-j)n/N} \;=\; \delta_{jk}\,,$$

woraus unsere Behauptung folgt. □

Nun kommen wir zur Umkehrung der DFT.

Satz 5.6. *(Inverse der DFT)*
Sei $N\in\mathbb{N}$, $f\colon C_N\longrightarrow\mathbb{C}$ und $\omega=\mathrm{e}^{2\pi\mathrm{i}/N}$. Mit der diskreten Fourier-Transformation Df gemäß Gl. (28) *gilt*

$$f(n) \;=\; \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}(Df)(k)\,\underbrace{\mathrm{e}^{2\pi\mathrm{i}\,kn/N}}_{=\omega^{kn}}$$

für alle $n\in C_N$.

Beweis. Das rechnen wir wieder nach:

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}(Df)(k)\,\omega^{kn} \;&=\; \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}\frac{1}{\sqrt{N}}\sum_{j=0}^{N-1}f(j)\,\overline{\omega}^{jk}\,\omega^{kn} \;=\; \frac{1}{N}\sum_{k=0}^{N-1}\sum_{j=0}^{N-1}f(j)\,\omega^{(n-j)k}\\ &=\; \frac{1}{N}\sum_{j=0}^{N-1}f(j)\;\underbrace{\sum_{k=0}^{N-1}\omega^{(n-j)k}}_{=N\delta_{nj}\quad\text{(vgl. Lemma 5.3)}} \;=\; f(n)\,,\end{aligned}$$

was für beliebiges $n\in C_N$ stimmt. □

Satz 5.6 kann man ebenfalls (und letztlich eleganter) mit der Matrix-Form der DFT herleiten. Dafür muss nämlich lediglich die Gleichung $M_N\,f=Df$ mit $\overline{M}_N$ multipliziert werden, und es ergibt sich

$$\underbrace{\overline{M}_N M_N}_{=\mathbb{1}} f \;=\; \overline{M}_N(Df)\,,$$

was im Grunde eine Anwendung der Unitarität von M_N ist.

Bemerkung 5.7. Wir wissen aus Lemma 5.5, dass die DFT eine unitäre Abbildung ist. Dies ist für $M_1=1$ natürlich trivial. Man kann nun (für $N\geq 2$) nachrechnen, dass

$$(M_N)^2 \;=\; \begin{pmatrix}1&0&\cdots&0&0\\0&0&\cdots&0&1\\0&0&\cdots&1&0\\\vdots&\vdots&\iddots&\vdots&\vdots\\0&1&\cdots&0&0\end{pmatrix}$$

N	$\lambda = 1$	$\lambda = \mathrm{i}$	$\lambda = -1$	$\lambda = -\mathrm{i}$
$4m$	$m+1$	$m-1$	m	m
$4m+1$	$m+1$	m	m	m
$4m+2$	$m+1$	m	$m+1$	m
$4m+3$	$m+1$	m	$m+1$	$m+1$

Tabelle 1: Die Eigenwerte von M_N und ihre Vielfachheiten.

gilt. Die Rechnung ist analog zu vielen der obigen Rechnungen (Übung). Daraus folgt nun $(M_N)^4 = \mathbb{1}$, wobei man auch leicht nachprüft, dass 4 für $N \geq 3$ die minimale (positive) Potenz mit dieser Eigenschaft ist, während für $N = 2$ bereits $(M_2)^2 = \mathbb{1}$ gilt.

Die möglichen Eigenwerte der DFT sind ± 1 und $\pm \mathrm{i}$, wobei ab $N = 5$ alle vorkommen. Diese waren, zusammen mit ihren Vielfachheiten, bereits Gauß bekannt. Sie weisen eine Struktur modulo 4 auf und sind in Tabelle 1 zusammengefasst.

Allerdings sind die Eigenvektoren der DFT nicht in geschlossener Form bekannt.

Auch im diskreten Bereich spielt die Faltung wieder eine wichtige Rolle.

Definition 5.8. (diskrete Faltung)
Die diskrete Faltung $f * g$ von $f, g \in L^1(C_N)$ ist für jedes $k \in C_N$ definiert durch

$$(f * g)(k) \;=\; \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} f(j)\, g(k-j)\,,$$

wobei die Argumente wieder modulo N berechnet werden.

Das folgende Resultat überrascht uns dann nicht mehr:

Satz 5.9. *(Faltungssatz)*
Für beliebige $f, g \in L^1(C_N)$ gilt $D(f * g) = (Df) \cdot (Dg)$.

Beweis. Wir rechnen das für ein beliebiges $n \in C_N$ nach:

$$\begin{aligned}
D(f*g)(n) &= \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} (f*g)(k)\, \overline{\omega}^{kn} \\
&= \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} f(j)\, g(k-j)\, \overline{\omega}^{kn} \\
&= \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} f(j)\, \overline{\omega}^{jn} \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} g(k-j)\, \overline{\omega}^{(k-j)n} \\
&= \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} f(j)\, \overline{\omega}^{jn} \frac{1}{\sqrt{N}} \sum_{\ell=0}^{N-1} g(\ell)\, \overline{\omega}^{\ell n} \\
&= (Df)(n) \cdot (Dg)(n)\,,
\end{aligned}$$

wobei der vorletzte Schritt durch eine Umsummierung entsteht, die wir durchfüren können, weil die Argumente ja Elemente der Gruppe (!) C_N sind. □

Für die weiteren Abschnitte führen wir zunächst noch zwei neue Bezeichnungen ein, die auch in der Literatur weit verbreitet sind, nämlich

$$d_k \;=\; \mathcal{F}_k(f) \;=\; (Df)(k) \;=\; \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} \mathrm{e}^{-2\pi \mathrm{i}\, \frac{kj}{N}} f_j \,.$$

Dies bedeutet nur, dass wir die parallele Verwendung von Abbildungen und Vektoren jetzt 'ernst' nehmen.

5.1 Trigonometrische Interpolationsapproximation

Bislang ging es bei der DFT in erster Linie um die Approximation der Fourier-Transformation von Funktionen bzw. um die Transformation eines endlichen Datensatzes. Jetzt wollen wir uns ein wenig mit der Frage befassen, wie wir die obigen Resultate für die Approximation einer Funktion f auf einem Interval $[0, L]$ mittels Interpolation einsetzen können. Dabei ist ein trigonometrisches Polynom p auf $[0, L]$ gesucht, so dass für eine gegebene Menge an Punkten $\{(x_0, f_0), ..., (x_{N-1}, f_{N-1})\}$ mit (festem) $N \in \mathbb{N}$ die Gleichung $p(x_j) = f_j$ für alle $0 \le j \le N-1$ gilt. Außerdem soll natürlich p auch sonst so nahe an f verlaufen wie möglich. Wir wollen hier nur den Fall äquidistanter Stützstellen betrachten, also $x_j = \frac{jL}{N}$ für $0 \le j \le N-1$, und außerdem noch annehmen, dass $f(0) = f(L)$ gilt.

Gesucht ist also ein trigonometrisches Polynom der Form

$$p(x) \;=\; \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} d_k \, \mathrm{e}^{2\pi \mathrm{i} kx/L}. \tag{29}$$

Die Aufgabe ist nun zunächst, die Zahlen d_k so zu wählen, dass alle Stützstellenbedingungen erfüllt sind. Danach wird zu sehen sein, ob sich damit im Rest des Intervalls eine befriedigende Approximation der Funktion f ergibt, oder ob wir den Ansatz noch geeignet modifizieren müssen.

Satz 5.10. *Sei $N \in \mathbb{N}$ beliebig, aber fest. Zu äquidistanten Stützstellen $x_j \in [0, L]$, also $x_j = \frac{jL}{N}$ für $0 \le j \le N-1$, und beliebigen Stützwerten $f_j \in \mathbb{C}$ besitzt das trigonometrische Polynom aus Gl.* (29) *die Interpolationseigenschaft $p(x_j) = f_j$ genau dann, wenn gilt:*

$$\begin{pmatrix} d_0 \\ \vdots \\ d_{N-1} \end{pmatrix} \;=\; M_N \begin{pmatrix} f_0 \\ \vdots \\ f_{N-1} \end{pmatrix}.$$

Beweis. Dies ist eine direkte Konsequenz des Inversionssates (Satz 5.6) für die DFT. □

Beispiel 5.11. Gegeben sei die Dreiecksfunktion

$$f(x) = \begin{cases} x, & 0 \le x < \frac{1}{2}, \\ 1-x, & \frac{1}{2} \le x < 1, \\ 0, & \text{sonst.} \end{cases}$$

Mit Hilfe einer Interpolationsapproximation mit $N = 4$ erhält man für p die Situation aus Abbildung 14.

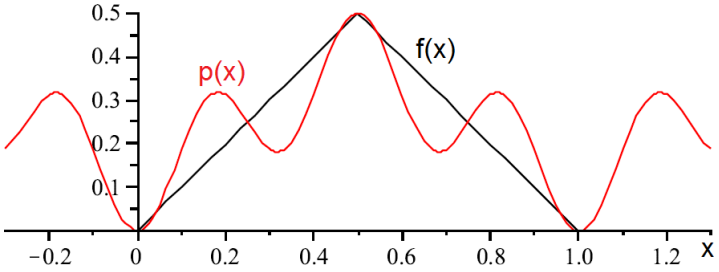

Abbildung 14: Interpolationsapproximation der Dreiecksfunktion durch das Polynom $p(x)$ aus Gl. (29) mit $L = 1$ und $N = 4$.

An diesem Beispiel ist sofort zu erkennen, dass p zwar an den Stützstellen exakt stimmt, die restliche Funktion aber nur sehr schlecht approximiert wird. Wählt man weitere Stützstellen um die Qualität der Approximation zu verbessern, oszilliert die Funktion p allerdings noch stärker. Abbildung 15 zeigt die Interpolationsapproximation der Dreiecksfunktion mit $N = 16$, die klarerweise noch weniger brauchbar ist als die mit $N = 4$. Wir sehen also, dass der erste Versuch mit Gl. (29) nicht wirklich geeignet ist.

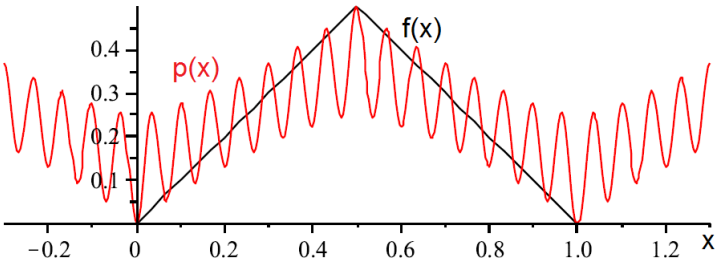

Abbildung 15: Interpolationsapproximation der Dreiecksfunktion mit $N = 16$. Die Approximation oszilliert dabei noch stärker als für $N = 4$.

Beispiel 5.11 zeigt, dass man andere Polynome benötigt. Dazu betrachten wir den

folgenden (in gewisser Weise symmetrisierten) Ansatz für N gerade:

$$\begin{aligned} r(x) &:= \frac{1}{\sqrt{N}} \sum_{k=-N/2}^{N/2-1} d_{k+N/2}\, \mathrm{e}^{2\pi \mathrm{i} kx/L} \\ &= \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} d_k\, \mathrm{e}^{2\pi \mathrm{i}(k-N/2)x/L} \\ &= \underbrace{\frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} d_k\, \mathrm{e}^{2\pi \mathrm{i} kx/L}}_{\text{entspricht } p(x)}\, \mathrm{e}^{-\pi \mathrm{i} Nx/L}. \end{aligned} \tag{30}$$

Zunächst ergeben sich die Parameter wie folgt.

Satz 5.12. *Sei $N \in \mathbb{N}$ beliebig, aber fest. Zu äquidistanten Stützstellen $x_j \in [0, L]$ und beliebigen Stützwerten $f_j \in \mathbb{C}$ für $0 \le j \le N-1$ erfüllt das trigonometrische Polynom r aus Gl. (30) die Interpolationsgleichungen $r(x_j) = f_j$ genau dann, wenn gilt:*

$$\begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ \vdots \\ d_{N-1} \end{pmatrix} = M_N \begin{pmatrix} f_0 \\ -f_1 \\ f_2 \\ \vdots \\ (-1)^{N-1} f_{N-1} \end{pmatrix}.$$

Beweis. Das Resultat folgt wieder aus dem Eindeutigkeitssatz der DFT (Satz 5.6), wenn man noch beachtet, dass $\mathrm{e}^{k\pi \mathrm{i}} = (-1)^k$ für $k \in \mathbb{Z}$ gilt. □

Für das obige Beispiel ergeben sich nun deutlich bessere Approximationen an die Dreiecksfunktion, wie Abbildung 16 illustriert.

