Mathematische Methoden für Biowissenschaften III
Fourieranalysis und Anwendungen

Skript WS 2009/10

PD Dr Dirk Frettlöh
Fakultät Mathematik
Universität Bielefeld

February 26, 2010

Contents

1 Die schwingende Saite **4**

2 Fourierreihen **6**

2.1 Approximation von periodischen Funktionen 7

2.2 Gibbssches Phänomen 9

3 Konvergenz von FRn **10**

4 Hilberträume **12**

4.1 ON-Systeme 17

4.2 Die Gaußsche Approximationsaufgabe 18

5 Fouriertransformation (FT) **21**

5.1 Poissons Summenformel (PSF) 27

6 DFT **29**

6.1 Trigonometrische Interpolation 31

6.2 Bildkompression 35

6.3 FFT 37

6.4 Schnelle Multiplikation 40

7 FT und DGL **43**

7.1 Die Wärmegleichung 45

7.2 Die mehrdimensionale Wellengleichung 46

8 Elementares zu Funktionalgleichungen **50**

8.1 Die Matrixexponentialfunktion 52

8.2 Lineare DGL mit konstanten Koeffizienten 55

Vorab:

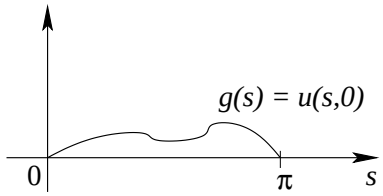
Dieses Skript entstand im Rahmen der Vorlesung Mathematische Methoden für Biowissenschaften III an der Uni Bielefeld. Darin soll laut Modulbeschreibung hauptsächlich Fourieranalysis behandelt werden, sowie stochastische Prozesse. In Absprache mit den Modulverantwortlichen wurden stochastische Prozesse weggelassen, um Zeit für Anwendung von Fourieranalysis (Bildkompression, schnelle Multiplikation, Behandlung von Differentialgleichungen mit Fouriermethoden...) sowie ein kurzes Kapitel zu Funktionalgleichungen zu geben.

Dank:

Bedanken möchte ich mich an dieser Stelle bei Uwe Schwerdtfeger für viele hilfreiche Tipps und tatkräftige Unterstützung; sowie bei den fünf tapferen Teilnehmern der Vorlesung, die mich auf etliche Fehler im Skript hinwiesen und es so entscheidend verbesserten.

1 Die schwingende Saite

Betrachte eine Saite, eingespannt zwischen 0 und π:



Zur Zeit $t = 0$ befinde sich die Saite in Lage $g(s)$. Dabei ist g eine Funktion $g : [0,\pi] \to \mathbb{R}$, und $g(s)$ beschreibt die Auslenkung der Saite nach oben (positiv) oder unten (negativ) an der Stelle s. Um den Zustand der Saite zu beschreiben, brauchen wir auch noch den Impuls (=die Geschwindigkeit) zur Zeit 0 (positiv: Saite schwingt gerade nach oben, negativ: Saite schwingt gerade nach unten). Das bezeichnen wir mit h, also $h : [0,\pi] \to \mathbb{R}$. Die Funktionen g und h sollen "sinnvoll" sein. Etwa soll g keine Sprungstellen haben (dann wäre die Saite ja an dieser Stelle gerissen). Daher nehmen wir erstmal an, g und h seien stetig.

Die Funktion $u(s,t) : [0,\pi] \times \mathbb{R} \to \mathbb{R}$ soll dann die Lage der Saite an der Stelle s ($0 \le s \le \pi$) zur Zeit t ($t \ge 0$) beschreiben. Auf Grund der Newtonschen Mechanik (s. Physikbuch) gilt dann die partielle Differentialgleichung (PDE)

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial s^2} \quad (c \in \mathbb{R}) \tag{1.1}$$

In Worten: die Funktion u, zweimal nach t abgeleitet, ist gleich c^2 mal der Funktion u, zweimal nach s abgeleitet.

Mögliche Lösungen (aber nicht unbedingt alle!) liefert der Separationsansatz (siehe Differentialgleichungsbuch). Dazu nehmen wir an dass $u(s,t)$ von der Form $v(s)w(t)$ ist (was ja nicht der Fall sein muss, aber wir nehmen's mal an, weil's klappen wird). Dann gilt:

$$(1.1) \Rightarrow v(s)w''(t) = c^2 v''(s)w(t) \tag{1.2}$$

$$\Rightarrow \frac{w''(t)}{w(t)} = c^2 \frac{v''(s)}{v(s)} \tag{1.3}$$

$$\Rightarrow \frac{w''(t)}{w(t)} = -ac^2 \quad \wedge \quad \frac{v''(s)}{v(s)} = -a \tag{1.4}$$

$$\Rightarrow w''(t) + ac^2 w(t) = 0 \quad \wedge \quad v''(s) + av(s) = 0 \tag{1.5}$$

für eine Konstante $a \in \mathbb{R}$, denn: in der zweiten Gleichung hängt die linke Seite von t ab, die rechte nicht. D.h., egal wie ich t ändere, der Wert der linken Seite bleibt gleich der rechten, also konstant. Dito für die rechte Seite bzgl. s. Also sind beide Seiten konstant (sagen wir, die rechte ist gleich $-a$), und die linke Seite ist genau c^2 mal die rechte. (Warum wir hier $-a$ statt a nahmen: dann sparen wir uns im Folgenden ein Minuszeichen.)

Die allgemeine Lösung der zwei Differentialgleichungen (1.5) ist nun einfach (raten, oder DGL-Buch, oder [WIK]):

$$w(t) = C\sin(c\sqrt{a}t) + \tilde{C}\cos(c\sqrt{a}t) \quad (C,\tilde{C} \in \mathbb{R}) \tag{1.6}$$

$$v(s) = D\sin(\sqrt{a}s) + \tilde{D}\cos(\sqrt{a}s) \quad (D,\tilde{D} \in \mathbb{R}) \tag{1.7}$$

Obacht, dies liefert nicht alle Lösungen des ursprünglichen Problems (1.1)! Aber diese Lösungen reichen uns erst mal. Wir interessieren uns überdies nur für reelle Lösungen, also wollen wir $a \geq 0$. Wir müssen nun außerdem noch weitere Bedingungen erfüllen, unter anderem soll die Saite ja zu jedem Zeitpunkt t in $s = 0$ und $s = \pi$ in Position 0 festsitzen, also:

$$v(0) = 0 \quad \wedge \quad v(\pi) = 0 \tag{1.8}$$

Also folgt aus (1.7) mit $v(0) = 0$: $\tilde{D} = 0$, und dann, wegen $v(\pi) = 0$, entweder $D = 0$, dass interessiert uns aber nicht, denn dann würde die Saite immer still stehen (konstante Lösung 0). Also muss gelten $\sin(\sqrt{a}\pi) = 0$, und das heißt: $\sqrt{a} \in \mathbb{N}$! Also hat das Gleichungssystem (1.6) & (1.7) & (1.8) nur Lösungen für bestimmte Werte von a, nämlich für $a = n^2, n \in \mathbb{N}$. (Diese heißen auch "Eigenwerte" des Gl.-systems.) Alle Lösungen von (1.6) & (1.7) & (1.8) sind dann diese:

$$v_n(s) = D_n \sin(ns) \tag{1.9}$$

$$w_n(t) = C_n \sin(cnt) + \tilde{C}_n \cos(cnt) \tag{1.10}$$

Die Lösungen (sie heißen auch "Eigenfunktionen" des Gl.-systems) lassen sich also "nummerieren", mit $n = 0, 1, 2, \ldots$. Die entsprechenden Lösungen von (1.1) mit den Anfangsbedingungen (1.8) sind somit

$$\boxed{u_n(s,t) = \sin(ns)\big(B_n \sin(cnt) + A_n \cos(cnt)\big) \qquad (D)}$$

Der Einfachheit halber setzen wir $A_n = \tilde{C}_n D_n, B_n = C_n D_n$, dann sparen wir uns eine Konstante in der Formel.

Aber: das ursprüngliche Problem enthielt noch g und h! Wir suchen also jene Lösungen u_n, die auch noch

$$u(s,0) = g(s) \quad \wedge \quad \frac{\partial u}{\partial t}(s,0) = h(s) \tag{1.11}$$

erfüllen. Die obigen u_n tun das sehr, sehr selten, egal, wie man A_n, B_n wählt. TRICK: Setze

$$u(s,t) = \sum_{n=1}^{\infty} u_n(s,t).$$

Falls das konvergiert, und je zweimal nach s und nach t diff-bar ist, dann ist das eine Lösung für das ursprüngliche Problem (denn jeder einzelne Summand ist ja eine). Die Frage nach einer Lösung für das ursprüngliche Problem haben wir also reduziert auf die Frage: Können wir A_n, B_n in (D) so wählen, dass

$$u(s,0) = \sum_{n=1}^{\infty} \sin(ns) = g(s) \quad \text{und} \tag{1.12}$$

$$\frac{\partial u}{\partial t}(s,0) = \sum_{n=1}^{\infty} cnB_n \sin(ns) = h(s) \quad \text{gilt?} \tag{1.13}$$

$$\tag{1.14}$$

Die Frage nach Konvergenz, allgmein und gegen g bzw h, hat Hunderte von Jahren im Raum gestanden und viele Zweige der Mathematik befruchtet. Sie ist knifflig, daher verschieben wir sie auf später. Das andere Problem, das Bestimmen der A_n und B_n, konnte bereits Euler (s. [WIK]).

2 Fourierreihen

Betrachten wir das Problem in etwas allgemeinerer Form. Gegeben eine Funktion $f : [-\pi, \pi] \to \mathbb{R}$. Gesucht a_n, b_n so dass gilt:

$$f(s) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \big(a_n \cos(ns) + b_n \sin(ns)\big) \tag{2.1}$$

Diese Reihe sei überdies gleichmäßig konvergent (s. unten oder [WIK]). (Warum wir hier $\frac{a_0}{2}$ schreiben und nicht a_0 wird später deutlich.) TRICK: Es gilt

$$\int_{-\pi}^{\pi} \cos(nt)\sin(mt)dt = 0 \quad (m, n \in \mathbb{N}_0) \tag{2.2}$$

$$\int_{-\pi}^{\pi} \cos(nt)\cos(mt)dt = \begin{cases} 0 & : m \neq n \in \mathbb{N}_0 \\ \pi & : n = m \in \mathbb{N} \end{cases} \tag{2.3}$$

$$\int_{-\pi}^{\pi} \sin(nt)\sin(mt)dt = \begin{cases} 0 & : m \neq n \in \mathbb{N}_0 \\ \pi & : n = m \in \mathbb{N} \end{cases} \tag{2.4}$$

(Beweis: Übungsaufgabe 2, Blatt 1). Damit erhalten wir aus (2.1):

$$f(s)\cos(ms) = \frac{a_0}{2}\cos(ms) + \sum_{n=1}^{\infty} \big(a_n \cos(ns)\cos(ms) + b_n \sin(ns)\cos(ms)\big) \tag{2.5}$$

$$\Rightarrow \int_{-\pi}^{\pi} f(s)\cos(ms)ds = \int_{-\pi}^{\pi} \frac{a_0}{2}\cos(ms) + \sum_{n=1}^{\infty} \big(a_n \cos(ns)\cos(ms) + b_n \sin(ns)\cos(ms)\big)ds \tag{2.6}$$

$$= \frac{a_0}{2}\int_{-\pi}^{\pi} \cos(ms)ds + \sum_{n=1}^{\infty} \big(a_n \int_{-\pi}^{\pi} \cos(ns)\cos(ms)ds + b_n \int_{-\pi}^{\pi} \sin(ns)\cos(ms)ds\big) \tag{2.7}$$

Hier brauchen wir gleichmäßige Konvergenz: Dann ist es erlaubt, Summe und Integral zu vertauschen. Das Schöne ist nun: Wegen (2.2), (2.3), (2.4) werden für fast alle Werte von m die Summanden zu Null. Die Integrale, die beides, also sowohl sin- als auch cos-Terme, enthalten, werden alle zu Null, von den anderen überleben nur die mit $m = n$. Also erhalten wir für $m = 0$:

$$\int_{-\pi}^{\pi} f(s)\cos(0)ds = \int_{-\pi}^{\pi} \frac{a_0}{2}\cos(0)ds = \frac{a_0}{2}\int_{-\pi}^{\pi} 1ds = a_0\pi,$$

und für $m \geq 1$ (mit $n = m$):

$$\int_{-\pi}^{\pi} f(s)\cos(ns)ds = \int_{-\pi}^{\pi} a_n \cos(ns)\cos(ns)ds = a_n\pi.$$

Damit haben wir eine einfache Darstellung für die a_n. Mit einer komplett analogen Rechnung (Mult. mit $\sin(ms)$ statt $\cos(ms)$) erhalten wir die b_n, und insgesamt:

$$\boxed{a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(s)\cos(ns)ds \quad (n \in \mathbb{N}_0) \qquad b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(s)\sin(ns)ds \quad (n \in \mathbb{N})}$$

Definition 2.1. Sei $f : [-\pi, \pi] \to \mathbb{R}$. Die a_n, b_n oben heißen *Fourierkoeffizienten* von f, die Reihe

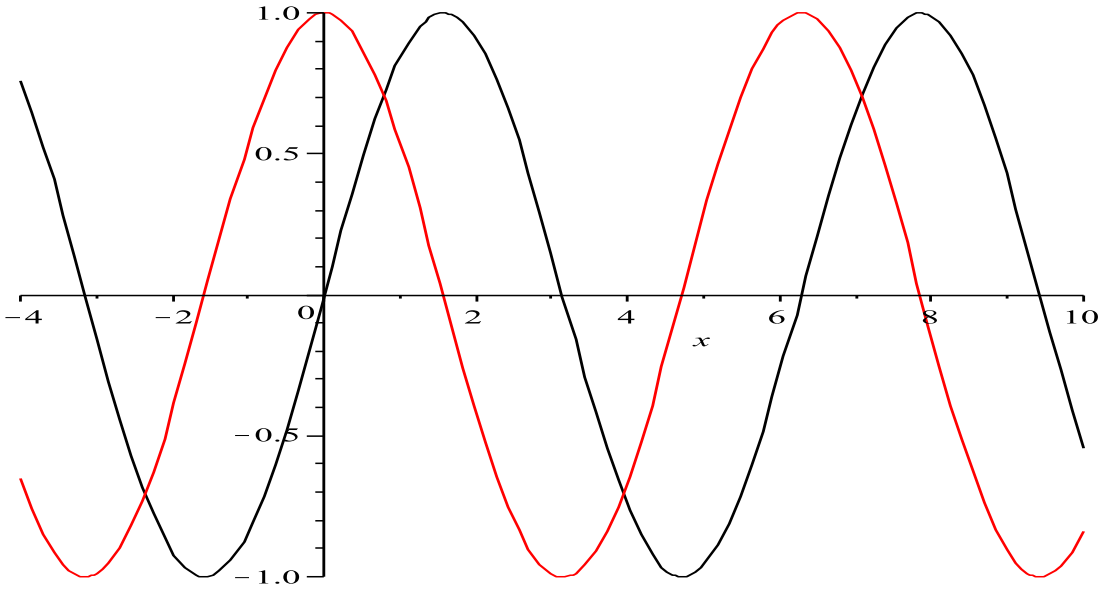
$$\frac{a_0}{2} + \sum_{n=1}^{\infty} \big(a_n \cos(nx) + b_n \sin(nx)\big)$$

heißt *Fourierreihe* (kurz: FR) von f.

(Zu Herrn Jean Baptiste Joseph Fourier s. [WIK].) Die Frage der Konvergenz der Fourierreihe ist immer noch offen; bisher wissen wir also nicht, ob (2.1) gilt, ob also die Fourierreihe von f wirklich (und für alle s bzw x) gegen f konvergiert. Aber falls es klappt, dann ist diese Funktion ja auf ganz $\mathbb{R}$ definiert. Genauer: Wegen $\sin(x) = \sin(x + 2n\pi)$ und $\cos(x) = \cos(x + 2n\pi)$ für alle $x \in \mathbb{R}, n \in \mathbb{Z}$ ist dann auch $f(x) = f(x + 2n\pi)$ für alle $x \in \mathbb{R}, n \in \mathbb{Z}$. Das heißt, die Funktion ist 2π-periodisch:

Definition 2.2. Gegeben eine Funktion $f : \mathbb{R} \to \mathbb{R}$.
(a) f heißt *T-periodisch*, falls für alle $x \in \mathbb{R}$ gilt: $f(x) = f(x + T)$. Das (betragsmäßig) kleinste solcher T heißt *Periode* von f.
(b) f heißt *gerade* [bzw *ungerade*], falls $f(-x) = f(x)$ [bzw $-f(-x) = f(x)$].

Beispiel 2.3. Die Sinusfunktion ist 2π-periodisch und ungerade (schwarzer Graph):

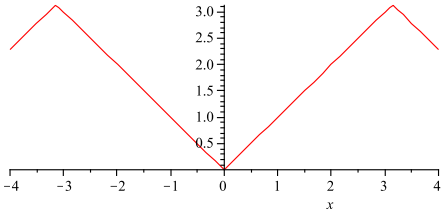


Die Cosinusfunktion ist 2π-periodisch und gerade (roter [bzw. grauer] Graph).

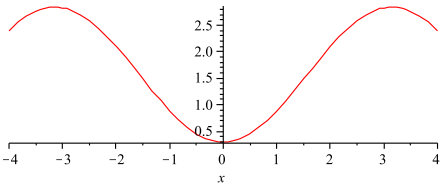
2.1 Approximation von periodischen Funktionen

Wir stellen uns folgendem Problem: eine gegebene 2π-periodische Funktion soll durch einfache Funktionen approximiert werden. Einfache Funktionen sind z.B. Polynome. Leicht überlegt man sich aber, warum Polynome hier ungeeignet sind (warum?). Geeignet sind *trigonometrische Polynome*, das sind Funktionen der Form

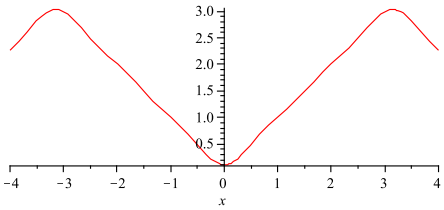
$$f(x) = \tilde{a}_0 + \sum_{n=1}^{N} \big(a_n \cos(nx) + b_n \sin(nx)\big).$$



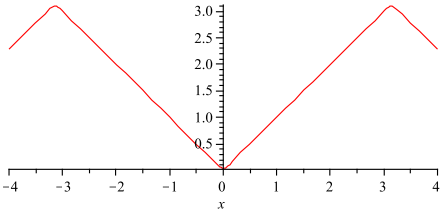
(a)



(b)



(c)



(d)

Figure 1: Approximationen der Dreiecksfunktion (a) durch die ersten 2 Terme seiner Fourierreihe (b); bzw die ersten 4 Terme (c); bzw die ersten 9 Terme (d).

(Ein Ingenieur würde das Problem so formulieren, dass er ein vorgegebenes Schwingungsmuster [z.B. Sägezahnschwingung, Rechteckschwingung] durch Überlagerung "reiner" Schwingungen erzeugt, oder zumindest annähert. Ein Musiker würde sich dafür interessieren, welche hochfrequenten Töne (Obertöne) er durch gleichzeitiges Erklingen lassen von Grundtönen erzeugen kann; s. [WIK].)

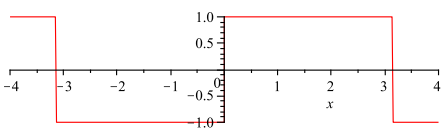
Und da haben wir ja bereits eine mögliche Lösung (bis auf die immer noch offenen Konvergenzfragen): Wir nehmen die ersten N Summanden der Fourierreihe von f. Probieren wir es einfach mal aus:

Beispiel 2.4. Sei $\tilde{f} : [-\pi, \pi] \to \mathbb{R}$, $f(x) = |x|$; und sei f die 2π-periodische Fortsetzung von $\tilde{f}$ auf ganz $\mathbb{R}$ ("Dreiecksfunktion", siehe Fig. 1 (a)).