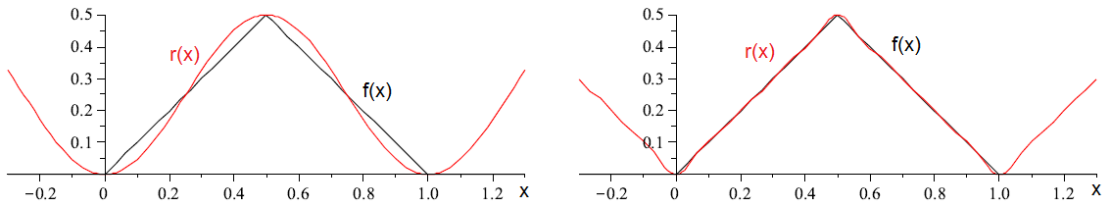

Abbildung 16: Beispiel der Interpolationsapproximation der Dreiecksfunktion durch das verbesserte Polynom r für $N = 4$ (links) und $N = 16$ (rechts). Die Approximation ist deutlich besser als durch das vorherige Polynom p.

Die obige Diskussion soll auch klarmachen, dass man bei Anwendungen der DFT Vorsicht walten lassen muss.

5.2 Schnelle Fourier-Transformation

Dieses Kapitel beschäftigt sich mit der effizienten Berechnung der DFT für einen gegebenen Datensatz $(f_0, f_1, \ldots, f_{N-1})$. Der Algorithmus der schnellen Fourier-Transformation (kurz: FFT, für 'fast Fourier transform') wird auch in vielen Computer-Programmen zur Fourier-Transformation verwendet. Das Problem der DFT, wie sie bisher dargestellt wurde, ist die Matrix-Multiplikation, welche sehr rechenintensiv ist. Der Aufwand liegt dabei in $\mathcal{O}(N^2)$. Mit einem einfachen Trick lässt sich dies besser gestalten.

Für $N = 2M$ und eine Umsortierung der Summanden der k-ten Komponente $\mathcal{F}_k$ erhält man zunächst

$$\begin{aligned}
&\mathcal{F}_k[g_0, g_M, g_1, g_{M+1}, \ldots, g_{M-1}, g_{2M-1}] \\
&\quad = \frac{1}{\sqrt{2M}}\Bigg(\sum_{j=0}^{M-1} g_j \, \mathrm{e}^{-2\pi \mathrm{i}\, k(2j)/2M} + \sum_{j=0}^{M-1} g_{M+j} \, \mathrm{e}^{-2\pi \mathrm{i}\, k(2j+1)/2M}\Bigg) \\
&\quad = \frac{1}{\sqrt{2M}}\Bigg(\sum_{j=0}^{M-1} g_j \, \mathrm{e}^{-2\pi \mathrm{i}\, kj/M} + \mathrm{e}^{-\pi \mathrm{i}\, k/M} \sum_{j=0}^{M-1} g_{M+j} \, \mathrm{e}^{-2\pi \mathrm{i}\, kj/M}\Bigg),
\end{aligned}$$

und man erkennt, dass auf der rechten Seite Ausdrücke stehen, die wir aus der DFT eines Datensatzes der Länge M kennen. Allgemeiner gilt nun folgender Zusammenhang.

Satz 5.13. *Sei $M \in \mathbb{N}$ beliebeig, aber fest und $N = 2M$. Aus der DFT der beiden Datensätze $(g_0, g_1, ..., g_{M-1})$ und $(g_M, g_{M+1}, ..., g_{2M-1})$ der Länge M bekommt man die DFT des Datensatzes $(g_0, g_M, g_1, g_{M+1}, ..., g_{M-1}, g_{2M-1})$ der Länge $2M$ wie folgt:*

$$\begin{aligned}
\mathcal{F}_k[g_0, g_M, ..., g_{2M-1}] &= \tfrac{1}{\sqrt{2}}\big(\mathcal{F}_k[g_0, g_1, ..., g_{M-1}] + \mathrm{e}^{-\pi \mathrm{i}\, k/M} \mathcal{F}_k[g_M, g_{M+1}, ..., g_{2M-1}]\big), \\
\mathcal{F}_{k+M}[g_0, g_M, \ldots, g_{2M-1}] &= \tfrac{1}{\sqrt{2}}\big(\mathcal{F}_k[g_0, g_1, ..., g_{M-1}] - \mathrm{e}^{-\pi \mathrm{i}\, k/M} \mathcal{F}_k[g_M, g_{M+1}, ..., g_{2M-1}]\big),
\end{aligned}$$

jeweils mit $0 \le k < M$.

Beweis. Die erste Relation folgt sofort aus obiger Rechnung. Die zweite Relation rechnet man analog nach, wobei man noch folgende Identität beachtet:

$$\begin{aligned}
\mathrm{e}^{-2\pi \mathrm{i}\,(k+M)\ell/2M} &= \mathrm{e}^{-\pi \mathrm{i}\, \ell} \, \mathrm{e}^{-\pi \mathrm{i}\,(k\ell)/M} = (-1)^\ell \mathrm{e}^{-\pi \mathrm{i}\,(k\ell)/M} \\
&= \begin{cases} \mathrm{e}^{-2\pi \mathrm{i}\,(kj)/M}, & \ell = 2j, \\ -\mathrm{e}^{-\pi \mathrm{i}\, k/M} \, \mathrm{e}^{-2\pi \mathrm{i}\,(k\ell)/M}, & \ell = 2j+1. \end{cases}
\end{aligned}$$

Damit ist die behauptete Reduktion gezeigt. □

Mit Hilfe von Satz 5.13 lässt sich der Aufwand zur Berechnung der DFT nun deutlich reduzieren. Allerdings ist dies mittels Satz 5.13 zunächst nur für Datensätze mit einer bestimmten Länge von 2^n hilfreich. Die Verbesserung ist dafür aber bemerkenswert, denn die Berechnung kann nun in $\mathcal{O}(n \log(n))$ ablaufen. Für andere Primzahlen gibt es analoge Reduktionen, so dass man am Ende für jeden endlichen Datensatz einen entsprechenden FFT-Algorithmus bekommt, indem man die Primfaktorzerlegung der Länge heranzieht.

Bemerkenswert ist hier, dass dieses Verfahren – in voller Allgemeinheit – bereits Gauß bekannt war. Es findet sich in seinen Notizbüchern, und zwar zu einem Datum, das *vor* der Veröffentlichung der Fourier-Reihen (durch Fourier) liegt.

Das Problem des Umfangs des Datensatzes ist auch in der Praxis oft gar nicht hinderlich, da die Datensätze in der Regel manuell gewählt werden und so die Anzahl im Vorfeld sinnvoll bestimmt werden kann. Abbildung 17 zeigt ein Beispiel für die komplette Behandlung eines Datensatzes der Länge 8. Man sieht, dass man lediglich den anfänglichen Datensatz einmalig umsortieren muss, und dann in drei Schritten am Ziel ist (da $f_j = \mathcal{F}[f_j]$ für Datensätze der Länge 1) — entsprechend der Relation $8 = 2^3$. Die Umordnung nimmt man über eine binäre Darstellung des Index mit nachfolgender Bitumkehr vor (Details: Übung). Weiterführende Literatur ist [6].

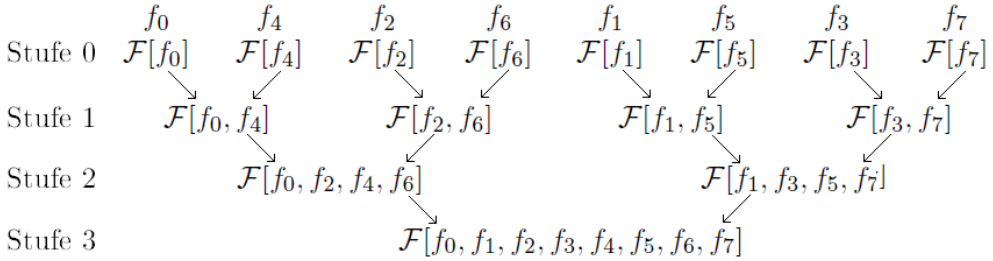

Abbildung 17: Beispiel für die FFT eines Datensatzes der Länge 8 mit einmaliger Umsortierung des ursprüglichen Datensatzes.

5.3 Beispiel: Bildkompression

Mit Hilfe der DFT bzw. der FFT ist die effiziente Kompression von Bildern möglich. Um das Prinzip näher zu erläutern, soll ein Bild zunächst nur aus Grauwerten bestehen. Die Idee ist nun, jede Bildzeile getrennt zu betrachten, und den Grauwert eines einzelnen Pixels k einer Zeile als Funktionswert f_k zu interpretieren. Daraus entsteht eine diskrete Funktion f, auf welche die DFT angewendet werden kann.

Um nun eine tatsächliche Kompression zu erhalten, muss abgeschätzt werden, auf welchem Anteil der so erhaltenen Transformation die 'wichtige' Informationen für das ursprüngliche Bild liegen. Z.B. kann man nur die ersten 20% speichern unter der Annahme, dass die letzten 80% vernachlässigbar sind. Dadurch benötigt man dann nur etwa 20% an Speicherbedarf der ursprünglichen Pixel-Werte f_k und kann durch die Inverse der DFT annähernd das vollständige Bild wiederherstellen. Dies ist eine sehr rudimentäre Herangehensweise, welche noch stark verbessert werden kann. Eine ausführliche Erklärung optimierter Verfahren findet sich in [9, Kap. 13].

6 Differentialgleichungen – eine kurze Wiederholung

Viele Zusammenhänge in der Natur lassen sich durch Differentialgleichungen beschreiben. Dazu gehören Vorgänge wie die Wärmeleitung, die Diffusion in Flüssigkeiten, Wellenphänomene aller Art, oder auch die zeitliche Entwicklung von Tierpopulationen. Eine Differentialgleichung beschreibt dabei, wie Ableitungen einer Funktion (etwa nach der Zeit oder nach räumlichen Komponenten) wieder von der ursprünglichen Funktion und evtl. der Zeit und den Koordinaten abhängig sind. Hier wollen wir nur gewöhnliche Differentialgleichungen ansehen, es kommen also keine partiellen Ableitungen vor. Allerdings wollen wir auch Differentialgleichungen höherer Ordnung ansehen, wie sie etwa bei Schwingungsphänomenen vorkommen.

Eine allgemeine Differentialgleichung (DGL) 1. Ordnung besitzt die Form

$$\dot{x}(t) \;:=\; \frac{\mathrm{d}}{\mathrm{d}t}\, x(t) \;=\; f(x(t), t)\,, \tag{31}$$

wobei wir hier die Zeit als Parameter nehmen. Oft schreibt man auch kurz $\dot{x} = f(x, t)$. In der Regel liegt weiter ein *Anfangswertproblem* (oder auch *Cauchy-Problem*) vor, d.h. es ist die Lösung dieser DGL gesucht mit einen gegebenen Anfangswert, $x(t_0) = x_0$. In einfachen Fällen kann man ein solches Problem explizit lösen. Falls nicht, gibt es verschiedene Verfahren aus der Numerik für eine näherungsweise (numerische) Lösung. An dieser Stelle wird ein iteratives Verfahren nach Volterra vorgestellt.

6.1 Volterra-Integration

Die Volterra-Integralgleichung entsteht durch Integration von Gl. (31) und lautet

$$x(t) \;=\; x_0 \,+\, \int_{t_0}^{t} f(x(\tau), \tau)\, \mathrm{d}\tau\,.$$

Diese Gleichung, deren Lösungen im Wesentlichen dieselben sind wie diejenigen des obigen Anfangswertproblems, kann iterativ eingesetzt werden. Mit $x_0(t) \equiv x_0$ als Startwert setzen wir dazu

$$x_{n+1}(t) \;:=\; \underbrace{x_0 \,+\, \int_{t_0}^{t} f(x_n(\tau), \tau)\, \mathrm{d}\tau}_{=:\,(\Phi(x_n))(t)}$$

für $n \in \mathbb{N}_0$. Dabei ist $\Phi\colon C^0[a,b] \to C^0[a,b]$ mit $t_0 \in [a,b]$ eine Abbildung, die wir noch genauer ansehen werden. Die Iteration wird — je nach Text — *Picard-Iteration* oder *Volterra-Iteration* genannt.

Beispiel 6.1. Die DGL $\dot{x}(t) = f(t)$ mit Anfangswert $x(t_0) = x_0$ kann einfach durch Integration gelöst werden,

$$x(t) \;=\; x_0 \,+\, \int_{t_0}^{t} f(\tau)\, \mathrm{d}\tau\,.$$

Etwas schwieriger ist die DGL $\dot{x} = f(x)$, deren Lösung mit der Methode der Trennung der Variablen berechnet werden kann. Durch Integration bekommt man

$$\int_{x_0}^{x(t)} \frac{\mathrm{d}\xi}{f(\xi)} \;=\; \int_{t_0}^{t} \frac{\dot{x}(\tau)}{f(x(\tau))}\,\mathrm{d}\tau \;=\; \int_{t_0}^{t} 1\,\mathrm{d}\tau \;=\; (t - t_0)$$

wobei angenommen sei, dass der Nenner nirgends verschwindet (sonst muss man natürlich anders vorgehen). Wenn man nun das linke Integral ausrechnen kann, löst man danach nach $x(t)$ auf. Diese Methode kommt etwa bei der logistischen Gleichung zum Einsatz; vgl. [12] für Details.

Wichtiger als die bisherigen Beispiele ist die folgende DGL, die wir später noch vektorwertig verallgemeinern werden.