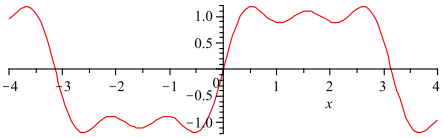
Die Fourierkoeffizienten sind

$$a_0 = \pi, \quad b_n = 0, \quad a_n = \begin{cases} -\frac{4}{\pi n^2} & : n \text{ ungerade} \\ 0 & : n \text{ gerade} \end{cases}$$

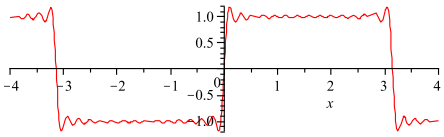
also ist $f(x) = |x| \approx \frac{\pi}{2} + \sum_{n=1}^{N} \frac{-4\cos\left((2n-1)x\right)}{\pi(2n-1)^2}$. (Zu Details s. Übungsaufgabe 3, Blatt 1.)



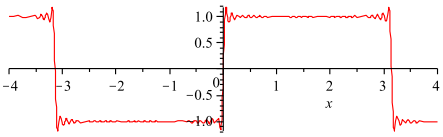
(a)



(b)



(c)



(d)

Figure 2: Approximationen der Rechtecksfunktion (a) durch die ersten 4 Terme seiner Fourierreihe (b); bzw die ersten 16 Terme (c); bzw die ersten 30 Terme (d).

Einige approximierenden Funktionen sind in Bild 1 dargestellt: Für $N = 1$ ($\frac{\pi}{2} - \frac{4}{\pi}\cos(x)$, also zwei Summanden), $N = 3$ (4 Summanden), $N = 8$ (neun Summanden).

Insbesondere fällt im letzten Beispiel auf: Die zu approximierende Funktion war gerade, und die Koeffizienten b_n der (ungeraden) Sinusfunktionen sind alle 0. Das ist ein allgemeines Phänomen:

Bemerkung 2.5. Ist f eine gerade Funktion, so gilt für all seine Sinus-Fourierkoeffizienten $b_n = 0$. Ist f eine ungerade Funktion, so gilt für all seine Cosinus-Fourierkoeffizienten $a_n = 0$.

2.2 Gibbssches Phänomen

Die Funktion im letzten Beispiel war stetig, und die Approximation hat offenbar gut geklappt. Betrachten wir mal ein Beispiel mit einer unstetigen Funktion, der "Rechtecksfunktion" (s. Fig. 2 (a)). Dazu sei

$$\tilde{f} : [-\pi, \pi] \to \mathbb{R},\ \tilde{f}(x) = \begin{cases} 1 & : x > 0 \\ 0 & : x = 0 \\ -1 & : x < 0 \end{cases},$$

und f sei seine 2π-periodische Fortsetzung auf ganz $\mathbb{R}$.

f ist ungerade, also sind alle $a_n = 0$, und es ist

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)dx = \frac{2}{\pi}\int_0^{\pi}\sin(nx)dx = -\frac{2}{\pi n}\cos(nx)\Big|_0^{\pi}$$
$$= -\frac{2}{\pi n}\big(\cos(n\pi)-\cos(0)\big) = \frac{2}{\pi n}\big(1-(-1)^n\big) = \begin{cases} 0 & : n \text{ gerade} \\ \frac{4}{\pi n} & : n \text{ ungerade} \end{cases}$$

Also ist $f(x)$ (hoffentlich) gleich seiner Fourierreihe $\frac{4}{\pi}\big(\sin(x)+\frac{1}{3}\sin(3x)+\frac{1}{5}\sin(5x)+\cdots\big)$. In Fig. 2 sind die Approximationen mittels der ersten 4 bzw 16 bzw 30 Termen der FR dargestellt. Offenbar nähert sich das insgesamt unserem f an. Aber an den Sprungstellen will es nicht richtig konvergieren: Es gibt immer "Überschießer" um etwa 0,2. Diese rücken zwar immer näher an die Sprungstelle heran, aber niedriger werden sie nicht! Dies beschreibt die folgende Bemerkung.

Bemerkung 2.6 (Gibbssches Phänomen)**.** Ist f eine 2π-periodische Funktion, und hat f eine Sprungstelle bei x, und die Höhe des Sprungs ist a. Dann schießt die N-te Fourierreihen-Approximation von f in der Nähe von x um $a\cdot 0,08949...$ über den wahren Funktionswert hinaus. Dieses Maximum wandert mit wachsendem N immer näher gegen x, aber die Höhe bleibt dieselbe, unabhängig von N. (Für Unterschießer analog mit Minima).

(Der allgemeine Beweis ist sehr technisch. Der Beweis für unser Beispiel steht auf [WIK].)

Im Beispiel oben ist die Sprunghöhe 2, also sind die Werte der Maxima (= die Höhe der Spitzen) etwa 1,18. Nun kann man sich fragen: Ist das nun konvergent? Oder nicht?

3 Konvergenz von FRn

Definition 3.1. Sei $f_n : D \to \mathbb{R}$, also $(f_n)_{n\in\mathbb{N}}$ eine Folge von Funktionen. Dann heißt

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

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

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

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

Bemerkung 3.2. $(f_n)_{n\in\mathbb{N}}$ ist glm kgt genau dann, wenn $\|f_n - f\|_\infty \to 0 \quad (n\to\infty)$

Dabei ist $\|\cdot\|_\infty$ die *Supremumsnorm*: $\|f\|_\infty := \sup_{x\in D}|f(x)|$.

Proof. (zu Bem. 3.2) Dass $\|f_n - f\|_\infty \to 0 \quad (n\to\infty)$ gilt, heißt ja:

$$\forall \varepsilon > 0\ \exists N\ \forall n \geq N :\quad \sup_{x\in D}|f_n(x) - f(x)| < \varepsilon$$

Das Supremum über alle $x \in D$ (der "worst-case") des Betrags ist also schon kleiner als ε, also alle andern Beträge auch:

$$\forall \varepsilon > 0\ \exists N\ \forall n \geq N, x \in D :\quad |f_n(x) - f(x)| < \varepsilon.$$

□

Beispiel 3.3. (Zu Supremumsnorm)

- $f:\mathbb{R}\to\mathbb{R},\ f(x)=\cos(x)$. Dann ist $\|f\|_\infty = 1$.
- $f:\mathbb{R}\to\mathbb{R},\ f(x)=x^2$. Dann ist $\|f\|_\infty = +\infty$.
- $f:[-\pi,\pi]\to\mathbb{R},\ f(x)=x^2$. Dann ist $\|f\|_\infty = \pi^2$.

Das Beispiel in 2.6 zeigt: die n-ten Approximanten $f_n(x) = \frac{a_0}{2} + \sum_{k=1}^n \big(a_k\cos(ks) + b_k\sin(ks)\big)$ konvergieren zwar punktweise gegen f, aber nicht glm. Warum? Auf unmathematisch:

Punktweise: Nehmen wir irgendein $0 < x < \pi$, beliebig nahe an 0. Die Spitzen auf dem Plateau wandern für sehr hohe n irgendwann ganz nah an 0. Egal, wie klein x ist, irgendwann sind die Spitzen nach links an x vorbeigewandert, und die FR nähert sich an der Stelle x dem richtigen Wert 1 an. Analog für alle anderen Stellen, die selber keine Sprungstellen sind. An der Sprungstelle $x = 0$ hat dir FR immer den korrekten Wert, nämlich 0 (nachrechnen!). Also geht für jedes x die FR an dieser Stelle irgendwann gegen den richtigen Wert $f(x)$.

Die FR konvergiert aber nicht gleichmäßig: Der Überschießer ist immer 0,178... zu hoch. Also ist $\|f_n - f\|_\infty = \max\limits_{x\in D} |f_n(x) - f(x)| = 0,178\ldots$ für jedes n. Daher $\lim\limits_{n\to\infty} \|f_n - f\|_\infty = 0,178\ldots \neq 0$.

Es folgen nun einige Resultate zu Konvergenz von FRn bzgl dieser beiden Konvergenzbegriffe.

Definition 3.4. Eine Funktion $f:[a,b]\to\mathbb{R}$ heißt von *beschränkter Variation*, falls es ein $M > 0$ gibt, so dass für alle Zerlegungen $x_0 = a < x_1 < x_2 < \cdots < x_n = b$ gilt:

$$\sum_{k=1}^n |f(x_k) - f(x_{k-1})| \le M.$$

Satz 3.5. *Ist $f:[-\pi,\pi]\to\mathbb{R}$ von beschränkter Variation, so konvergiert die FR von f für jedes $x\in]-\pi,\pi[$ gegen $s(x) = \frac{1}{2}\big(f(x^+) + f(x^-)\big)$. Insbesondere konvergiert die FR punktweise auf ganz $]-\pi,\pi[$. Ist f in x außerdem stetig, konvergiert die FR an der Stelle x gegen $f(x)$.*

(Der Beweis ist lang und technisch, machen wir hier nicht, s. [H3].)

Dabei bezeichnet $f(x^+)$ den rechtsseitigen Grenzwert von f an der Stelle x (und $f(x^-)$ den linksseitigen). Damit kann man zeigen:

Satz 3.6. *Ist f diff.-bar auf $[-\pi,\pi]$, oder Lipschitzstetig, oder monoton steigend (bzw fallend), so konvergiert die FR von f punktweise.*

Eine Funktion $f:D\to\mathbb{R}$ heißt *Lipschitzstetig*, falls

$$\exists L > 0\ \forall x,y\in D: \quad |f(x) - f(y)| \le L|x-y|.$$

Aus f Lipschitzstetig folgt f stetig.

Proof. (Satz 3.6) (Nur für Lipschitzstetig, monoton: Übung) Es gilt, in der Situation des Satzes,

$$\begin{aligned}\sum_{k=1}^n |f(x_k) - f(x_{k-1})| &\le \sum_{k=1}^n L|x_k - x_{k-1}| = L\sum_{k=1}^n (x_k - x_{k-1}) \\ &= L(x_1 - x_0 + x_2 - x_1 + \cdots x_{n-1} - x_{n-2} + x_n - x_{n-1}) = L(x_n - x_0) = L(a-b)\end{aligned}$$

Mit $M := L(b-a)$ ist die Definition beschränkter Variation erfüllt. Mit Satz 3.5 folgt die Behauptung. □

Bemerkung 3.7. Die Forderung "f stetig" reicht nicht einmal für punktweise Konvergenz der FR!

Aber aus einem Resultat von Carleson (dazu später) von 1966 (Wolf-Preis 1992, Abel-Preis 2006) folgt dieser Satz:

Satz 3.8. *Sei $f : [-\pi,\pi] \to \mathbb{R}$ stetig. Dann konvergiert die FR von f fast überall gegen f.*

Also nicht "punktweise" (= überall), sondern nur fast überall. Dabei ist "fast überall" ein genau definierter Begriff: D.h., dass die Menge aller $x \in [-\pi,\pi]$, für die dies gilt, Maß 2π hat. Also z.B. für alle außer endlich viele Punkte in $[-\pi,\pi]$.

Einschub: Fast überall

Erinnerung: In Wahrscheinlichkeitstheorie gibt es Ereignisse, die mit W. 0 auftreten. Z.B. wird beim wiederholten Werfen einer Münze irgendwann Mal "Kopf" geworfen. Das Ereignis "Es wird unendlich oft Zahl geworfen" ist zwar nicht unmöglich, hat aber W. 0. "Fast immer" — d.h. in diesem Fall: mit W. 1 — wird irgendwann Kopf fallen.

"Fast überall" in $[-\pi,\pi]$ heißt: Überall, außer in einer Menge mit Maß 0. Wenn es z. B. überall gilt außer in endlich vielen Punkten, dann gilt es fast überall.

Es gibt aber sogar überabzählbare Ausnahmemengen, z.B. die Cantormenge.

Zur glm Kgz gilt folgendes Resultat:

Satz 3.9. *Ist $f : [-\pi,\pi] \to \mathbb{R}$ stetig differenzierbar (also diff-bar mit stetiger Ableitung), so ist die FR von f glm kgt gegen f.*

(Bew.: [H2], Satz 136.5 und folgende Bemerkung.)

Ein analoges Ergebnis gilt für stückweise stetige Funktionen f: Ist f auf $[-\pi,\pi]$ stückweise stetig diff-bar, so konvergiert die FR auf jedem abgeschlossenen Teilintervall von $[-\pi,\pi]$, das keine Unstetigkeitsstelle enthält.

Punktweise Konvergenz gilt also oft, ist aber (s. Gibbssches Ph.) keine starke Eigenschaft. Glm Kgz dagegen ist eine zu starke Eigenschaft. Es gibt einen befriedigenderen Konvergenzbegriff: Kgz im quadratischen Mittel. D.h. im Wesentlichen, dass die Fläche der Graphen der Differenzen $f_n - f$ gegen 0 geht. Das wird sehr einfach und natürlich, sobald man Hilberträume kennt.

4 Hilberträume

Erinnerung: Ein Euklidischer Vektorraum (VR) ist ein VR mit einem Innenprodukt (oft auch "Skalarprodukt" geheißen). Es ist völlig OK, sich dazu $\mathbb{R}^2$ oder $\mathbb{R}^3$ vorzustellen mit dem Standardskalarprodukt (hier für $\mathbb{R}^2$):

$$\langle x,y\rangle = \sum_{n=1}^{2} x_n y_n = x_1 y_1 + x_2 y_2.$$

Die *Länge* (oder *Norm*) eines Vektors x ist dann $\|x\| = \sqrt{\langle x,x\rangle}$. Die *Distanz* zwischen zwei Punkten x und y ist dann

$$d(x,y) = \sqrt{\langle x-y, x-y\rangle} = \|x-y\|.$$

Zwei Vektoren x,y sind *orthogonal* (oder senkrecht) zueinander, falls gilt $\langle x,y\rangle = 0$. Auch die Konvergenz von Vektoren lässt sich mittels der Norm $\|\cdot\|$ erklären.

Fein. Nehmen wir nun statt Vektoren Funktionen, und hoffen, dass wir diese mittels einer geeigneten Norm zu einem euklidischen VR machen können. Konkreter: Sei $C([-\pi,\pi])$ die Menge aller stetigen Funktionen mit Definitionsbereich $[-\pi,\pi]$ und Werten in $\mathbb{R}$. Damit es ein VR wird, müssen wir zwei Elemente (hier also Funktionen) addieren dürfen, das klappt: $f+g$ ist erklärt durch $(f+g)(x) = f(x)+g(x)$. Und wir müssen eine Funktion mit einer reellen Zahl (einem *Skalar*) multiplizieren dürfen, das geht auch: für $\alpha \in \mathbb{R}$ ist αf erklärt durch $(\alpha f)(x) = \alpha f(x)$.

Inspiriert von (2.2), oder von der Darstellung der Fourierkoeffizienten (s. Def. 2.1), setzen wir nun

$$\langle f,g\rangle = \int_{-\pi}^{\pi} f(x)g(x)dx \tag{4.1}$$

Dann ist, analog zu oben, die *Norm* einer Funktion

$$\|f\|_2 = \sqrt{\langle f,f\rangle} = \sqrt{\int_{-\pi}^{\pi} f^2(x)dx}$$

und die *Distanz* (auch: Abstand) zweier Funktionen

$$d(f,g) = \sqrt{\langle f-g, f-g\rangle} = \sqrt{\int_{-\pi}^{\pi} (f(x)-g(x))^2 dx}$$

Wenn klar ist, von welcher Norm wir gerade sprechen, schreiben wir auch $\|f\|$ statt $\|f\|_2$. Norm ist ein ganz allgemeiner Begriff:

Definition 4.1. Sei V ein VR über $\mathbb{R}$ (bzw $\mathbb{C}$). Eine Abbildung $N : V \to \mathbb{R}$ heißt *Norm* (auf V), falls für alle $f,g \in V$ und alle $\alpha \in \mathbb{R}$ (bzw $\mathbb{C}$) gilt:

(N1) $N(f) \geq 0 \quad$ und $\quad N(f) = 0 \Leftrightarrow f = 0.$

(N2) $N(\alpha f) = |\alpha| N(f)$

(N3) $N(f+g) \leq N(f) + N(g)$

Die Ungleichung (N3) heißt auch “Dreiecksungleichung”.

Das oben definierte ist tatsächlich eine Norm, wie man nachrechnen kann. Dazu ist es hilfreich, zunächst einige Eigenschaften des Innenprodukts festzuhalten:

Bemerkung 4.2. Für obiges Innenprodukt, wie allgemein für Innenprodukte über $\mathbb{R}$, gilt:

(I1) $\langle f+g,h\rangle = \langle f,h\rangle + \langle g,h\rangle$

(I2) $\langle \alpha f,g\rangle = \alpha\langle f,g\rangle$

(I3) $\langle f,g\rangle = \langle g,f\rangle$

(Das rechnet man sehr leicht nach).

Proof. (Zu: $\|\cdot\|_2$ ist Norm auf $C([-\pi,\pi])$.)

Zu (N1): Es ist immer $f(x)^2 \geq 0$, also auch $\int_{-\pi}^{\pi} f^2(x)dx \geq 0$. Wenn das Integral gleich 0 ist, dann muss (s. Maßtheorie) $f(x) = 0$ fast überall gelten. Da f stetig ist, ist es dann überall 0. Also ist f die Nullfunktion, und diese spielt hier natürlich die Rolle des Nullelements (des neutralen Elements).

Zu (N2): $\|\alpha f\| = \sqrt{\int_{-\pi}^{\pi}(\alpha f(x))^2 dx} = \sqrt{\int_{-\pi}^{\pi}\alpha^2 f(x)^2 dx} = \sqrt{\alpha^2}\sqrt{\int_{-\pi}^{\pi} f(x)^2 dx} = |\alpha|\|f\|$. □

Für den Nachweis von (N3) brauchen wir die folgende wichtige und sehr grundlegende Aussage.

Satz 4.3 (Cauchy-Schwarzsche Ungleichung, CSU)**.** *In einem Vektorraum mit Innenprodukt gilt für alle f,g:*

$$|\langle f,g\rangle| \leq \sqrt{\langle f,f\rangle}\sqrt{\langle g,g\rangle} = \|f\|\|g\|.$$

Gleichheit herrscht genau dann, wenn f und g linear abhängig sind. (Was heißt das hier? $f = \alpha g$)

Proof. Wegen (N1), (I1) und (I2) wissen wir bereits:

$$0 \leq \langle f+\alpha g, f+\alpha g\rangle = \langle f,f\rangle + \alpha\langle f,g\rangle + \alpha\langle g,f\rangle + \alpha^2\langle g,g\rangle = \langle f,f\rangle + 2\alpha\langle f,g\rangle + \alpha^2\langle g,g\rangle$$

Mit $\alpha = -\langle f,g\rangle/\langle g,g\rangle$ wird daraus

$$0 \leq \langle f,f\rangle - 2\frac{\langle f,g\rangle^2}{\langle g,g\rangle} + \frac{\langle f,g\rangle^2}{\langle g,g\rangle} = \langle f,f\rangle - \frac{\langle f,g\rangle^2}{\langle g,g\rangle}, \quad \text{also} \quad \langle f,g\rangle^2 \leq \langle f,f\rangle\langle g,g\rangle,$$

daraus folgt die Behauptung. (Hier: $g \neq 0$, falls $g = 0$ ist's trivial.) Falls f,g linear abhängig sind, rechnet man Gleichheit sehr leicht nach. □

Proof. (Forts. zu oben) Zu (N3):

$$\|f+g\|^2 = \langle f+g, f+g\rangle = \langle f,f\rangle + \langle f,g\rangle + \langle g,f\rangle + \langle g,g\rangle = \|f\|^2 + 2\langle f,g\rangle + \|g\|^2,$$

und wegen der CSU ist das kleiner oder gleich $\|f\|^2 + 2\|f\|\|g\| + \|g\|^2 = (\|f\| + \|g\|)^2$. Wurzelziehen liefert die Behauptung. □

Analog zu oben sind zwei Funktionen f,g *orthogonal* zueinander, kurz: $f \perp g$, falls $\langle f,g\rangle = 0$.

Beispiel 4.4. Sei $f_n : [-\pi,\pi] \to \mathbb{R}$, $f_n(x) = \cos(nx)$ und $g_n : [-\pi,\pi] \to \mathbb{R}$, $g_n(x) = \sin(nx)$. Nach (2.2), (2.3) und (2.4) gilt: $f_n \perp g_m$ für alle $m,n \in \mathbb{N}$, und $f_n \perp f_m$ sowie $g_n \perp g_m$ für alle $m \neq n \in \mathbb{N}_0$.

Nun kommen wir endlich zum versprochenen Konvergenzbegriff. Den liefert uns nun natürlich die Norm $\|\cdot\|_2$. Die heißt auch L^2-Norm.

Definition 4.5. Eine Folge von Funktionen f_n in $C([-\pi,\pi])$ heißt *konvergent im quadratischen Mittel* (oder kurz L^2-konvergent) gegen f, falls $\|f_n - f\|_2 \to 0 \quad (n \to \infty)$.