Beispiel 6.2. Betrachten wir $\dot{x} = a\,x$ mit $x(0) = 1$, wobei $a \in \mathbb{R}$ beliebig, aber fest sei. Dann kann man induktiv nachrechnen, dass die Picard-Iteration auf

$$x_n(t) \;= \sum_{\ell=0}^{n} \frac{(a\,t)^\ell}{\ell!}$$

für alle $n \in \mathbb{N}_0$ führt (Übung: nachrechnen!). Aus der Analysis wissen wir, dass dann $\lim_{n\to\infty} x_n(t) = \mathrm{e}^{a\,t}$ gilt, mit punktweiser Konvergenz für jedes $t \in \mathbb{R}$ und gleichmäßiger Konvergenz auf jedem kompakten Intervall. In der Tat ist $\mathrm{e}^{a\,t}$ die *eindeutige* Lösung unseres Anfangswertproblems.

Die Situation im letzten Beispiel hat eine wichtige Verallgemeinerung, den Existenz- und Eindeutigkeitssatz für gewöhnliche DGLen.

Satz 6.3. *(Picard–Lindelöf)*
Gegeben sei eine Funktion $f\colon \mathbb{R} \times [a,b] \longrightarrow \mathbb{R}$. Sei f stetig und zudem Lipschitz bzgl. der x-Koordinate, mit Lipschitz-Konstante L. Sei außerdem $t_0 \in [a,b]$. Dann besitzt das Anfangswertproblem $\dot{x} = f(x,t)$ mit $x(t_0) = x_0$ eine eindeutige Lösung, die in t bis zum Rand des Intervalls fortgesetzt werden kann. Diese Lösung kann durch die Picard-Iteration beliebig genau approximiert werden.

Beweisskizze. Der klassische Beweis beruht auf einer Reduktion auf den Banach'schen Fixpuktsatz. Um ihn anwenden zu können, definieren wir uns zunächst eine geeignete Norm,

$$\|x\| \;:=\; \max_{t\in[a,b]} |x(t)|\,\mathrm{e}^{-\alpha t}$$

für ein (noch geschickt zu wählendes) $\alpha > 0$. Nun rechnet man nach, dass für die Abbildung Φ aus der Picard-Iteration die Ungleichung

$$\|\Phi(x) - \Phi(y)\| \;\le\; \frac{L}{\alpha}\,\|x - y\|$$

gilt, wobei L die Lipschitz-Konstante von f ist. Wählt man nun $\alpha > L$, so ist Φ eine Kontraktion, mit Kontraktionskonstante $\frac{L}{\alpha}$. Satz 2.13 liefert nun sowohl die Existenz als auch die Eindeutigkeit der Lösung, und zudem die exponentiell schnelle Konvergenz der Picard-Iteration gegen diese Lösung. □

Man kann eine ganze Reihe von Verallgemeinerungen dieses Satzes formulieren. So muss f i.A. in der x-Koordinate nicht auf ganz $\mathbb{R}$ definiert sein, oder die Lipschitz-Eigenschaft wird nur lokal gefordert — dementsprechend ändert sich dann immer ein wenig am Resultat. Dies ist sehr schön im Buch von Walter [12] ausgeführt. Außerdem kann man Gl. (31) auch als Vektor-Gleichung lesen — und alles Weitere danach lässt sich analog entwickeln. Man spricht dann von *Systemen* von DGLen erster Ordnung.

Beispiel 6.4. Das für uns wichtigste System ist das lineare, mit konstanten Koeffizienten. Dies schreiben wir als $\dot{\boldsymbol{x}} = A\boldsymbol{x}$ mit einer konstanten Matrix A, die i.A. komplexe Einträge haben darf. Explizit geschrieben sieht das so aus:

$$\dot{\boldsymbol{x}} \;=\; \begin{pmatrix} \dot{x}_1 \\ \vdots \\ \dot{x}_n \end{pmatrix} \;=\; A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \;=\; A\boldsymbol{x},$$

und die Anfangsbedingung lautet wieder $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$.

Die eindeutige Lösung lautet in diesem Fall $\boldsymbol{x}(t) = \exp\big((t-t_0)\,A\big)\boldsymbol{x}_0$, und man kann auch diese explizit aus der Picard-Iteration herleiten. Dabei verwendet man die Reihendefinition für e^A, die wir im nächsten Kapitel noch genauer besprechen werden. Die Eindeutigkeit kann man auch beweisen, indem man annimmt, dass $\boldsymbol{y}(t)$ eine Lösung ist und dann $\boldsymbol{z}(t) := \exp\big(-(t-t_0)A\big)\boldsymbol{y}(t)$ setzt. Nun zeigt man, dass $\dot{\boldsymbol{z}}(t) \equiv \mathbf{0}$ gilt, also $\boldsymbol{z}(t) = \mathit{const}$. An der Stelle t_0 kennen wir den Wert aber, das ist $\boldsymbol{z}(t_0) = \boldsymbol{x}_0$, und dann folgt $\boldsymbol{y}(t) = \exp\big((t-t_0)A\big)\,\boldsymbol{x}_0$.

6.2 Lineare Differentialgleichungen höherer Ordnung

Differentialgleichungen höherer Ordnung können behandelt werden, indem man sie auf DGL-Systeme 1. Ordnung zurückführt.

Für die Lösung von DGLs 2. Ordnung soll dies zunächst an einem Beispiel erläutert werden, nämlich anhand der Schwingungsgleichung $\ddot{x}+x = 0$. Diese besitzt die allgemeine Lösung $x(t) = \alpha \cdot \cos(t) + \beta \cdot \sin(t)$. Die Parameter α und β werden dann eindeutig festgelegt durch die Vorgabe von

$$\begin{aligned} x(t_0) &= \alpha\cos(t_0) + \beta\sin(t_0)\,, \\ \dot{x}(t_0) &= -\alpha\sin(t_0) + \beta\cos(t_0)\,. \end{aligned}$$

Durch Auflösen bekommt man

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \;=\; \begin{pmatrix} \cos(t_0) & -\sin(t_0) \\ \sin(t_0) & \cos(t_0) \end{pmatrix} \begin{pmatrix} x(t_0) \\ \dot{x}(t_0) \end{pmatrix}.$$

Damit ist die Lösung dann eindeutig bestimmt.

Dieses Ergebnis erhält man durch Rückführung auf ein Paar an Differentialgleichungen 1. Ordnung mittels

$$x_1 := \dot{x} \quad \text{und} \quad x_2 := x\,.$$

Damit bekommt man dann das System

$$\begin{pmatrix}\dot{x}_1\\ \dot{x}_2\end{pmatrix} \;=\; \underbrace{\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}}_{=:A} \begin{pmatrix}x_1\\ x_2\end{pmatrix}. \tag{32}$$

mit den Anfangswerten $x_1(t_0) = \dot{x}(t_0)$ und $x_2(t_0) = x(t_0)$. Für eine Differentialgleichung n-ter Ordnung funktioniert diese Rückführung auf ein n-Tupel von Differentialgleichungen 1. Ordnung analog.

Die Lösung des Systems (32) kann man nun folgendermaßen berechnen. Zunächst zeigt man, dass $A^{2m} = (-1)^m \mathbb{1}$ und $A^{2m+1} = (-1)^m A$ gilt, für alle $m \in \mathbb{N}_0$. Damit können wir $\exp(tA)$ berechnen, und bekommen durch Aufspalten der Summe nach geraden und ungeraden Exponenten

$$\mathrm{e}^{tA} = \sum_{n=0}^{\infty} \frac{t^n}{n!} A^n = \underbrace{\sum_{m\geq 0} (-1)^m \frac{t^{2m}}{(2m)!}}_{=\cos(t)} \mathbb{1} + \underbrace{\sum_{\ell\geq 0} (-1)^\ell \frac{t^{2\ell+1}}{(2\ell+1)!}}_{=\sin(t)} A = \begin{pmatrix}\cos(t) & -\sin(t)\\ \sin(t) & \cos(t)\end{pmatrix}.$$

Daraus folgt nun die zuvor bereits angegebene Lösung der DGL 2. Ordnung (mit Anfangswerten):

$$\begin{pmatrix}x_1\\ x_2\end{pmatrix} = \exp\big((t-t_0)A\big) \begin{pmatrix}x_1(t_0)\\ x_2(t_0)\end{pmatrix} = \begin{pmatrix}\cos(t-t_0) & -\sin(t-t_0)\\ \sin(t-t_0) & \cos(t-t_0)\end{pmatrix} \begin{pmatrix}x_1(t_0)\\ x_2(t_0)\end{pmatrix}.$$

Dies stimmt mit der obigen Lösung überein.

Eine lineare Differentialgleichung höherer Ordnung mit konstanten Koeffizienten besitzt die allgemeine Form

$$x^{(n)} + a_1 x^{(n-1)} + a_2 x^{(n-2)} + \cdots + a_{n-1}\dot{x} + a_n x = 0. \tag{33}$$

Dies kann man nun wie oben erwähnt umformen in ein System n-ter Ordnung, und erhält dann eine $n \times n$-Matrix A, deren Eigenwerte und ggf. Eigenvektoren zur Lösung herangezogen werden.

Das kann man aber auch abkürzen, mit einem Ansatz von Euler. Dabei wird ein Polynom n-ter Ordnung aufgestellt, welches nach dem Fundamentalsatz der Algebra genau n Nullstellen $\lambda_1, \ldots, \lambda_n$ in $\mathbb{C}$ besitzt. Mit $x(t) := \mathrm{e}^{\lambda t}$ bekommen wir durch Einsetzen in Gl. (33) in der Tat

$$\underbrace{(\lambda^n + a_1\lambda^{n-1} + \ldots + a_{n-1}\lambda + a_n)}_{=:p(\lambda)} \underbrace{\mathrm{e}^{\lambda t}}_{\neq 0 \text{ auf ganz } \mathbb{C}} = 0.$$

Diese Gleichung kann nur stimmen, wenn $p(\lambda) = 0$ erfüllt ist, was uns die n Nullstellen (mit Vielfachheiten gezählt) verschafft.

Falls nun die n Nullstellen paarweise verschieden sind (was der generische Fall ist), so lautet die allgemeine Lösung der DGL (33) einfach

$$x(t) = \sum_{\ell=1}^{n} c_\ell \, \mathrm{e}^{\lambda_\ell t}.$$

Mit Hilfe der Anfangsbedingungen $x(t_0), \dot{x}(t_0), \ldots, x^{(n-1)}(t_0)$ lässt sich ein lineares Gleichungssystem aufstellen, das die Koeffizienten c_ℓ eindeutig festlegt.

Wenn Multiplizitäten bei den Nullstellen auftreten, so hängt die Form der allgemeinen Lösung davon ab, ob die Matrix A, die wir hier nicht explizit hingeschrieben haben, diagonalisierbar ist oder nicht. Für Details hierzu sei auf [12] verwiesen.

7 Markov-Ketten

Um das Matrix-Exponential im nächsten Kapitel einsetzen zu können, sind zunächst einige Vorüberlegungen notwendig.

7.1 Matrix-Normen und Matrix-Exponential

Für den Vektorraum $V = \mathbb{R}^n$ (oder $\mathbb{C}^n$) mit einer beliebigen Norm $\|\cdot\|$ und einer linearen Abbildung $A\colon V \to V$ ist

$$|||A||| := \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|}$$

die zugehörige *Operatornorm*. Dabei gilt

$$|||A||| = \sup_{\|x\|=1} \|Ax\| = \max_{\|x\|=1} \|Ax\|.$$

In der Gleichung ist das Supremum gleich dem Maximum, da die Einheitskugel $\{\|x\| \leq 1\}$ im $\mathbb{R}^n$ (bzw. im $\mathbb{C}^n$) kompakt ist, und daher ihr Rand, der eine abgeschlossene Teilmenge ist, auch. Dieser Zusammenhang ist nicht trivial und bedeutet, dass der Extremwert für ein x mit $\|x\| = 1$ wirklich angenommen wird. Es gelten weiter folgende Eigenschaften:

$$|||A||| \geq 0, \text{ mit } |||A||| = 0 \Leftrightarrow A = \mathbf{0} \tag{34}$$
$$|||cA||| = |c|\,|||A||| \text{ für alle } c \in \mathbb{R}\ (c \in \mathbb{C}) \tag{35}$$
$$|||A+B||| \leq |||A||| + |||B||| \tag{36}$$
$$\|Ax\| \leq |||A|||\,\|x\| \tag{37}$$
$$|||AB||| \leq |||A|||\,|||B||| \tag{38}$$

Dabei beschreiben die Eigenschaften (34), (35) und (36) die üblichen Norm-Axiome und (37) ist eine Verträglichkeits-Bedingung. Schließlich ist (38) eine wichtige neue Eigenschaft. Man überlege sich, warum hier i.A. keine Gleichheit gelten kann!

Als Beispiele werden an dieser Stelle die Matrixnormen zu bereits bekannten Normen aufgestellt für eine $n \times n$-Matrix A mit Elementen in $\mathbb{C}$ (kurz: $A \in \mathrm{Mat}(n, \mathbb{C})$). Die wichtigsten Beispiele sind in Tabelle 2 zusammengefasst. Für eine dieser Normen wird noch ein neuer Begriff benötigt.