Mit diesem Konvergenzbegriff wird nun vieles "schön". Aber zunächst:

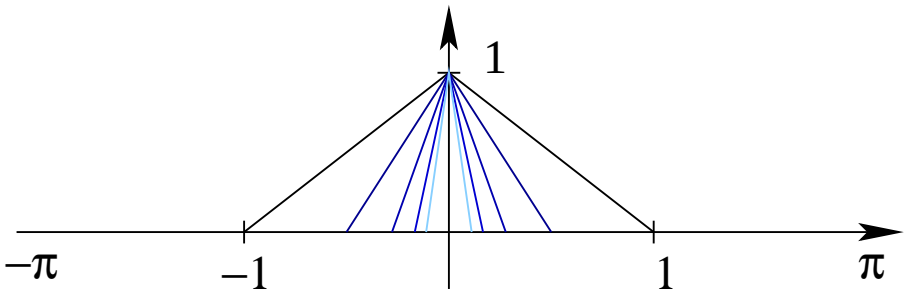


Figure 3: Eine Funktionenfolge, die im quadratischen Mittel konvergiert, aber nicht punktweise.

Beispiel 4.6. (1) Beim Gibbschen Phänomen konvergieren die Approximanten f_n (die Summe der ersten n Terme der FR der Rechteckfunktion f) im quadratischen Mittel gegen f.

(2) Die Funktionenfolge

$$f_n : [-\pi,\pi] \to \mathbb{R},\ f(x) = \begin{cases} nx+1 & : -\frac{1}{n} \le x < 0 \\ -nx+1 & : 0 \le x \le \frac{1}{n} \\ 0 & : \text{ sonst} \end{cases}$$

(s. Fig. 3) konvergiert nicht punktweise gegen 0, aber im quadratischen Mittel schon. (Punktweise konvergiert's gegen etwas anderes: f mit $f(0) = 1, f(x) = 0$ sonst.)

(3) Die Folge $f_n : [-\pi,\pi] \to \mathbb{R},\ f_n(0) = n,\ f(x) = 0$ sonst, konvergiert im quadratischen Mittel auch gegen die Nullfunktion, aber punktweise konvergiert sie gar nicht (Divergenz im Punkt 0!).

(4) Es geht auch anders: Die in Fig. 4 dargestellte Funktionenfolge konvergiert punktweise gegen die Nullfunktion, aber nicht im quadratischen Mittel!

Insgeamt bestehen folgende Zusammenhänge für Kgz von Funktionenfolgen in $C([-\pi,\pi])$ bzw allgemein in $C([a,b])$:

punktweise Kgz	$\Leftarrow$	Kgz bzgl $\|\cdot\|_\infty$	$\Leftrightarrow$	glm Kgz	$\Rightarrow$	Kgz im quadr Mittel

Nun ist also $C([a,b])$ mit $\|\cdot\|_2$ (oder auch mit $\|\cdot\|_\infty$) ein euklidischer VR. Er hat aber einen Schönheitsfehler: Eine Funktionenfolge in $C([a,b])$ kann gegen etwas konvergieren, das nicht in $C([a,b])$ liegt (s. Bsp. 4.6 (1) und (2)). Wir wollen einen Raum, der in diesem Sinne abgeschlossen ist. Diese Eigenschaft heißt "vollständig" und ist genau genommen so erklärt:

Definition 4.7. Ein Raum X mit einer Norm heißt *vollständig*, falls jede Cauchyfolge in X konvergent ist. Eine Folge $(a_n)_{n\in\mathbb{N}}$ heißt *Cauchyfolge*, falls

$$\forall \varepsilon > 0\ \exists N \in \mathbb{N}\ \forall m,n \ge N : \|a_n - a_m\| < \varepsilon.$$

Jede konvergente Folge ist eine Cauchyfolge.

Beispiel 4.8. (a) (Blödes Beispiel) Im Raum $X = \mathbb{Q}$ (!) ist die Folge $(a_n) = (3;\ 3,1;\ 3,14;\ 3,141;\ 3,1415;\ 3,14159;\ 3,141592;\ 3,1415926\ldots)$ eine Cauchyfolge (denn sie konvergiert ja gegen π), aber, da $\pi \notin \mathbb{Q}$, ist sie nicht konvergent (in $\mathbb{Q}$): Es gibt kein $a \in \mathbb{Q}$ mit $|a_n - a| \to 0$.

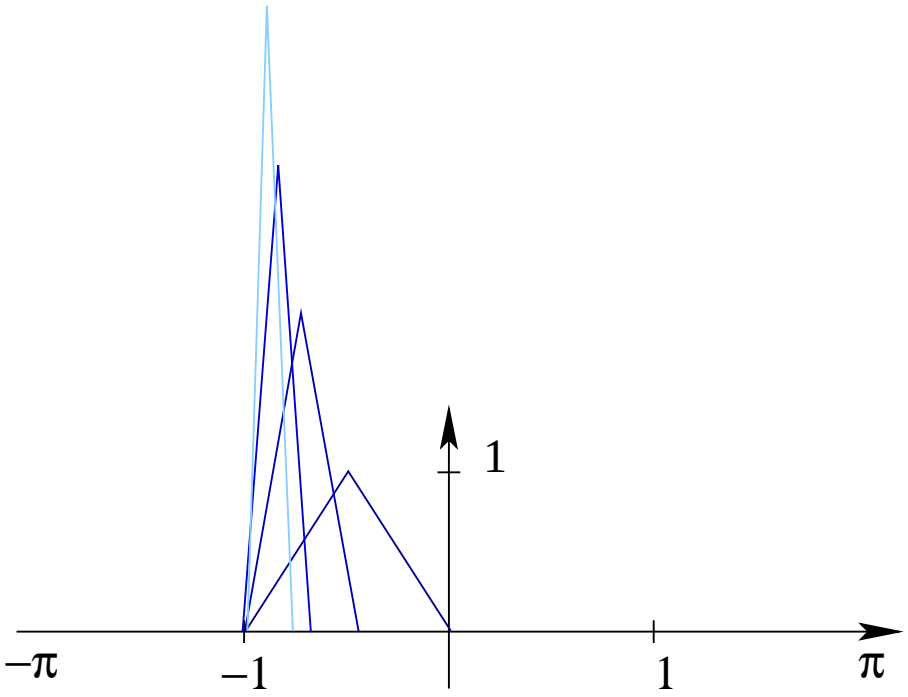


Figure 4: Eine Funktionenfolge, die punktweise konvergiert, aber nicht im quadratischen Mittel.

(b) (Etwas weniger blödes Beispiel) Der Raum $C([a,b])$ ist nicht vollständig, siehe oben.

(c) (Gutes Beispiel) Die Funktionen in Fig. 4 sind keine Cauchyfolge bzgl $\|\cdot\|_2$, vgl Aufg 12, Blatt 4.

Definition 4.9. Ein VR mit Innenprodukt, der bezüglich der dadurch induzierten Norm vollständig ist, heißt *Hilbertraum*.

Bemerkung 4.10. Ein allgemeinerer Begriff sind *Banachräume*. Das sind auch vollständige normierte Vektorräume. Aber in einem Hilbertraum muss die Norm durch das Innenprodukt gegeben sein. Falls das nicht so ist, hat man einen Banachraum, aber keinen Hilbertraum. (Genauer: Ein Banachraum ist ein Hilbertraum genau dann, wenn in ihm die Parallelogrammgleichung gilt, vgl Aufgabe 10, Blatt 3). Im Folgenden brauchen wir aber nur Hilberträume.

Beispiel 4.11. (a) Der $\mathbb{R}^d$ wird mit dem Standardskalarprodukt $\langle (a_1,\dots,a_n),(b_1,\dots,b_n)\rangle := \sum_{n=1}^{d} a_n b_n$ zu einem Hilbertraum.

(b) Die Menge aller "unendlichen" Vektoren $(a_1,a_2,a_3,\dots)$ ist ein VR, aber kein Hilbertraum bzgl $\langle (a_n)_n,(b_n)_n\rangle := \sum_{n=1}^{\infty} a_n b_n$. Aber dieser: Es sei ℓ^2 der Raum aller Folgen $(a_n)_n = (a_1,a_2,\dots)$ (wobei $a_n \in \mathbb{R}$) mit $\sum\limits_{n=1}^{\infty} a_n^2 < \infty$. Durch

$$\langle (a_n)_n,(b_n)_n\rangle := \sum_{n=1}^{\infty} a_n b_n$$

ist ein Innenprodukt auf ℓ^2 gegeben. Damit wird ℓ^2 zu einem Hilbertraum.

Nun zur ersten schönen Eigenschaft:

Satz 4.12 (und Definition). *Den Raum aller Funktionen $f : [a,b] \to \mathbb{R}$ mit $\|f\|_2 < \infty$, versehen mit dem Innenprodukt*

$$\langle f,g\rangle = \int_a^b f(x)g(x)dx \qquad (\textit{wobei } \|f\|_2 < \infty,\ \|g\|_2 < \infty),$$

bezeichnen wir mit $L^2([a,b])$. ("Raum aller quadratintegrierbaren Funktionen auf $[a,b]$").
$L^2([a,b])$ ist ein Hilbertraum.

(Beweis: es bleibt die Vollständigkeit zu zeigen. Das ist länglich, s. [H2], Satz 130.5.)

Obacht: In Wirklichkeit ist L^2 mit der Norm $\|\cdot\|_2$ eine Bedingung verletzt, nämlich (N1), 2. Teil: Aus $\|f\|_2 = 0$ folgt nicht zwingend, dass f die Nullfunktion ist. Mit

$$f : [-\pi,\pi] \to \mathbb{R},\ f(x) = \begin{cases} 0 & : x \neq 0 \\ 1 & : x = 0 \end{cases}$$

ist $\|f\|_2 = \int_{-\pi}^{\pi} f(x)dx = \int_{-\pi}^{0} 0dx + \int_0^0 1dx + \int_0^{\pi} 0dx = 0+0+0 = 0$. Daher werden wir im Folgenden immer Funktionen f,g als gleich auffassen, für die gilt: $\|f-g\|_2 = 0$. (Die unterscheiden sich dann lediglich auf einer Menge vom Maß 0.)

4.1 ON-Systeme

In einem normierten VR ist es nützlich, eine Basis zu haben. Besonders praktisch ist eine Orthonormal-Basis (ON-Basis), das ist eine Basis $\{b_1, b_2, \dots b_d\}$ mit $b_n \perp b_m$ für $m \neq n$ (also $\langle b_n, b_m\rangle = 0$), und $\|b_n\| = 1$ für alle $1 \leq n \leq d$. Bei $L^2([a,b])$ stehen wir aber vor dem Problem, dass es gar keine endliche Basis gibt! Nicht einmal, im klassichen Sinne, eine abzählbar unendliche, wie in ℓ^2.

Definition 4.13. Eine Menge $\{b_1, b_2, b_3, \dots\}$ in einem Hilbertraum heißt *Orthonormalsystem* (ONS), falls $b_n \perp b_m$ für alle m,n mit $m \neq n$, und $\|b_n\| = 1$ für alle n.
Ein ONS heißt *maximal*, falls einzig das Nullelement orthogonal zu allen b_n des ONS ist.

"Orthonormal" ist zusammengesetzt aus "orthogonal" (also $b_n \perp b_m$) und "normal" (also $\|b_n\| = 1$).

Satz 4.14. *Die Funktionen $\frac{1}{2\sqrt{\pi}}, \frac{1}{\sqrt{\pi}}\sin(\cdot), \frac{1}{\sqrt{\pi}}\cos(\cdot), \frac{1}{\sqrt{\pi}}\sin(2\cdot), \frac{1}{\sqrt{\pi}}\cos(2\cdot), \dots$ sind ein maximales ONS in $L^2([-\pi,\pi]$.*

Die Orthogonalität ist uns bereits bekannt, siehe (2.2), (2.3), (2.4). Die Vorfaktoren $\frac{1}{\sqrt{\pi}}$ sorgen für die Normalität:

$$\|\frac{1}{\sqrt{\pi}}\cos(n\cdot)\|_2 = \sqrt{\int_{-\pi}^{\pi} \frac{1}{\pi}\cos^2(nx)dx} = \frac{1}{\sqrt{\pi}}\sqrt{\int_{-\pi}^{\pi}\cos^2(nx)dx} = 1,$$

vgl (2.3). Die Maximalität ist etwas knifflig, siehe [H2], Satz 141.3.

Zur zweiten schönen Eigenschaft: Wir sahen bereits, dass wir eine 2π-periodische Funktion auf $\mathbb{R}$ (oder allgemeiner, durch Skalierung, irgendeine periodische Funktion auf $\mathbb{R}$) oft gut approximieren können durch die Teilsummen seiner FR. Wie oft, und was heißt "gut"?

4.2 Die Gaußsche Approximationsaufgabe

Die Gaußsche Approximationsaufgabe in einem Innenproduktraum X ist diese: Gegeben $f \in X$ und ein ONS $\{e_1, e_2, e_3, \ldots, e_n\}$. Gesucht sind $\alpha_1, \alpha_2, \ldots, \alpha_n$, so dass

$$\| f - \sum_{k=1}^{n} \alpha_k e_k \|$$

minimal wird. In $L^2 = L^2([-\pi, \pi])$ wird daraus: Gegeben $f \in L^2$. Wir setzen $e_1 = \frac{1}{2\sqrt{\pi}}$, $e_2 = \frac{1}{\sqrt{\pi}} \sin(\cdot)$, $e_3 = \frac{1}{\sqrt{\pi}} \cos(\cdot)$, $e_4 = \frac{1}{\sqrt{\pi}} \sin(2\cdot)$, $e_5 = \frac{1}{\sqrt{\pi}} \cos(2\cdot)$ usw. Gesucht sind α_k, so dass für ein vorgegebenes $n \in \mathbb{N}$

$$\| f - \sum_{k=1}^{n} \alpha_k e_k \|_2$$

minimal wird.

Satz 4.15. *In L^2 mit dem ONS $O = \{e_1, e_2, \ldots, e_n\}$ wird die Gaußsche Approximationsaufgabe eindeutig gelöst durch $\alpha_k = \langle f, e_k \rangle$ ($k = 1, \ldots, n$).*
Dabei ist $f - \sum_{k=1}^{n} \alpha_k e_k$ orthogonal zu O.

Proof. Seien jetzt c_k irgendwelche Koeffizienten. Wir wollen zeigen, dass $c_k = \alpha_k$ optimal ist. Sei $F_n = \sum_{k=1}^{n} c_k e_k$.

$$\begin{aligned}
\| f - F_n \|^2 &= \langle f - F_n, f - F_n \rangle = \langle f, f \rangle - 2\langle f, F_n \rangle + \langle F_n, F_n \rangle && (4.2)\\
&= \| f \|^2 - 2\langle f, \sum_{k=1}^{n} c_k e_k \rangle + \langle \sum_{k=1}^{n} c_k e_k, \sum_{m=1}^{n} c_m e_m \rangle && (4.3)\\
&= \| f \|^2 - 2 \sum_{k=1}^{n} |c_k| \langle f, e_k \rangle + \sum_{k=1}^{n} \sum_{m=1}^{n} |c_k| |c_m| \langle e_k, e_m \rangle && (4.4)\\
&= \| f \|^2 - \sum_{k=1}^{n} 2|c_k| \langle f, e_k \rangle + \sum_{k=1}^{n} |c_k|^2 && (4.5)\\
&= \| f \|^2 - \sum_{k=1}^{n} |\langle f, e_k \rangle|^2 + \sum_{k=1}^{n} |c_k - \langle f, e_k \rangle|^2. && (4.6)
\end{aligned}$$

(Die letzte Gleichheit folgt aus der binomischen Formel: $-2ab + b^2 = (a-b)^2 - a^2$.) Nun hängt der Ausdruck, den wir minimieren wollen, nur noch im letzten Summanden von den c_k ab. Und wann wird der am kleinsten? Genau für $c_k = \langle f, e_k \rangle$.

Zur Orthogonalitätsaussage siehe Aufgabe 16, Blatt 5. □

Damit steht nun fest: **Die Teilsummen der FR zu $f \in L^2$ liefern immer die beste Approximation an f mit trigonometrischen Polynomen!** Und zwar "beste" im Sinne der L^2-Norm (also des "Abstands" bzgl L^2). Mit den Bezeichnungen aus Def. 2.1 haben wir:

$$\alpha_{2k} = a_k, \quad \alpha_{2k-1} = b_k \tag{4.7}$$

Außerdem gilt:

Satz 4.16. *Sei $f \in L^2$. Die FR von f konvergiert gegen f bzgl $\|\cdot\|_2$.*

Dazu brauchen wir folgendes Resultat. Für $\|\cdot\|_2$ schreiben wir kurz $\|\cdot\|$, das spart Tinte.

Proposition 4.17 (Besselsche Gleichung)**.** *Sei F_n die Summe der ersten n Summanden der FR von $f \in L^2$, also $F_n = \sum_{k=1}^{n} \langle f, e_k\rangle e_k$. Dann gilt die* Besselsche Gleichung

$$\|f - \sum_{k=1}^{n} \langle f, e_k\rangle e_k\|^2 = \|f\|^2 - \sum_{k=1}^{n} |\langle f, e_k\rangle|^2$$

sowie die Besselsche Ungleichung

$$\sum_{k=1}^{n} |\langle f, e_k\rangle|^2 \leq \|f\|^2$$

Proof. Dazu werden wir die Überlegungen aus dem Beweis von 4.15 recyceln.

$$\begin{aligned}
\|f - F_n\|^2 = \langle f - F_n, f - F_n\rangle &= \langle f, f\rangle - 2\langle f, F_n\rangle + \langle F_n, F_n\rangle && (4.8)\\
&= \|f\|^2 - 2\langle f, \sum_{k=1}^{n} \langle f, e_k\rangle e_k\rangle + \langle \sum_{k=1}^{n} \langle f, e_k\rangle e_k, \sum_{m=1}^{n} \langle f, e_m\rangle e_m\rangle && (4.9)\\
&= \|f\|^2 - 2\sum_{k=1}^{n} |\langle f, e_k\rangle||\langle f, e_k\rangle| + \sum_{k=1}^{n}\sum_{m=1}^{n} |\langle f, e_k\rangle||\langle f, e_m\rangle|\langle e_k, e_m\rangle && (4.10)\\
&= \|f\|^2 - 2\sum_{k=1}^{n} |\langle f, e_k\rangle|^2 + \sum_{k=1}^{n} |\langle f, e_k\rangle|^2 && (4.11)\\
&= \|f\|^2 - \sum_{k=1}^{n} |\langle f, e_k\rangle|^2. && (4.12)
\end{aligned}$$

Also gilt die Gleichung. Für die Ungleichung beachten wir nur, dass die linke Seite ≥ 0 ist, also folgt für die rechte Seite: $\|f\|^2 \geq \sum_{k=1}^{n} \langle f, e_k\rangle^2$. □

Insbesondere stellen wir fest, dass die Ungleichung für alle $n \in \mathbb{N}$ gilt, immer ist die Summe kleiner oder gleich $\|f\|^2 < \infty$. D.h., die Reihe konvergiert. Daher muss die Folge $|\langle f, e_k\rangle|^2$ eine Nullfolge sein! Und damit ist auch die Folge $(\alpha_k)_k$ mit $\alpha_k = \langle f, e_k\rangle$ eine Nullfolge. Damit bekommen wir praktisch geschenkt (vgl (4.7)):

Satz 4.18 (Riemann-Lebesgue)**.** *Für die Fourierkoeffizienten a_n, b_n einer Funktion $f \in L^2$ gilt:* $a_n \to 0,\ b_n \to 0\ (n \to \infty)$.

Proof. (Satz 4.16) The circle has closed. The Jedi will rule again. STAR WARS

Es ist

$$\|\langle f, e_k\rangle e_k\|^2 = \langle \langle f, e_k\rangle e_k, \langle f, e_k\rangle e_k\rangle = \langle f, e_k\rangle^2 \langle e_k, e_k\rangle = \langle f, e_k\rangle^2,$$

da ja $\|e_k\| = 1$. Nach dem (verallgemeinerten) Satz des Pythagoras gilt dann für $n \geq m$:

$$\left\| \sum_{k=m}^{n} \langle f, e_k\rangle e_k \right\|^2 = \sum_{k=m}^{n} \|\langle f, e_k\rangle e_k\|^2 = \sum_{k=m}^{n} \langle f, e_k\rangle^2$$

Wegen der Besselschen Ungleichung (vgl Anmerkung oben) ist ja die Reihe $\sum_{k=1}^{\infty} \langle f, e_k\rangle^2$ konvergent. Also kann man zu jedem $\varepsilon > 0$ ein m finden, so dass für alle $n \geq m$ gilt: $\sum_{k=m}^{n} \langle f, e_k\rangle^2 < \varepsilon$. Die letzte Gleichung besagt daher, dass die Folge der Partialsummen der Reihe (also die folge $(\sum_{k=1}^{n} \langle f, e_k\rangle e_k)_{n\in\mathbb{N}}$ eine Cauchyfolge (bzgl L^2) ist. Weil nun L^2 vollständig ist (!), ist diese Reihe auch konvergent gegen ein $g \in L^2$.