Norm $\|.\|$	Operatornorm	$\|\|\|.\|\|\|$
$\|.\|_1$	Spaltensummennorm:	$\|\|\|A\|\|\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} \|a_{ij}\|$
$\|.\|_2$	Spektralnorm:	$\|\|\|A\|\|\|_2 = \sqrt{\varrho(A^+A)}$, mit $A^+ = \overline{A^t}$
$\|.\|_\infty$	Zeilensummennorm:	$\|\|\|A\|\|\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} \|a_{ij}\|$

Tabelle 2: Einige gängige Matrixnormen

Definition 7.1. Das *Spektrum* einer Matrix $A \in \mathrm{Mat}(n, \mathbb{C})$ ist die Menge ihrer Eigenwerte, also

$$\sigma(A) \;=\; \{\lambda \in \mathbb{C} \mid \lambda \text{ ist EW von } A\}\,.$$

Dann heißt

$$\varrho(A) \;=\; \max_{\lambda\in\sigma(A)} |\lambda|$$

der *Spektralradius* von A und definiert den kleinsten Kreis um 0 in $\mathbb{C}$, in dem alle Eigenwerte der Matrix A liegen.

Es gibt verschiedene Ansätze, das Matrix-Exponential e^A zu betrachten. Wir beginnen mit der Definition mittels der Potenzreihe der Exponential-Funktion, und erhalten (zunächst formal)

$$\mathrm{e}^A \;=\sum_{n=0}^{\infty} \frac{A^n}{n!}\,. \tag{39}$$

Wie auch bei der Exponential-Funktion aus Analysis I ist die Konvergenz der Reihe von großer Bedeutung. Daher wird zunächst das Konvergenzverhalten geprüft.

$$\Big\|\!\Big| \sum_{m=0}^{N} \frac{A^m}{m!} \Big|\!\Big\| \le \sum_{m=0}^{N} \frac{\|\!|A^m|\!\|}{m!} \le \sum_{m=0}^{N} \frac{\|\!|A|\!\|^m}{m!} \le \sum_{m=0}^{\infty} \frac{\|\!|A|\!\|^m}{m!} \;=\; \mathrm{e}^{\|\!|A|\!\|} \;<\; \infty\,.$$

Daraus folgt die Konvergenz der Reihe für jede gegebene Matrix A, weil wir hier eine Majorante für jedes Element der durch Gl. (39) definierten Matrix berechnet haben (man mache sich die Details als Übung klar).

Beispiel 7.2. Wir berechnen e^{tA} für $A = \left(\begin{smallmatrix} 1 & 1 \\ -1 & -1 \end{smallmatrix}\right)$. Da $A^2 = 0$, wird die Reihe einfach:

$$\mathrm{e}^{tA} \;=\; \mathbb{1} + tA + 0 \;=\; \begin{pmatrix} 1+t & t \\ -t & 1-t \end{pmatrix}.$$

Für $A = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ ist die Rechnung etwas komplizierter. Da $A^2 = \mathbb{1}$, können wir die Summe nach geraden und ungeraden Indices aufteilen, und bekommen (Details: Übung):

$$\mathrm{e}^{tA} \;=\; \cosh(t)\,\mathbb{1} + \sinh(t)\,A \;=\; \begin{pmatrix} \cosh(t) & \sinh(t) \\ \sinh(t) & \cosh(t) \end{pmatrix}.$$

Aus unseren obigen Überlegungen folgt auch, dass e^{tA} für jedes $t \in \mathbb{R}$ wohldefiniert ist. Dabei gilt sogar gleichmäßige Konvergenz auf kompakten Intervallen für den Parameter t. Insbesondere gilt $\mathrm{e}^{0\cdot A} = \mathrm{e}^{\mathbf{0}} = \mathbb{1}$.

Nun betrachter wir die Ableitung:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{tA} \;&=\; \frac{\mathrm{d}}{\mathrm{d}t} \sum_{m\ge 0} \frac{t^m A^m}{m!} \;=\; \sum_{m\ge 1} \frac{t^{m-1}}{(m-1)!}\, A^m \\ &=\; A \underbrace{\sum_{m\ge 1} \frac{(tA)^{m-1}}{(m-1)!}}_{=\mathrm{e}^{tA}} \;=\; A\,\mathrm{e}^{tA} \;=\; \mathrm{e}^{tA} A\,. \end{aligned}$$

Man beachte, dass der letzte Schritt gilt, weil wir im Schritt davor genauso gut ein A nach hinten hätten herausziehen können. Die Matrizen A und e^{tA} vertauschen also für alle $t \in \mathbb{R}$. Diese Rechnung legt nun außerdem das folgende Resultat nahe.

Satz 7.3. *Sei $A \in \mathrm{Mat}(n, \mathbb{C})$ beliebig, aber fest. Dann besitzt das matrixwertige Anfangswertproblem $\dot{M} = AM$ mit $M(0) = \mathbb{1}$ die eindeutige Lösung $M(t) = \mathrm{e}^{tA}$.*

Beweis. Die Lösungseigenschaft ist nach obiger Rechnung klar, die Eindeutigkeit könnte mittels Picard–Lindelöf gezeigt werden. Der Beweis ist auch durch eine konkrete Rechnung möglich. Dazu sei $N(t)$ eine beliebige Lösung. Wir betrachten nun $R(t) = \mathrm{e}^{-tA}\, N(t)$ und werden nachweisen, dass R eine konstante Matrixfunktion ist.

Dazu berechnen wir den Differenzenquotienten (für kleine $h > 0$) wie folgt:

$$\begin{aligned}
\frac{1}{h}\big(R(t+h) - R(t)\big) &= \frac{1}{h}\big(\mathrm{e}^{-(t+h)A}\,[N(t+h) - N(t)] + [\mathrm{e}^{-(t+h)A} - \mathrm{e}^{-tA}]N(t)\big) \\
&= \mathrm{e}^{-(t+h)A}\,\frac{[N(t+h) - N(t)]}{h} + \frac{[\mathrm{e}^{-(t+h)A} - \mathrm{e}^{tA}]}{h}\,N(t) \\
&\xrightarrow{h\searrow 0} \mathrm{e}^{-tA}\,\underbrace{\dot{N}(t)}_{=A\cdot N(t)} - A\mathrm{e}^{-tA}N(t) \;\equiv\; 0 \;\equiv\; \dot{R}(t)\,.
\end{aligned}$$

Also ist $R(t)$ konstant und es folgt, dass $R(t) = R(0) = \mathbb{1}$ gilt. Es ergibt sich also $N(t) = \mathrm{e}^{tA}$ wie behauptet. □

Eine alternative Herleitung (wie bereits angedeutet) liefert die Lösung einer Differentialgleichung mittels Picard-Iteration. Definiere dafür e^{tA} als Lösung des Anfangswertproblems $\dot{M} = AM$ mit $M(0) = \mathbb{1}$. Daraus ergibt sich

$$M(t) \;=\; \underbrace{M(0)}_{=\mathbb{1}} + \int_0^t A\,M(\tau)\,\mathrm{d}\tau$$

durch Integration. Nun können wir wieder iterieren:

$$M^{(0)}(t) \;=\; \mathbb{1} \quad \text{und} \quad M^{(n+1)}(t) \;:=\; \mathbb{1} + \int_0^t A\,M^{(n)}(\tau)\,\mathrm{d}\tau \quad \text{für alle } n \in \mathbb{N}_0\,.$$

Analog zu unserer eindimensionalen Picard-Iteration ergibt sich hier induktiv (nachrechnen!) die Formel

$$M^{(N)}(t) \;=\; \sum_{m=0}^{N} \frac{t^m A^m}{m!} \;\xrightarrow{N\to\infty}\; \sum_{m=0}^{\infty} \frac{t^m A^m}{m!},$$

und wir bekommen wieder die Potenzreihe, mit der wir weiter oben begonnen hatten. Welchen Zugang man bevorzugt, ist Geschmacksache — sie sind äquivalent.

Es gibt noch eine andere wichtige Eigenschaft der Exponentialfunktion, die über eine Funktionalgleichung beschrieben wird.

Satz 7.4. *Sei* $T\colon \mathbb{R}_{\geq 0} \to \mathrm{Mat}(n,\mathbb{C})$ *stetig, und erfülle die Funktionalgleichung*

$$T(t+s) \;=\; T(t)\,T(s)$$

für alle $t, s \geq 0$ *zusammen mit der Anfangsbedingung*

$$T(0) \;=\; \mathbb{1}\,.$$

Dann existiert eine Matrix $A \in \mathrm{Mat}(n,\mathbb{C})$*, so dass* $T(t) = \mathrm{e}^{tA}$ *für alle* $t \geq 0$*.*

Für einen Beweis verweisen wir auf [8, S. 3 und S. 11]. Der entscheidende Punkt in dieser Aussage ist, dass es eine *feste* Matrix A gibt, so dass $T(t) = \mathrm{e}^{tA}$ gilt, wobei diese Matrix dann aus anderen Daten bestimmt werden muss, die nicht in der Funktionalgleichung vorkommen. Diese Situation tritt etwa in der Phylogenie auf, wenn man Mutations-Generatoren in Stammbäumen bestimmen muss.

Nun kommen wir zu einer Besonderheit in der Berechnung von Matrix-Exponentialen. Sie leitet sich daraus ab, dass der Kommutator zweier Matrizen A und B, also

$$[A,B] \;:=\; AB - BA$$

i.A. nicht verschwindet, weil die Matrizen nicht miteinander vertauschen. Falls $[A,B] = 0$ gilt, können wir den binomischen Lehrsatz in der üblichen Form benutzen und bekommen

$$\begin{aligned}
\mathrm{e}^{A+B} \;&=\; \sum_{m\geq 0} \frac{(A+B)^m}{m!} \;=\; \sum_{m\geq 0} \frac{1}{m!} \sum_{\ell=0}^{m} \binom{m}{\ell} A^\ell B^{m-\ell} \\
&=\; \underbrace{\sum_{\ell=0}^{\infty} \frac{A^\ell}{\ell!}}_{=\mathrm{e}^A} \underbrace{\sum_{m\geq \ell} \frac{B^{m-\ell}}{(m-\ell)!}}_{=\sum_{k\geq 0}\frac{B^k}{k!}=\mathrm{e}^B} \;=\; \mathrm{e}^A\,\mathrm{e}^B\,.
\end{aligned}$$

Gilt aber $[A,B] \neq 0$, ist der zweite Schritt nicht durchführbar, da AB nicht mit BA vertauscht werden darf:

$$\begin{aligned}
(A+B)^2 \;&=\; A^2 + AB + BA + B^2 \;=\; A^2 + 2AB + B^2 - [A,B] \quad \text{und} \\
(A+B)^3 \;&=\; A^3 + 2A^2B + 2AB^2 + B^3 - [A,[A,B]] + [B,[A,B]]\,.
\end{aligned}$$

Entsprechene Formeln gelten für höhere Potenzen.

Satz 7.5. *(BCH-Formel nach Baker, Campbell, Hausdorff)*

$$\underbrace{\log(\mathrm{e}^A\,\mathrm{e}^B) \;=\; A + B + \tfrac{1}{2}[A,B]}_{\textit{häufig eingesetzte Approximation}} + \tfrac{1}{12}([A,[A,B]] - [B,[A,B]]) - \tfrac{1}{24}[B,[A,[A,B]]] + \ldots$$

In Zusammenhang mit der allgemeinen Formel ist auch Dynkin (1947) zu nennen.

Es existieren weitere Formeln, welche zur Approximation eingesetzt werden und verwandt zur BCH-Formel sind:

$$\text{Zassenhaus:}\quad \mathrm{e}^{t(A+B)} \;=\; \underbrace{\mathrm{e}^{tA}\,\mathrm{e}^{tB}\,\mathrm{e}^{-\frac{1}{2}t^2[A,B]}}_{\text{häufige Approximation}}\;\mathrm{e}^{\frac{1}{6}t^3(\ldots)}\cdot\ldots$$

$$\text{Hadamard:}\quad \mathrm{e}^{A}\,B\,\mathrm{e}^{-A} \;=\; \underbrace{B+[A,B]}_{\text{häufige Approx.}}\;+\tfrac{1}{2}[A,[A,B]]+\tfrac{1}{6}[A,[A,[A,B]]]+\ldots$$

Diese und weitere Formeln finden sich auch in den online verfügbaren Lexika.

Um später besser mit speziellen Matrizen (den Markov-Generatoren) umgehen zu können, folgen an dieser Stelle einige allgemeine Rechenregeln und Definitionen.

$$\det(\mathrm{e}^{A}) \;=\; \mathrm{e}^{\mathrm{tr}(A)} \quad \text{mit der Spur} \quad \mathrm{tr}(A)=\sum_i A_{ii}\,. \tag{40}$$

Einen Beweis dieser sehr nützlichen Formel kann man rasch für diagonalisierbare Matrizen erbringen, denn dann hat man $A = S^{-1}DS$ mit $D = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$. Also bekommt man

$$S\mathrm{e}^{A}S^{-1} \;=\; \mathrm{e}^{SAS^{-1}} \;=\; \mathrm{e}^{\mathrm{diag}(\lambda_1,\ldots,\lambda_n)} \;=\; \mathrm{diag}(\mathrm{e}^{\lambda_1},\ldots,\mathrm{e}^{\lambda_n})$$

und folglich auch

$$\det(\mathrm{e}^{A}) \;=\; \det(S\mathrm{e}^{A}S^{-1}) \;=\; \det\big(\mathrm{diag}(\mathrm{e}^{\lambda_1},\ldots,\mathrm{e}^{\lambda_n})\big) \;=\; \prod_{i=1}^{n}\mathrm{e}^{\lambda_i} \;=\; \mathrm{e}^{\sum_{i=1}^{n}\lambda_i} \;=\; \mathrm{e}^{\mathrm{tr}(A)},$$

wobei wir uns daran erinnern, dass die Spur eine Matrix mit der Summe ihrer Eigenwerte (mit Multiplizitäten) übereinstimmt.

Um die Formel (40) in voller Allgemeinheit ableiten zu können, kann man die Jordan'sche Normalform komplexer Matrizen einsetzen. Eine Variante besagt, dass man A stets als $A = B + N$ mit B diagonalisierbar und N nilpotent zerlegen kann, so dass $[B,N] = \mathbf{0}$ gilt. Nilpotent bedeutet, dass $N^m = \mathbf{0}$ für ein $m \in \mathbb{N}$ gilt. Dann folgt

$$\mathrm{e}^{A} \;=\; \mathrm{e}^{B+N} \;=\; \mathrm{e}^{B}\,\mathrm{e}^{N},$$

wobei wir schon wissen, wie wir e^{B} und seine Determinante berechnen. Wegen der Nilpotenz von N ist e^{N} ein Polynom in N mit Grad kleiner als m, und man zeigt dann noch, dass $\det(\mathrm{e}^{N}) = 1$. Wegen $\det(\mathrm{e}^{A}) = \det(\mathrm{e}^{B})\det(\mathrm{e}^{N}) = \det(\mathrm{e}^{B})$ folgt dann unsere Behauptung.

7.2 Markov-Matrizen und Markov-Generatoren

Wir beginnen mit einer kurzen Wiederholung aus der Wahrscheinlichkeitstheorie, wo wie uns schon genauer mit Markov-Ketten in diskreter Zeit befasst hatten. Eine gute allgemeine Referenz für den Hintergrund zu diesem Kapitel ist das Buch von Norris [4].

Definition 7.6. Eine Matrix $M \in \mathrm{Mat}(n, \mathbb{R})$, notiert als $M = (M_{ij})_{1 \le i,j \le n}$, heißt *Markov-Matrix*, falls die folgenden beiden Eigenschaften erfüllt sind:

(1) $M_{ij} \ge 0$ gilt für alle $1 \le i, j \le n$;

(2) $\sum_j M_{ij} = 1$ gilt für alle $1 \le i \le n$.

Im weiteren Verlauf wurden absorbierende Zustände diskutiert, aber auch wichtige Kriterien für verschiedene Stufen von Irreduzibilität besprochen (eine Wiederholung des Stoffes ist dringend empfohlen!), nämlich

$$\begin{aligned} \text{irreduzibel:} &\quad \forall\, i, j \colon \exists\, m \in \mathbb{N} \colon (M^m)_{ij} > 0\,, \\ \text{primitiv:} &\quad \exists\, m \colon \forall\, i, j \colon (M^m)_{ij} > 0\,, \\ \text{aperiodisch:} &\quad \mathrm{ggT}(\text{Zykel-Längen in } \mathcal{G}_M) = 1\,. \end{aligned}$$

Dabei bezeichnet $\mathcal{G}_M$ den M zugeordneten Graphen, der n Knoten besitzt und eine gerichtete Kante von i nach j genau dann, wenn $M_{ij} > 0$. Man kann die obigen Eigenschaften rasch anhand von $\mathcal{G}_M$ ermitteln. Irreduzibel etwa bedeutet, dass man von jedem Knoten zu jedem Knoten entlang der 'Einbahnstraßen' kommen kann. Es sei noch angemerkt, dass die obige, vereinfachte Charakterisierung von 'aperiodisch' so i.A. nur für irreduzible Matrizen stimmt, was uns hier aber ausreicht. Die Eigenschaften stehen dabei in folgender Relation:

Satz 7.7. *Eine Markov-Matrix ist primitiv genau dann wenn sie irreduzibel und aperiodisch ist.*

Nun entwickeln wir das Analogon für stetige Zeit.

Definition 7.8. Eine Matrix $A \in \mathrm{Mat}(n, \mathbb{R})$ heißt *Markov-Generator*, falls die folgenden beiden Eigenschaften erfüllt sind:

(1) $A_{ij} \ge 0$ gilt für alle $i \ne j$;

(2) $\sum_j A_{ij} = 0$ gilt für alle i.

Ein Beispiel für einen Markov-Generator ist

$$A = \begin{pmatrix} -4 & 1 & 3 \\ 0 & -1 & 1 \\ 1 & 1 & -2 \end{pmatrix}. \tag{41}$$

Um auch einer solchen Matrix einen Graphen zuordnen zu können, vereinbaren wir, dass der Graph $\mathcal{G}_A$ eine gerichtete Kante von i nach j genau dann besitzt, falls $A_{ij} > 0$. Es

kann also niemals eine Kante von i nach i geben, da $A_{ii} \leq 0$, und somit gibt es keine Schleifen. Abb. 18 illustriert $\mathcal{G}_A$, und wir können hier die Irreduzibilität ablesen, da wir von jedem Knoten starten können und doch entlang der gerichteten Kanten zu jedem Knoten gelangen können. Die Beziehung zwischen irreduzibel und primitiv wird sich hier etwas anders gestalten, wie wir gleich sehen werden.

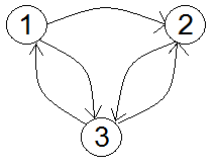

Abbildung 18: Graph $\mathcal{G}_A$ zum Markov-Generator A von Gl. (41).

Satz 7.9. *Ist A ein Markov-Generator, so ist $\{\mathrm{e}^{tA} \mid t \geq 0\}$ eine Halbgruppe (sogar Monoid) von Markov-Matrizen. Ist e^{tA} Markov für alle $t \geq 0$, so ist A ein Markov-Generator. Es reicht sogar, dass e^{tA} Markov-Matrix für alle t in einem kleinen Intervall $[0, \varepsilon]$ mit $\varepsilon > 0$ ist.*

Beweis. Für diesen Beweis benötigen wir die Exponentialreihe, also

$$\mathrm{e}^{tA} \;=\; \mathbb{1} + tA + \frac{t^2}{2}A^2 + \ldots\,,$$

die wir in mehrfacher Weise einsetzen werden.

Sei nun A ein Markov-Generator. Dann gilt für ein beliebiges $1 \leq i \leq n$:

$$\begin{aligned}\sum_j (\mathrm{e}^{tA})_{ij} \;&=\; 1 + \sum_{m\geq 1} \frac{t^m}{m!} \sum_{j=1}^{n} (A^n)_{ij} \;=\; 1 + \sum_{m\geq 1} \frac{t^m}{m!} \sum_{\ell,j=1}^{n} (A^{m-1})_{i\ell}\, A_{\ell j} \\ &=\; 1 + \sum_{m\geq 1} \frac{t^m}{m!} \sum_{\ell=1}^{n} (A^{m-1})_{i\ell} \underbrace{\sum_{j=1}^{n} A_{\ell j}}_{=0} \;=\; 1\,,\end{aligned}$$

woraus sich die Zeilensummeneigenschaft von $M(t) := \mathrm{e}^{tA}$ ergibt. Um zu zeigen, dass $M_{ij}(t) \geq 0$ für alle i, j gilt, beachten wir, dass für kleine $t > 0$ gilt:

$$M(t) \;=\; \mathrm{e}^{tA} \;=\; \mathbb{1} + tA + \mathcal{O}(t^2)\,, \tag{42}$$

wobei dies genauer so zu lesen ist, dass es für ein gegebenes $\varepsilon > 0$ zu jedem Indexpaar i, j eine Konstante $c_{ij} > 0$ gibt, so dass das Restglied in $M_{ij}(t) = \delta_{ij} + tA_{ij} + R_{ij}^{(2)}(t)$ die Ungleichung

$$|R_{ij}^{(2)}(t)| \;\leq\; c_{ij}\, t^2$$

erfüllt, und zwar für alle $t \in [0, \varepsilon]$.

Falls nun $A_{ij} > 0$, was nur für $i \neq j$ der Fall sein kann, so wählen wir $t > 0$ so klein, dass $tA_{ij} > c_{ij}t^2$ gilt. Das ist dann für *alle* hinreichend kleinen t wahr, und wir sehen, dass dann $(\mathrm{e}^{tA})_{ij} > 0$ für alle hinreichend kleinen $t \geq 0$. Betrachten wir nun den Fall $A_{ij} = 0$ mit $i \neq j$. Hier ist natürlich $M_{ij}(0) = 0$. Wenn nun $M_{ij}(t) < 0$ für $0 < t < \eta$ (und ein sehr kleines $\eta > 0$) wäre, so bekämen wir aus Gl. (42) als Konsequenz aber, dass $A_{ij} < 0$ gelten müsste — was im Widerspruch zur Generator-Eigenschaft von A stünde. Also kann es so ein Intervall nicht geben, und wir haben $M_{ij}(t) \geq 0$ für (mindestens) ein kleines Intervall $[0, \eta']$ mit $\eta' > 0$.

Für die Diagonale, wo wir ja $A_{ii} \leq 0$ haben, können wir Gl. (42) noch vereinfachen zu $M_{ii}(t) = 1 + \mathcal{O}(t)$, und es ist unmittelbar klar, dass sogar $M_{ii}(t) > 0$ für alle hinreichend kleinen t gelten muss. Damit haben wir insgesamt nachgewiesen, dass $M(t)$ für alle hinreichend kleinen t eine Markov-Matrix ist.

Nun bedienen wir uns eines Tricks, und bemerken, dass

$$M(t) \;=\; \mathrm{e}^{tA} \;=\; \bigl(\mathrm{e}^{\frac{t}{n}A}\bigr)^n \;=\; M\bigl(\tfrac{t}{n}\bigr)^n$$

gilt, und zwar für jedes $n \in \mathbb{N}$. Somit können wir für ein gegebenes $t > 0$ immer ein n finden, so dass $M(\frac{t}{n})$ nach unserem obigen Argument eine Markov-Matrix ist. Dann muss aber auch $M(t)$ eine Markov-Matrix sein, und die eine Richtung des Beweises ist fertig.

Für die Rückrichtung sei nun $\varepsilon > 0$ beliebig, aber fest, und $M(t) = \mathrm{e}^{tA}$ sei Markov für alle $t \in [0, \varepsilon]$. Dann gilt $\sum_j M_{ij}(t) = 1$ für alle $1 \leq i \leq n$ und alle $t \in [0, \varepsilon]$, und wir bekommen durch Differentiation:

$$0 \;=\; \frac{\mathrm{d}}{\mathrm{d}t} \sum_{j=1}^{n} (\mathrm{e}^{tA})_{ij} \;=\; \sum_{j=1}^{n} (\mathrm{e}^{tA}\, A)_{ij} \;=\; \sum_{\ell=1}^{n} (\mathrm{e}^{tA})_{i\ell} \sum_{j=1}^{n} A_{\ell j}\,.$$

Diese Relation können wir nun bei jedem $t \in [0, \varepsilon]$ einsetzen, insbesondere also auch bei $t = 0$. Da nun $(\mathrm{e}^{tA})_{i\ell}\big|_{t=0} = \delta_{i\ell}$ gilt, bekommen wir damit

$$0 \;=\; \sum_{\ell=1}^{n} \delta_{i\ell} \sum_{j=1}^{n} A_{\ell j} \;=\; \sum_{j=1}^{n} A_{ij} \;=\; 0\,,$$

womit wir die Zeilensummenbedingung bestätigt haben. Für $i \neq j$ betrachten wir noch einmal die Exponentialreihe, und bemerken, dass wir A_{ij} dann folgendermaßen als Grenzwert bekommen können,

$$A_{ij} \;=\; \lim_{h \searrow 0} \frac{1}{h} \bigl(\mathrm{e}^{hA} - \mathbb{1}\bigr)_{ij} \;\overset{i \neq j}{=}\; \lim_{h \searrow 0} \frac{1}{h} \bigl(\mathrm{e}^{hA}\bigr)_{ij} \;\geq\; 0\,.$$

Daraus ergibt sich die fehlende Bedingung, und A ist ein Markov-Generator. □

Die Nichtnegativität der $M_{ij}(t)$ kann man auch anders (und sogar schneller) einsehen, wenn man verwendet, dass

$$\mathrm{e}^{tA} \;=\; \lim_{n \to \infty} \bigl(\mathbb{1} + \tfrac{t}{n} A\bigr)^n$$

gilt, was eine Verallgemeinerung der bekannten skalaren Identität für die e-Funktion auf Matrizen ist. Für jedes beliebige, aber feste $t \geq 0$ gibt es sicher ein $n \in \mathbb{N}$, so dass $\mathbb{1} + \frac{t}{n}A$ eine Matrix mit ausschließlich nichtnegativen Elementen ist, weil A ja negative Einträge nur auf der Diagonalen haben kann. Natürlich ist dann auch $(\mathbb{1} + \frac{t}{n}A)^n$ nichtnegativ, und das bleibt so, wenn n wächst. Diese Eigenschaft ist nun auch im Limes $n \to \infty$ erhalten, und man bekommt $M_{ij}(t) \geq 0$ für alle $t \geq 0$ in einem Schritt.