Es bleibt zu zeigen, dass $g = f$. Es gilt

$$\langle g - f, e_k\rangle = \langle g, e_k\rangle - \langle f, e_k\rangle = \left\langle \sum_{j=1}^{\infty} \langle f, e_j\rangle, e_k \right\rangle - \langle f, e_k\rangle = \sum_{j=1}^{\infty} \langle f, e_j\rangle\langle e_j, e_k\rangle - \langle f, e_k\rangle,$$

und da $\langle e_j, e_k\rangle = 0$ für $j \neq k$ gilt, überlebt nur ein Summand der Summe, der k-te. Damit ist $\langle g - f, e_k\rangle = \langle f, e_k\rangle - \langle f, e_k\rangle = 0$. Also ist $g - f$ orthogonal zu allen e_k. Wir wissen aber (Satz 4.14), dass unsere e_k ein maximales ONS bilden. Daher muss ein Element, das orthogonal zu allen e_k steht, das Nullelement sein. Also ist $\|g - f\| = 0$, und bzgl L^2 ist also $f = g$ (vgl dazu die Bemerkung nach Satz 4.12). □

Bemerkung 4.19. L^2 kann keine Basis $\{b_1, b_2, \ldots\}$ haben in dem Sinne, dass jedes $f \in L^2$ dargestellt werden kann als Linearkombination der b_k: $f = \sum_{k\in\mathbb{N}} \beta_k b_k$ klappt im Allgemeinen nicht (vgl Beispiele 4.6). Aber wegen Satz 4.16 sind die e_k von oben eine ON-Basis in dem Sinne, dass

$$\|f - \sum_{k\in\mathbb{N}} \alpha_k e_k\|_2 = 0, \quad (\alpha_k = \langle f, e_k\rangle). \tag{4.13}$$

Das passt perfekt zu unserer Sichtweise, nach der wir Funktionen f, g als gleich auffassen, falls gilt: $\|f - g\|_2 = 0$. In dem Sinne bezeichnen wir $\{e_1, e_2, \ldots\}$ als ON-Basis von L^2.

Nun wissen wir aus Satz 4.16, dass (4.13) gilt. Damit wird aber die linke Seite der Besselschen Gleichung (siehe 4.17) gleich 0. Also gilt:

Satz 4.20 (Parsevalsche Gleichung)**.** *Für jedes $f \in L^2$ gilt:*

$$\sum_{k=1}^{\infty} \langle f, e_k\rangle^2 = \|f\|^2, \quad \textit{also} \quad \frac{1}{2}a_0^2 + \sum_{k=1}^{\infty} (a_k^2 + b_k^2) = \frac{1}{\pi}\|f\|,$$

wobei a_k, b_k die Fourierkoeffizienten von f sind (vgl Def 2.1).

Abschließend noch zwei Bemerkungen.

Zu $L^p([a,b])$: Für jedes $1 \leq p < \infty$ kann man die entsprechende Norm definieren:

$$\|f\|_p := \Big(\int_a^b |f(x)|^p dx\Big)^{1/p}$$

und erhält einen Banachraum. Aber nur für $p = 2$ gehört diese Norm zu einem Innenprodukt, also hat man nur für $p = 2$ auch einen Hilbertraum. Trotzdem geht für $1 < p < \infty$ sonst fast alles hier Diskutierte gut. Für $p = 1$ wird vieles anders; auch dort gibt es eine ausgefeilte Theorie, aber keine so schöne wie für $p = 2$.

Zu $L^2([-\infty,\infty])$: Hier ist das Skalarprodukt gegeben durch

$$\langle f,g\rangle = \int_{-\infty}^{\infty} f(x)g(x)dx.$$

Damit wird auch $L^2([-\infty,\infty])$ ein Hilbertraum. Eine ON-Basis (also ein maximales ONS) ist gegeben durch die *Hermite-Funktionen* Ψ_n:

$$\Psi_n(x) = \frac{1}{\sqrt{2^n n!\sqrt{\pi}}} e^{-x^2/2} H_n(x),$$

wobei

$$H_n(x) = (-1)^n e^{x^2} \frac{d}{dx^n} e^{-x^2}.$$

Die H_n sind dabei einfach Polynome vom Grad n:

$$H_0(x) = 1,\ H_1(x) = 2x, H_2(x) = 4x^2 - 2, H_3(x) = 8x^3 - 12x, H_4(x) = 16x^4 - 48x^2 + 12, \ldots$$

5 Fouriertransformation (FT)

> Die Neigung des Menschen, kleine Dinge für wichtig zu halten, hat sehr viel Großes hervorgebracht.
> G C Lichtenberg

Was wir jetzt bereits können: Gegeben ein Impuls (Signal, Welle) f. Falls es zusammengesetzt ist aus reinen Wellen (sin, cos), deren Frequenzen Vielfache voneinander sind (1,2,3,... bzw $\sin(x), \sin(2x)$ usw), dann liefern uns die Fourierkoeffizienten den jeweiligen Anteil (vgl etwa Bsp. 2.4). Wenn wir nun aber ganz allgemeine Wellen untersuchen, so können diese ja Kombinationen von reinen Wellen sein, die ganz verschieden sind ("inkommensurabel": auf Deutsch in etwa: nicht vergleichbar, auf mathematisch: in irrationalem Verhältnis stehend, wie etwa 1 und $\sqrt{2}$).

Wir wollen nun alle möglichen Werte $t \in \mathbb{R}$ zulassen. Dann haben wir nicht mehr abzählbar viele Koeffizienten, und können nicht mehr mit Reihen der Form $\sum_{k\in\mathbb{N}} a_k \cos(kx) + b_k \sin(kx)$ arbeiten. Stattdessen bieten sich Integrale an.

Überdies wird vieles systematischer, wenn wir zu komplexwertigen Funktionen übergehen. Es ist ja

$$e^{ix} = \cos(x) + i\sin(x)$$

Statt $a_k\cos(kx) + b_k\sin(kx)$ können wir also auch schreiben $c_k e^{ikx}$, wobei $c_k = a_k - ib_k$. Denn, mit $c = r + si$, ergibt sich für den Realteil von ce^{ix}:

$$\begin{aligned}\mathrm{Re}(ce^{ix}) &= \mathrm{Re}\left((r+si)(\cos(x)+i\sin(x))\right)\\ &= \mathrm{Re}(r\cos(x) + si\cos(x) + ri\sin(x) - s\sin(x)) = r\cos(x) - s\sin(x).\end{aligned}$$

(Aber Obacht, unten (ab Lemma 5.14) benutzen wir zweiseitig unendliche Reihen, und daher letztendlich andere c_k). Denkt man das konsequent durch, führt das zu folgender Definition.

Definition 5.1. Sei $f : \mathbb{R} \to \mathbb{C}$ integrierbar über R, d.h. $f \in L^1(\mathbb{R}) := \{f : \mathbb{R} \to \mathbb{C} \mid \int_{\mathbb{R}} |f(t)|dt < \infty$. Die Fouriertransformierte (FT) $\widehat{f}$ von f ist gegeben durch

$$\widehat{f}(k) = \int_{\mathbb{R}} f(x)e^{-ikx}dx$$

Manchmal ist es auch praktisch, die FT von f zu schreiben als $FT(f)$.

Obacht: Die Definition der FT ist bei weitem nicht einheitlich! In vielen Büchern steht das obige mit einem Vorfaktor $\frac{1}{2\pi}$ oder $\frac{1}{\sqrt{2\pi}}$, oder aber im Exponenten kommt noch ein Faktor 2π dazu.

Satz 5.2. *Seien $f, g \in L^1$, $\alpha \in \mathbb{C}$, $a \in \mathbb{R}$. Für die FT gelten folgende Eigenschaften:*

1. **Linearität:** $\widehat{f+g} = \widehat{f} + \widehat{g}$, $\widehat{\alpha f} = \alpha \widehat{f}$.
2. **Translationsformel:** *Ist $g(x) = f(x-a)$, dann ist $\widehat{g}(k) = e^{-iak}\widehat{f}(k)$.*
3. **Modulationsformel:** *Ist $g(x) = e^{iax}f(x)$, dann ist $\widehat{g}(k) = \widehat{f}(k-a)$.*
4. **Streckungsformel:** *Ist $g(x) = f(ax)$, dann ist $\widehat{g}(k) = \frac{1}{|a|}\widehat{f}\left(\frac{k}{a}\right)$.*
   Für $a = -1$ erhält man mit $g(x) = f(-x)$ daraus: $\widehat{g}(k) = \widehat{f}(-k)$.
5. **Konjugationsformel:** $\widehat{\overline{f}}(k) = \overline{\widehat{f}(-k)}$.
   Falls f nur reelle Werte hat, wird daraus $\widehat{f}(-k) = \overline{\widehat{f}(k)}$.