Nun kommen wir zum wichtigen Zusammenhang zwischen 'irreduzibel' und 'primitiv', der bei Markov-Ketten in stetiger Zeit etwas überraschend ist.

Satz 7.10. *Ist A ein irreduzibler Markov-Generator, so ist e^{tA} für alle $t > 0$ primitiv.*

Beweis. Ist $A_{ij} > 0$, so auch $(\mathrm{e}^{tA})_{ij} > 0$ für hinreichend kleine t, wie wir aus dem vorigen Beweis gesehen haben, und damit überhaupt für alle $t > 0$. Da A irreduzibel ist, existiert für alle i, j ein (gerichteter) Weg von i nach j. Nehmen wir an, ein solcher Weg sei $i = i_0 \to i_1 \to \ldots \to i_m = j$. Dabei sind dann alle $A_{i_\ell, i_{\ell+1}} > 0$. Daher bekommen wir dann die Ungleichung

$$\begin{aligned}(\mathrm{e}^{tA})_{ij} &= \left((\mathrm{e}^{\frac{t}{m}A})^m\right)_{ij} = \sum_{j_1=1}^{n} \cdots \sum_{j_{m-1}=1}^{n} (\mathrm{e}^{\frac{t}{m}A})_{i_0 j_1} (\mathrm{e}^{\frac{t}{m}A})_{j_1 j_2} \cdots (\mathrm{e}^{\frac{t}{m}A})_{j_{m-1} i_m} \\ &\geq \underbrace{(\mathrm{e}^{\frac{t}{m}A})_{i_0 i_1}}_{>0} \underbrace{(\mathrm{e}^{\frac{t}{m}A})_{i_1 i_2}}_{>0} \cdots \underbrace{(\mathrm{e}^{\frac{t}{m}A})_{i_{m-1} i_m}}_{>0} > 0.\end{aligned}$$

Es existiert also mindestens ein Term, in dem jeder Faktor > 0 ist und damit auch $(\mathrm{e}^{tA})_{ij} > 0$. Da dies nun für alle $t > 0$ und alle i, j gilt, ist die Behauptung klar. □

Aus dem Beweis des vorigen Satzes ergibt sich noch eine recht erstaunliche Folgerung:

Folgerung 7.11. *Ist e^{tA} irreduzibel für irgendein $t > 0$, so ist e^{tA} primitiv für alle $t > 0$. Insbesondere kann in diesem Fall kein Element von e^{tA} bei $t > 0$ verschwinden.*

Beispiel 7.12. Betrachte die Markov-Matrix $M = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, die irreduzibel, aber nicht primitiv ist. Aufgrund der obigen Folgerung kann sie also nie in einer Markov-Halbgruppe vorkommen.

Beispiel 7.13. Nun betrachten wir die Markov-Matrix $M = \left(\begin{smallmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{smallmatrix}\right)$, die primitiv ist, denn $M^2 = \left(\begin{smallmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \end{smallmatrix}\right)$. Dennoch kann auch M nicht in einer Markov-Halbgruppe vorkommen, denn in $M = \mathrm{e}^{tA}$ wäre ein Element 0.

Aus diesen Beispielen ergibt sich, dass es eine interessante (und für die Anwendungen wichtige) Frage ist, welche Markov-Matrizen in einer Markov-Halbgruppe vorkommen können. Dies wurde zuerst von Elfving (1936) betrachtet und ist als *Darstellungsproblem* oder *Einbettungsproblem* für Markov-Matrizen bekannt.

Definition 7.14. Eine Markov-Matrix M heisst *einbettbar* (oder *darstellbar*), wenn es einen Markov-Generator Q gibt, so dass $M = \mathrm{e}^{Q}$ gilt.

Man überlegt sich sofort, dass M genau dann einbettbar ist, wenn M in einer Markov-Halbgruppe der Form $\{\mathrm{e}^{tQ} \mid t \geq 0\}$ vorkommt. Dies folgt aus der Beobachtung, dass Q genau dann ein Markov-Generator ist, wenn dies auch für tQ mit beliebigen $t > 0$ gilt. Um das Einbettungsproblem besser zu verstehen, benötigen wir auch den Logarithmus einer Matrix, der aber nicht für jede Matrix definiert ist. Dazu beachten wir

$$\mathrm{e}^{tA} \;=\; \mathbb{1} + \underbrace{(\mathrm{e}^{tA} - \mathbb{1})}_{\text{‘klein’ für } t \searrow 0}$$

und erinnern uns an die Potenzreihe von $\log(1+x)$ aus Analysis I, nämlich

$$\log(1+x) \;=\; x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots \;=\; \sum_{m \geq 1} (-1)^{m+1} \, \frac{x^m}{m} \,.$$

Diese Reihe konvergiert bekanntlich für alle $x \in \mathbb{C}$ mit $|x| < 1$. Ist $B \in \mathrm{Mat}(n, \mathbb{C})$ nun eine Matrix mit Spektralradius $\varrho(B) < 1$, so ist

$$\log(\mathbb{1} + B) \;=\; \sum_{m \geq 1} (-1)^{m+1} \, \frac{B^m}{m}$$

wohldefiniert und konvergent.

Lemma 7.15. *Ist M eine Markov-Matrix mit Spektrum $\sigma(M)$ und*

$$\varrho(M - \mathbb{1}) \;=\; \max_{\lambda \in \sigma(M)} |\lambda - 1| \;<\; 1 \,,$$

so gibt es eine relle Matrix A mit verschwindenden Zeilensummen und $M = \mathrm{e}^A$.

Beweis. Wir setzen $A = \log(\mathbb{1} + (M - \mathbb{1}))$. Da die zugehörige Potenzreihe wegen der Annahme $\varrho(M - \mathbb{1}) < 1$ konvergiert, definiert dies eine reelle Matrix. Dabei ergibt sich $\sum_j A_{ij} = 0$, denn $M - \mathbb{1}$ ist selber ein Markov-Generator und wir können die Zeilensummen (analog zu früher) anhand der Potenzreihe ausrechnen. Weiter gilt dann:

$$\mathrm{e}^A \;=\; \mathrm{e}^{\log(\mathbb{1} + (M - \mathbb{1}))} \;=\; \mathrm{e}^{\log(M)} \;=\; M \,,$$

woraus sich die Behauptung konstruktiv ergibt. □

Im letzten Beweis haben wir benutzt, dass der Logarithmus die Umkehrfunktion der Exponentialfunktion ist. Um zu sehen, dass dies auch für Matrizen so ist, erinnern wir kurz daran, dass für $|x| < 1$ folgende Rechnung (durch Einsetzen der Reihen ineinander) gilt:

$$\begin{aligned}
\mathrm{e}^{\log(1+x)} \;&=\; 1 + \big(x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 + \mathcal{O}(x^4)\big) \\
&\quad + \tfrac{1}{2}\big(x^2 - x^3 + \mathcal{O}(x^4)\big) \\
&\quad + \tfrac{1}{6}\big(x^3 + \mathcal{O}(x^4)\big) \;+\; \mathcal{O}(x^4) \\
&=\; 1 + x + \underbrace{\big(-\tfrac{1}{2} + \tfrac{1}{2}\big)}_{=0} x^2 + \underbrace{\big(\tfrac{1}{3} - \tfrac{1}{2} + \tfrac{1}{6}\big)}_{=0} x^3 + \mathcal{O}(x^4) \\
&=\; 1 + x + \mathcal{O}(x^4) \,,
\end{aligned}$$

und analog ergibt sich das für alle höheren Ordnungen, d.h.

$$\mathrm{e}^{\log(1+x)} \;=\; 1 + x + \mathcal{O}(x^n)$$

für alle $4 \leq n \in \mathbb{N}$ und damit insgesamt also $\mathrm{e}^{\log(1+x)} = 1+x$. Die Rechnung für Matrizen verläuft analog.

Nun wollen wir das Einbettungsproblem für $n = 2$ behandeln. Dazu betrachten wir eine Markov-Matrix der Form $M = \left(\begin{smallmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{smallmatrix}\right)$ mit den Parametern $0 \leq \alpha, \beta \leq 1$. Ein Eigenwert von M ist stets 1. Da die Spur der Matrix gleich der Summe der Eigenwerte ist, ergeben sich für obiges M also die zwei Eigenwerte $\lambda_1 = 1$ und $\lambda_2 = 1-\alpha-\beta$. Weiter lässt sich M als Summe zerlegen in

$$M \;=\; \mathbb{1} + \underbrace{\left(\begin{smallmatrix} -\alpha & \alpha \\ \beta & -\beta \end{smallmatrix}\right)}_{=M-\mathbb{1}},$$

wobei $M - \mathbb{1}$ ein Markov-Generator ist. Wenn nun $0 < \alpha, \beta < 1$ und $\alpha + \beta < 1$, so gilt auch $\varrho(M-\mathbb{1}) < 1$, und nach Lemma 7.15 läge es nahe, dass ein solches M darstellbar ist. Man beachte aber, dass im Lemma nur die Existenz einer Matrix A mit verschwindenden Zeilensummen attestiert wird. Wir müssen also noch klären, ob es dabei tatsächlich um einen Markov-Generator handelt. Dies kann man durch eine genaue Untersuchung der obigen Logarithmus-Reihe nachweisen (Details: Übung). Es gibt hier auch noch folgenden alternativen Zugang, der ohne den Logarithmus auskommt.

Der allgemeinste Markov-Generator für $n = 2$ lautet

$$A \;=\; \begin{pmatrix} -\alpha' & \alpha' \\ \beta' & -\beta' \end{pmatrix},$$

mit $\alpha', \beta' \geq 0$. Wir wollen jetzt die zugehörige Markov-Halbgruppe berechnen, also e^{tA} für $t \geq 0$. Falls $\alpha' = \beta' = 0$, bekommen wir $A = \mathbf{0}$ und $\mathrm{e}^{tA} \equiv \mathbb{1}$. Sei also $\alpha' + \beta' > 0$. Durch Induktion (Übung!) weist man nach, dass

$$A^n \;=\; (-1)^{n-1}\,(\alpha' + \beta')^{n-1}\,A$$

für alle $n \in \mathbb{N}$ gilt. Einsetzen in die allgemeine Formel für e^{tA} liefert dann

$$\begin{aligned}
\mathrm{e}^{tA} &= \mathbb{1} + \sum_{n\geq 1} (-1)^{n-1} (\alpha' + \beta')^{n-1} \frac{t^n}{n!} A \\
&= \mathbb{1} - \frac{1}{\alpha' + \beta'} \underbrace{\left(\sum_{n\geq 1} \frac{(-t)^n (\alpha' + \beta')^n}{n!} \right)}_{=\mathrm{e}^{-t(\alpha'+\beta')}-1} A \\
&= \mathbb{1} + \underbrace{\frac{1 - \mathrm{e}^{-t(\alpha'+\beta')}}{\alpha' + \beta'}}_{=:\varphi(t)} A \;=\; \begin{pmatrix} 1-\varphi(t)\alpha' & \varphi(t)\alpha' \\ \varphi(t)\beta' & 1-\varphi(t)\beta' \end{pmatrix}.
\end{aligned}$$

Dabei gilt $\varphi(0) = 0$ und $0 < (\alpha' + \beta')\varphi(t) = 1 - \mathrm{e}^{-t(\alpha'+\beta')} < 1$ für alle $t > 0$. Man kann sich nun überlegen, dass hierdurch genau die Matrizen aus obigem Lemma erfasst werden, wobei für festes $t > 0$ die Wahl von α' und β' jeweils eindeutig ist (Details: Übung). Damit erhalten wir den folgenden wichtigen Satz. Für weitere Überlegungen verweisen wir auf die Literatur, insbesondere auf [10].

Satz 7.16. *(Kendall)*
Die Markov-Matrix $M = \left(\begin{smallmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{smallmatrix}\right)$ *ist genau dann einbettbar, wenn* $\alpha, \beta \geq 0$ *gilt zusammen mit* $\alpha + \beta < 1$. *In diesem Fall gibt es genau einen Markov-Generator* A *mit* $M = \mathrm{e}^A$, *nämlich* $A = \log\big(\mathbb{1} + (M - \mathbb{1})\big)$.

Beweis. Falls $\alpha = \beta = 0$, ist $M = \mathbb{1} = \mathrm{e}^{\mathbf{0}}$. Sei also $\alpha + \beta > 0$. Dann ist aus unserer obigen Rechnung klar, dass wir Parameter α' und β' finden können, so dass M in der Form e^{tA} z.B. mit $t = 1$ geschrieben werden kann. Die Parameter α', β' sind dann eindeutig bestimmt durch $\alpha' = \alpha/\varphi(1)$ und $\beta' = \beta/\varphi(1)$. □

Beispiel 7.17. Wir betrachten noch einmal die Matrizen

$$M = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix} \quad \text{und} \quad M^2 = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

aus Beispiel 7.13. Nach dem Satz von Kendall ist M^2 einbettbar, M aber nicht. Dabei ist also $M^2 = \mathrm{e}^A$, und man kann dann nachrechnen (Übung), dass

$$\exp\big(\tfrac{1}{2}A\big) = \begin{pmatrix} \frac{5}{6} & \frac{1}{6} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix} =: M'$$

gilt, wobei natürlich $(M')^2 = M^2$ gilt. Wir sehen hier also das Phänomen, dass eine Matrix verschiedene Markov-Quadratwurzeln haben kann. Man kann sich rasch überlegen, dass es in diesem Fall keine weiteren gibt.

Bevor wir uns weiter mit dem Einbettungsproblem befassen, werfen wir noch einen kleinen Blick auf die Gleichgewichte von (eingebetteten) Markov-Matrizen. Dabei beschränken wir uns hier auf den Fall, dass es nur ein Gleichgewicht gibt. In diesem Fall ist die *stationäre Verteilung* eine wichtige Kenngröße des Markov-Prozesses, insbesondere für sein asymptotisches Verhalten. Mathematisch ist die stationäre Verteilung der (statistisch normierte) Links-Eigenvektor von einer Markov-Matrix (und damit, wie wir noch sehen werden, sogar von allen Matrizen) aus der Halbgruppe zum Eigenwert 1.

Beispiel 7.18. Für den Markov-Generator $A = \left(\begin{smallmatrix} -\alpha & \alpha \\ \beta & -\beta \end{smallmatrix}\right)$ mit $\alpha+\beta > 0$ ist die stationäre Verteilung durch

$$\left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)$$

gegeben, wie man leicht nachrechnen kann.

Betrachten wir nun zunächst $\varphi(t)$ mit $\alpha + \beta > 0$ für $t \to \infty$, also

$$\varphi(t) = \frac{1 - \mathrm{e}^{-t(\alpha+\beta)}}{\alpha+\beta} \xrightarrow{t\to\infty} \frac{1}{\alpha+\beta},$$

so folgt daraus (wieder für $\alpha + \beta > 0$)

$$M(t) = \begin{pmatrix} 1-\varphi(t)\alpha & \varphi(t)\alpha \\ \varphi(t)\beta & 1-\varphi(t)\beta \end{pmatrix} \xrightarrow{t\to\infty} \begin{pmatrix} 1-\frac{\alpha}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\beta}{\alpha+\beta} & 1-\frac{\beta}{\alpha+\beta} \end{pmatrix} = \begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{pmatrix}.$$

Dies ist das erwartete asymptotische Verhalten für eine primitive Markov-Matrix.

Nun wollen wir die stationäre Verteilung einer irreduziblen Markov-Halbgruppe etwas allgemeiner ansehen.

Proposition 7.19. *Sei A ein irreduzibler Markov-Generator, mit Dimension ≥ 2. Dann besitzt die Markov-Halbgruppe $\{\mathrm{e}^{tA} \mid t \geq 0\}$ einen simultanen PF-Eigenvektor. Dieser liegt dann im Kern von A, der wegen der Irreduzibilität eindimensional ist.*

Beweis. Es ist e^{tA} primitiv für alle $t > 0$. Sei nun ein $t > 0$ fixiert. Nach den Satz von Perron–Frobenius existiert nun ein eindeutiger, strikt positiver Wahrscheinlichkeitsvektor p mit $p\mathrm{e}^{tA} = p$. Weiter gilt für ein beliebiges $s \geq 0$

$$(p\mathrm{e}^{sA})\mathrm{e}^{tA} = p\mathrm{e}^{(t+s)A} = (p\mathrm{e}^{tA})\mathrm{e}^{sA} = p\mathrm{e}^{sA},$$

also ist auch $p\mathrm{e}^{sA}$ ein Eigenvektor von e^{tA} zum Eigenwert 1. Aufgrund der Markov-Eigenschaft von e^{tA} ist dies ebenfalls ein Wahrscheinlichkeitsvektor, also gilt $p\mathrm{e}^{sA} = p$ wegen der Eindeutigkeit des normierten Eigenvektors. Somit ist p simultaner Eigenvektor für die gesamte Markov-Halbgruppe.

Durch Differenzieren (nach t) bekommen wir nun

$$p\mathrm{e}^{tA}A = 0,$$

da p ja konstant ist. Einsetzen von $t = 0$ liefert $pA = 0$, also liegt p im Kern von A (genauer sollten wir sagen, im Links-Kern), und dieser ist eindimensional, weil ein linear unabhängiges Element im Kern die Existenz eines weiteren Eigenvektors von e^{tA} zum Eigenwert 1 bedeuten würde, was ja nicht möglich ist. □

Nun wollen wir noch eine allgemeine Formel angeben, um diesen Gleichgewichtszustand auszurechnen. Für einen Markov-Generator $A = (A_{ij})_{1\leq i,j\leq n}$ setzen wir

$$v_i = \sum_{T\in\mathcal{R}^{(i)}} \prod_{\langle k,\ell\rangle\in T} A_{\ell k},$$

wobei $\mathcal{R}^{(i)}$ die Menge der bezifferten Bäume mit n Knoten und Wurzel (i) sind und $\langle k,\ell\rangle$ eine gerichtete Kante von k nach ℓ bezeichnet. Man beachte, dass durch die Auszeichnung eines Knotens als Wurzel eine natürliche Richtung auf jeder Kante im Baum entsteht. Damit können wir nun folgendes Resultat formulieren.

Satz 7.20. *Ist A ein irreduzibler Markov-Generator, so gilt:*

(1) *$vA = 0$ mit $v = (v_1, \ldots, v_n)$ und v_i wie oben definiert. Dabei sind alle $v_i > 0$.*

(2) $p_i = \frac{v_i}{\|v\|_1}$ *definiert einen Wahrscheinlichkeitsvektor. Dieser ist die stationäre Verteilung von der Markov-Halbgruppe.*

(3) $v_i = \det(-A_{(ii)})$. *Dabei entsteht die Matrix* $A_{(ii)}$ *aus* A *durch Entfernen von Zeile* i *und Spalte* i.

Beweis-Bausteine. Die erste Aussage folgt aus dem Matrix-Baum-Satz, der auf Kirchhoff (1859) zurückgeht, zusammen mit dem Satz von Perron–Frobenius. Die zweite Aussage ist klar nach unserer obigen Rechnung für Proposition 7.19. Die dritte Aussage folgt aus der Cauchy–Binet-Formel, die eine Verallgemeinerung der Laplace-Entwicklung ist. □

Weiterführende Literatur zu diesem Thema ist [11].

Nun wollen wir zum Einbettungsproblem zurückkehren und noch ein wenig den Fall $n > 2$ betrachten, der deutlich komplizierter ist. In der Phylogenie benötigt man bekanntlich $n = 4$, weshalb diese Erweiterung notwendig ist.

Lemma 7.21. *Sei* M *eine* $n\times n$ *Markov-Matrix mit* $\varrho(M - \mathbb{1}) < 1$. *Dann existiert ein* $B \in \mathrm{Mat}(n, \mathbb{R})$ *mit* $M = \mathrm{e}^B$ *und* $\sum_j B_{ij} = 0$ *für alle* $1 \le i \le n$. *Für* $n = 2$ *ist* B *dann stets ein Markov-Generator.*

Beweis. $B = \log\big(\mathbb{1} + (M - \mathbb{1})\big)$ ist wegen $\varrho(M - \mathbb{1}) < 1$ eine konvergente Reihe, und es gilt $\mathrm{e}^B = \exp\big(\log(\mathbb{1} + (M - \mathbb{1}))\big) = M$. Dabei ist $A := M - \mathbb{1}$ ein Markov-Generator, also $A_{ij} \ge 0$ für alle $i \ne j$ und $\sum_j A_{ij} = 0$ für alle $1 \le i \le n$.

Damit bekommen wir aber auch (für i beliebig, aber fest):

$$\sum_{j=1}^{n} B_{ij} = \sum_{j=1}^{n} \sum_{k\ge 1} \frac{(-1)^{k-1}}{k} \big(A^k\big)_{ij} = \sum_{k\ge 1} \frac{(-1)^{k-1}}{k} \underbrace{\sum_{\ell=1}^{n} \big(A^{k-1}\big)_{i\ell} \underbrace{\Big(\sum_j A_{\ell j}\Big)}_{=0}}_{=0} = 0.$$

Im Allgemeinen braucht $B_{ij} \ge 0$ für $i \ne j$ jedoch *nicht* zu gelten. Bei $n = 2$ stimmt das aber immer, wie wir aus der Herleitung des Satzes von Kendall bereits wissen. □

Beispiel 7.22. Sei $\beta > 0$ und $B = \left(\begin{smallmatrix} 0 & 0 \\ \beta & -\beta \end{smallmatrix}\right)$. Dann ist

$$\mathrm{e}^{tB} = \begin{pmatrix} 1 & 0 \\ 1 - \mathrm{e}^{-t\beta} & \mathrm{e}^{-t\beta} \end{pmatrix} = \begin{cases} \mathbb{1}, & \text{falls } t = 0, \\ \left(\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right), & \text{im Grenzfall } t \to \infty. \end{cases}$$

Die 0 in der oberen rechten Ecke illustriert ein allgemeines Phänomen: Es gilt $\big(\mathrm{e}^{tB}\big)_{ij} = 0$ für $i \ne j$ und ein $t > 0$ genau dann, wenn dies für *alle* $t > 0$ stimmt. In dem Fall steht die 0 auch bereits an der entsprechenden Stelle im Generator.

Folgerung 7.23. *Ist* B *ein Markov-Generator, so gilt:*

$$\mathrm{e}^{tB} \textit{ ist irreduzibel} \iff \mathrm{e}^{tB} \textit{ ist primitiv} \iff \mathrm{e}^{tB} \textit{ ist total positiv}.$$

Wir sehen hieraus, dass der Zusammenhang zwischen den verschiedenen Irreduzibilitätsbegriffen bei Matrizen aus Markov-Halbgruppen deutlich einfacher ist als bei allgemeinen Markov-Matrizen.

Beispiel 7.24. Betrachten wir die Matrix $B = 2\pi \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, die kein Generator ist. Per Induktion sieht man rasch, dass $B^{2m} = (-1)^m (2\pi)^{2m} \mathbb{1}$ und $B^{2m+1} = (-1)^m (2\pi)^{2m} B$ gilt, mit $m \in \mathbb{N}_0$. Damit kann man nun e^B ausrechnen,

$$\begin{aligned} \mathrm{e}^B &= \sum_{m\geq 0} \frac{(-1)^m (2\pi)^{2m}}{(2m)!}\, \mathbb{1} + \sum_{m\geq 0} \frac{(-1)^m (2\pi)^{2m+1}}{(2m+1)!} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \\ &= \cos(2\pi)\, \mathbb{1} + \sin(2\pi) \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \mathbb{1}, \end{aligned}$$

was nichts anderes als eine matrixwertige Version der Relation $\mathrm{e}^{2\pi \mathrm{i}} = 1$ ist. Also ist dieses B ebenfalls ein 'Logarithmus' von $\mathbb{1}$, aber mit nicht-verschwindenden Zeilensummen.

Man kann mit einem einfachen Argument (Übung) zeigen, dass es bei $n = 2$ ausser $\mathbf{0}$ keine weitere Lösung von $\mathbb{1} = \mathrm{e}^B$ mit B reell und verschwindenden Zeilensummen gibt. Bereits bei $n = 3$ wird auch dies schon komplizierter.

Wenn eine Markov-Matrix M einbettbar ist, ergeben sich sofort mehrere Konsequenzen, die somit auch notwendige Eigenschaften für eine Einbettbarkeit sind, nämlich:

1. Nach dem *spektralen Abbildungssatz* (nachschlagen !) für die Eigenwerte von M und B in $M = \mathrm{e}^B$ muss gelten

   $$\sigma(M) = \mathrm{e}^{\sigma(B)} = \{\mathrm{e}^\lambda \mid \lambda \in \sigma(B)\},$$

   wobei dies auch für mehrfache Eigenwerte von M stimmen muss.

2. Es gilt $0 < \det(M) \leq 1$, also ist 0 niemals Eigenwert einer einbettbaren Markov-Matrix. Ausserdem gilt $\det(M) = 1$ nur für $M = \mathbb{1}$ (diese Eigenschaften folgen aus der Formel $\det(\mathrm{e}^B) = \mathrm{e}^{\mathrm{tr}(B)}$ zusammen mit $\mathrm{tr}(B) \leq 0$).

3. Falls $1 \neq \lambda \in \sigma(M)$, so muss $|\lambda| < 1$ gelten (Satz von Elfving).

4. $M_{ij} > 0$ und $M_{jk} > 0$ implizieren $M_{ik} > 0$.

5. M ist entweder reduzibel oder total positiv (und dann auch primitiv).

6. Ist $\lambda \in \sigma(M)$ reell und negativ, so ist die algebraische Vielfachheit von λ gerade.

Leider sind auch alle sechs Eigenschaften zusammen genommen immer noch nicht hinreichend für die Einbettbarkeit. Dabei hat die letzte Bedingung eine besondere Bedeutung. Eine etwas einfachere Frage ist, wann eine relle Matrix M einen reellen Logarithmus besitzt, also $M = \mathrm{e}^R$ mit R reell gilt. Die Antwort gibt das folgende Kriterium.

Satz 7.25 (Culver, 1966)**.** *Für $S \in \mathrm{GL}(n,\mathbb{R})$ besitzt $S = \mathrm{e}^R$ eine Lösung $R \in \mathrm{Mat}(n,\mathbb{R})$ genau dann wenn jeder elementare Jordan-Block zu einem negativen Eigenwert von S, in der komplexen Jordan-Normalform von S, geradzahlig oft vorkommt.*

Ist S diagonalisierbar, vereinfacht sich dies zu der Bedingung, dass jeder negative Eigenwert von S gerade algebraische Vielfachheit besitzen muss.

Beweis-Skizze. Ist $S = \mathrm{e}^R$ und T invertierbar, so gilt bekanntlich

$$TST^{-1} = T\mathrm{e}^R T^{-1} = \exp\bigl(TRT^{-1}\bigr).$$

Nun wähle man $T \in \mathrm{GL}(n,\mathbb{C})$ so, dass TST^{-1} in (komplexer) Jordan-Normalform vorliegt. Ist λ ein positiver Eigenwert von S, kann man den zugehörigen Block von TRT^{-1} ausrechnen. Ist λ ein genuin komplexer Eigenwert, also $\lambda \in \mathbb{C} \setminus \mathbb{R}$, muss es ein komplex konjugiertes Paar von Jordan-Blöcken geben, weil S ja reell ist und somit auch $\overline{\lambda}$ Eigenwert von S ist, mit denselben algebraischen und geometrischen Vielfachheiten wie λ. Dann kann man wieder eine Lösung ausrechnen, wenn man sich an

$$\mathrm{diag}(\mathrm{i},-\mathrm{i}) \sim \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

erinnert, was wir oben (in Beispiel 7.24) schon einmal benutzt haben.

Der schwierige Fall ist ein Eigenwert $\lambda < 0$. Ist die Multiplizität der zugehörigen Jordan-Blöcke gleicher Größe gerade, kann man sich an Euler's Relation $\mathrm{e}^{\pi\mathrm{i}} = -1$ erinnern, die in Matrix-Form gerade

$$\exp\left(\pi \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\right) = -\mathbb{1}$$

bedeutet, und wieder eine Lösung ausrechnen. Schließlich muss man noch nachweisen, dass es bei ungerader Vielfachheit nicht geht, weil immer ein Block übrigbleibt. □

Wegen der Struktur mit der Ähnlichkeitstransformation T kann man nun auch sehen, dass $\mathbb{1}$ für $n \geq 2$ sogar kontinuierlich viele reelle Logarithmen besitzt: Jedes T vertauscht mit $\mathbb{1}$, aber nicht unbedingt mit R; man vergleiche das mit Beispiel 7.24.

Die *Vertauschungseigenschaften* von Matrizen sind also (wieder einmal) entscheidend. Ist $M = \mathrm{e}^R$, so muss natürlich $[M,R] = \mathbf{0}$ gelten, also auch:

$$R \in \mathrm{cent}(M) = \{T \in \mathrm{Mat}(n,\mathbb{R}) \mid [M,T] = \mathbf{0}\}.$$

Schreiben wir wieder $M = \mathbb{1} + A$ mit $A = M - \mathbb{1}$, so gilt auch $\mathrm{cent}(M) = \mathrm{cent}(A)$. Dies hilft aufgrund des folgenden Satzes aus der (gehobenen) linearen Algebra.

Satz 7.26 (Frobenius)**.** *Sei $B \in \mathrm{Mat}(n,\mathbb{R})$. Dann sind äquivalent:*

(1) *Das charakteristische Polynom von B, also $\det(B - x\mathbb{1})$, stimmt mit dem Minimalpolynom von B überein.*

(2) *Es gilt $\dim(V_\lambda) = 1$ für jedes $\lambda \in \sigma(B)$, wobei V_λ der zugehörige Eigenraum ist.*

(3) *Die Matrix B ist* zyklisch. *Dies bedeutet: Es gibt einen Vektor $u \in \mathbb{R}^n$, so dass $\{u, Bu, B^2u, \ldots, B^{n-1}u\}$ eine Basis des $\mathbb{R}^n$ ist.*

(4) *Der Matrix-Ring* $\operatorname{cent}(B)$ *ist abelsch.*

(5) *Es gilt:* $\operatorname{cent}(B) = \mathbb{R}[B]$*; dieser Ring besteht also aus allen Polynomen in B mit reellen Koeffizienten.*

Einfache Matrizen, also solche mit paarweise verschiedenen Eigenwerten, sind zyklisch. Somit ist die Eigenschaft, zyklisch zu sein, auch eine *generische* Eigenschaft, und zwar im maßtheoretischen Sinne: Fast alle Matrizen sind zyklisch. Darum wollen wir nun diese Klasse noch etwas genauer ansehen. Dabei ist klar, dass $M = \mathbb{1} + A$ genau dann zyklisch ist, wenn das auch für A der Fall ist, wobei natürlich $\mathbb{R}[M] = \mathbb{R}[A]$ gilt.

Uns interessieren nur reelle Logarithmen von M mit Zeilensummen 0. Eine allgemeine Matrix $T \in \mathbb{R}[A]$ ist nun von der Form

$$T \;=\; a_0\mathbb{1} + a_1A + \ldots + a_mA^m$$

für ein $m \in \mathbb{N}_0$ und geeignete relle Koeffizienten a_i. Wegen $A = M - \mathbb{1}$ wissen wir aber schon, dass $\sum_j T_{ij} = a_0$ gilt, und zwar für alle i. Eine solche Matrix T hat also verschwindende Zeilensummen genau für $a_0 = 0$. Daher setzen wir jetzt

$$\operatorname{alg}(A) \;:=\; \langle A^m \mid m \in \mathbb{N}\rangle_{\mathbb{R}} \;=\; \langle A, A^2, \ldots, A^{n-1}\rangle_{\mathbb{R}} \;\subsetneq\; \mathbb{R}[A],$$

wobei der zweite Schritt eine Konsequenz des Satzes von Cayley–Hamilton (nachschlagen!) ist. Für zyklische A gilt $\dim\big(\operatorname{alg}(A)\big) = n - 1$. Alle Elemente von $\operatorname{alg}(A)$ haben verschwindende Zeilensummen, sind aber i.A. keine Markov-Generatoren. Für zyklische Matrizen stellt sich dies nun gar nicht als Einschränkung heraus:

Lemma 7.27. *Ist M eine zyklische Markov-Matrix, so ist* 1 *ein einfacher Eigenwert von M. In diesem Fall besitzt jeder reelle Logarithmus von M automatisch Zeilensummen* 0.

Beweis. Für jede Markov-Matrix stimmt die algebraische Vielfachheit des Eigenwerts $\lambda = 1$ mit seine geometrischen Vielfachheit überein (Beweis per Jordan-Normalform, vgl. [2, Satz 13.10] für Details).

Ist die Matrix M zyklisch, muss 1 also ein einfacher Eigenwert sein, mit rechts-Eigenvektor $(1, \ldots, 1)^T$. Ist $M = \mathrm{e}^R$ mit R reell, gilt $R \in \mathbb{R}[A]$ mit $A = M - \mathbb{1}$. Ausserdem muss dann (nach dem spektralen Abbildungssatz) 0 ein einfacher Eigenwert von R sein, da dies die einzige relle Lösung von $\mathrm{e}^x = 1$ ist.

Sei nun u der zugehörige Eigenvektor von R zum Eigenwert 0. Dann gilt aber auch $Mu = \mathrm{e}^R u = u$, also folgt $u = \alpha(1, \ldots, 1)^T$ für ein $\alpha \neq 0$. Dies bedeutet aber gerade, dass in R alle Zeilensummen verschwinden. □

Aus dem Satz von Frobenius zusammen mit unserer obigen Überlegung zu den Vertauschungsrelationen ergibt sich nun unmittelbar die folgende Eigenschaft.

Folgerung 7.28. *Sei M eine zyklische $n\times n$ Markov-Matrix mit* $\det(M) \neq 0$. *Besitzt M einen reellen Logarithmus, $M = \mathrm{e}^R$, so ist $R \in \operatorname{alg}(A)$, mit $A = M - \mathbb{1}$.*

Damit können wir nun das Einbettungsproblem für eine wichtige Klasse von Matrizen voranbringen.

Satz 7.29. *Sei M eine zyklische Markov-Matrix mit reellem Spektrum. Dann besitzt M genau dann einen reellen Logarithmus, wenn alle Eigenwerte von M positiv sind.*

In diesem Fall, $M = \mathrm{e}^R$, ist R eindeutig, und erfüllt $R \in \mathrm{alg}(A)$ mit $A = M - \mathbb{1}$. Weiter ist M einbettbar genau wenn dieses R ein Markov-Generator ist.

Beweis-Skizze. Die erste Behauptung folgt auch dem Satz von Culver, und $R \in \mathrm{alg}(A)$ ist eine Konsequenz der voranstehenden Folgerung.

Die Eindeutigkeit bekommt man aus dem linearen Gleichungssystem, das man mittels des spektralen Abbildungssatz aus $\sigma(M) = \exp\big(\sigma(R)\big)$ ableitet, unter korrekter Verwendung der Jordan-Normalformen von M und R.

Die Generatoreigenschaft der Lösung bleibt offen, aber R ist der einzige Kandidat. □

Aus dem Beweis, der hier nicht im Detail geführt werden kann, ergibt sich noch, dass der Satz *konstruktiv* ist. Wir wissen, dass $R = \alpha_1 A + \ldots + \alpha_{n-1} A^{n-1}$ mit Koeffizienten $\alpha_i \in \mathbb{R}$ gelten muss, und diese Koeffizienten kann man aus dem im Beweis erwähnten linearen Gleichungssystem explizit berechnen. Die Generatoreigenschaft von R ist dann einfach nachzuweisen, indem man $R_{ij} \geq 0$ für alle $i \neq j$ überprüft.

Folgerung 7.30. *Sei M wie im vorigen Satz, mit positiven Eigenwerten. Dann ist die relle Matrix $R = \log(\mathbb{1} + A)$ der einzige Kandidat für die Einbettbarkeit von M.*

Beweis. Bei $\sigma(M) \subset \mathbb{R}_+$ gilt automatisch $\sigma(M) \subset (0, 1]$ nach dem Satz von Elfving. Dann ist aber der Spektralradius von A kleiner als 1, und $\log(\mathbb{1}+A)$ konvergiert. Aufgrund der Darstellung als Reihe in A ist dies eine relle Matrix, die auch in $\mathrm{alg}(A)$ liegt, und somit der einzige Kandidat für den gesuchten Markov-Generator ist. □

Nun bleibt noch festzuhalten, dass $\sigma(M)$ i.A. nicht reell sein muss. Beim Vorliegen von komplexen Eigenwerten (und dann von komplex-konjugierten Paaren) können Mehrdeutigkeiten auftreten, es kann also mehrere Lösungen für das Einbettungsproblem geben. Natürlich muss M nicht zyklisch sein. Wenn das passiert, wird das Einbettungsproblem noch einmal deutlich komplexer, und man benötigt Methoden aus der algebraischen Geometrie zu seiner Lösung. Bis $n = 4$ sind die möglichen Fälle inzwischen komplett studiert, darüberhinaus bleiben immer noch offene Fragen.

Literatur

[1] L.C. Evans, *Partial Differential Equations*, AMS, Providence (1998).

[2] F.R. Gantmacher, *Matrizentheorie*, Springer, Berlin (1986).

[3] H. Heuser, *Funktionalanalysis*, Teubner, Stuttgart (1992).

[4] J.R. Norris, *Markov Chains*, Cambridge University Press, Cambridge (1997).

[5] M. Pinsky, *Introduction to Fourier Analysis and Wavelets*, AMS, Providence (2009).

[6] R. Plato, *Numerische Mathematik kompakt*, 3. Aufl., Vieweg, Wiesbaden (2006).

[7] W. Rudin, *Real and Complex Analysis*, 3. Aufl., McGraw Hill, New York (1987).

[8] K.-J. Engel und R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations*, Springer, Berlin (2000).

[9] J. von zur Gathen und J. Gerhard, *Modern Computer Algebra*, Cambridge University Press, Cambridge (1999).

[10] J.F.C. Kingman, The imbedding problem for finite Markov chains, *Z. Wahrscheinlichkeitsth. verw. Geb.* **1** (1962) 14–24.

[11] M.I. Freidlin und A.D. Wentzell, *Random Perturbations of Dynamical Systems*, 2. Auflage, Springer, New York (1998).

[12] W. Walter Gewöhnliche Differentialgleichungen, 7. Auflage, Springer, Berlin (2000).