Proof. Zu (a):

$$\widehat{f+g}(k) = \int_{\mathbb{R}} (f(x)+g(x))e^{-ikx}dx = \int_{\mathbb{R}} f(x)e^{-ikx}dx + \int_{\mathbb{R}} g(x)e^{-ikx}dx = \widehat{f}(k) + \widehat{g}(k),$$

und

$$\widehat{\alpha f}(k) = \int_{\mathbb{R}} \alpha f(x)e^{-ikx}dx = \alpha \int_{\mathbb{R}} f(x)e^{-ikx}dx = \alpha \widehat{f}(k).$$

Zu (b):

$$\widehat{g}(k) = \int_{\mathbb{R}} f(x-a)e^{-ikx}dx = \int_{\mathbb{R}} f(y)e^{-ik(y+a)}dy = e^{-iak}\widehat{f}(k).$$

Zu (c),(d),(e): Aufgabe 21, Blatt 6. □

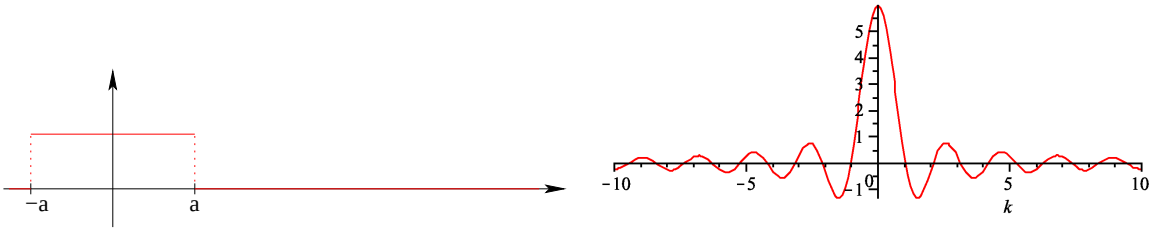


Figure 5: Die Rechtecksfunktion $1_{[-a,a]}$ und ihre FT (hier für $a = 3$)).

Beispiel 5.3. (a) Rechtecksfunktion $f:\mathbb{R}\to\mathbb{R}$, $f(x)=1_{[-a,a]}(x)=\begin{cases}1 & \text{für } x\in[-a,a]\\ 0 & \text{sonst.}\end{cases}$

(Obacht, nicht "Rechtecksschwingung", die ist periodisch, diese hier nicht.)

Es ist

$$\widehat{f}(k)=\int_{-a}^{a}1\cdot e^{-ikx}dx=(\frac{-1}{ik}e^{-ikx}\Big|_{x=-a}^{a})=\frac{-1}{ik}(e^{-ika}-e^{ika})=\frac{2}{k}\sin ka.$$

(Recall: $e^{ix}-e^{-ix}=2i\sin(x)$; vgl Aufgabe 17, Blatt 5).

(b) Die Gaußkurve der Standardnormalverteilung (siehe W-theorie) ist ihre eigene FT, bis auf einen konstanten Faktor. Genauer: Mit $f(x)=e^{-x^2/2}$ ist

$$\widehat{f}=\sqrt{2\pi}f.$$

[HIER STAND EIN ELEMENTARER BEWEIS DIESER AUSSAGE. DER WAR LEIDER FALSCH. EIN KORREKTER BEWEIS IST TECHNISCH ANSPRUCHSVOLL (BENUTZT Z.B. EINDEUTIGKEITSSÄTZE FÜR LÖSUNGEN VON DIFFERENTIALGLEICHUNGEN, ODER ERZEUGENDE FUNKTIONEN) UND FINDET SICH ETWA IN [PIN], ABSCHNITT 2.4.4.2.]

Allgemeiner gilt: Ist $f(x)=e^{-(x/\alpha)^2/2}$, dann ist $\widehat{f}(k)=\alpha\sqrt{2\pi}e^{-\alpha^2k^2/2}$.

Unter milden Voraussetzungen bekommen wir aus $\widehat{f}$ die ursprüngliche Funktion f zurück.

Satz 5.4 (Umkehrformel)**.** *Es seien $f\in L^1$ und $\widehat{f}\in L^1$. Dann gilt für jede Stetigkeitsstelle x von f*

$$f(x)=\frac{1}{2\pi}\int_{\mathbb{R}}\widehat{f}(k)e^{ikx}dk.$$

Proof. Wir wissen (Aufgabe 22, Blatt 6):

$$\int_{\mathbb{R}}f(k)\widehat{g}(k)dk=\int_{\mathbb{R}}\widehat{f}(k)g(k)dk \tag{5.1}$$

Setzen wir $g(k)=e^{-(k/\alpha)^2/2}$, wissen wir aus Bsp 5.3 (b):

$$\widehat{g}(k)=\sqrt{2\pi}\alpha e^{-\alpha^2k^2/2}.$$

In obige Gleichung eingesetzt liefert das

$$\int_{\mathbb{R}}\widehat{f}(k)e^{-(k/\alpha)^2/2}dk=\sqrt{2\pi}\int_{\mathbb{R}}\alpha f(k)e^{-\alpha^2k^2/2}dk=\sqrt{2\pi}\int_{\mathbb{R}}f(\frac{t}{\alpha})e^{-t^2/2}dt\quad(t=\alpha k).$$

Für $\alpha\to\infty$ ergibt sich daraus wg f stetig:

$$\int_{\mathbb{R}}\widehat{f}(k)dk=\sqrt{2\pi}\int_{\mathbb{R}}f(0)e^{-t^2/2}dt=2\pi f(0).$$

Das ist die Behauptung für $x=0$. Für beliebiges x folgt die Behauptung dann aus der Anwendung des obigen auf $h(t):=f(t+x)$ unter Benutzung der Translationsformel (und ein bisschen rechnen). □

Diese Operation bezeichnen wir auch als *inverse FT* (IFT). Bezeichnung:

$$\check{f}(x)=IFT(f)(x)=\frac{1}{2\pi}\int_{\mathbb{R}}f(k)e^{ikx}dk.$$

Damit lässt sich die Umkehrformel auch so formulieren:

$$\boxed{\check{\widehat{f}} = \widehat{\check{f}} = f.}$$

Insbesondere sagt die Umkehrformel, dass $f \in L^2$ durch $\widehat{f}$ eindeutig bestimmt ist (*).

Satz 5.5 (Plancherel)**.**

$$\int_{\mathbb{R}} \widehat{f}(k)\overline{\widehat{g}(k)}dk = 2\pi \int_{\mathbb{R}} f(x)\overline{g(x)}dx$$

Wir benutzen zum Beweis folgendes Ergebnis.

Lemma 5.6. $\widehat{\widehat{f}}(x) = 2\pi f(-x)$

Proof. Wegen (*) ist $\widehat{\widehat{f}}(x) = 2\pi f(-x)$ gleichbedeutend mit

$$\check{\widehat{\widehat{f}}}(x) = 2\pi \check{f}(-x), \quad \text{also} \widehat{f}(x) = 2\pi \check{f}(-x),$$

das heißt $\int_{\mathbb{R}} f(x)e^{-ikx}dx = 2\pi\frac{1}{2\pi}\int_{\mathbb{R}} f(-x)e^{ikx}dx$, was offenbar stimmt. □

Proof. (Plancherel) In der Gleichung (5.1) setzen wir $h = \overline{\widehat{g}}$. Damit erhalten wir

$$\int_{\mathbb{R}} \widehat{f}(k)\overline{\widehat{g}}(k)dk = \int_{\mathbb{R}} \widehat{f}(k)h(k)dk = \int_{\mathbb{R}} f(x)\widehat{h}(x)dx = \int_{\mathbb{R}} f(x)\widehat{\overline{\widehat{g}}}(x)dx,$$

und das ist mit der Konjugationsformel gleich

$$\int_{\mathbb{R}} f(x)\overline{\widehat{\widehat{g}}}(-x)dx$$

und nach Lemma 5.6 gleich

$$2\pi \int_{\mathbb{R}} f(x)\overline{g}(x)dx$$

□

Dieses Ergebnis bedeutet: Die durch das Skalarprodukt

$$\langle f, g\rangle = \int_{\mathbb{R}} f(x)\overline{g(x)}dx$$

gegebene Norm auf L^2 ist ja $\|f\| = \sqrt{\langle f, f\rangle}$. Plancherel sagt, dass die Norm von $\widehat{f}$ genau $\sqrt{2\pi}$ mal die Norm von f ist. Damit kann man zeigen, dass die FT eine bijektive Abbildung von L^2 auf L^2 ist (Details siehe [PIN]). Falls die FT definiert wird wie oben, nur mit dem Vorfaktor $\frac{1}{\sqrt{2\pi}}$, dann ist die FT sogar eine *Isometrie* (eine abstandserhaltende Abbildung, so wie eine Drehung oder eine Spiegelung im $\mathbb{R}^d$).

Lemma 5.6 bedeutet ja auch:

$$\widehat{\widehat{\widehat{f}}} = 4\pi^2 f$$

Also liefert viermalige Anwendung der FT das ursprüngliche, bis auf Skalierung um $4\pi^2$. Hätten wir die FT mit einem Vorfaktor $\frac{1}{\sqrt{2\pi}}$ definiert, wäre sogar $FT^4(f) = f$ für alle $f \in L^2$.

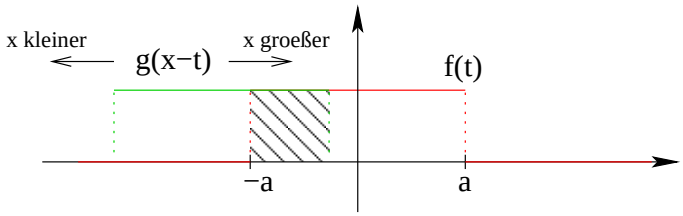


Figure 6: Die Faltung $f*g$ für $f=g$ die Rechteckfunktion. $f*g(x)$ ist die Fläche des Überlapps der beiden Graphen, wobei der Graph von g (gespiegelt wird und) um x verschoben wird.

Satz 5.7 (Riemann-Lebesgue)**.** *Ist $f \in L^1$ und stetig, dann gilt $\widehat{f}(k) \to 0$ für $|k| \to \infty$.*

Proof. Es ist

$$\widehat{f}(k) = \int_{\mathbb{R}} f(x)e^{-ikx}dx = -e^{-\pi ik/k}\int_{\mathbb{R}} f(x)e^{-ikx}dx = -\int_{\mathbb{R}} f(x)e^{-ik(x+\pi/k)}dx = -\int_{\mathbb{R}} f(x-\frac{\pi}{k})e^{-ikx}dx,$$

also ist

$$\widehat{f}(k) = \frac{1}{2}\widehat{f}(k) + \frac{1}{2}\widehat{f}(k) = \frac{1}{2}\int_{\mathbb{R}} \big(f(x) - f(x-\frac{\pi}{k})\big)e^{-ikx}dx.$$

Für $k \to \infty$ geht $f(x) - f(x-\frac{\pi}{k})$ gegen 0 (da f stetig). □

Ein wichtiges Ergebnis: Der Faltungssatz. Dazu erinnere man sich an MMfBW I (zahlentheoretische Faltung) und MMfBW II (Verteilungsdichte der Summe zweier unabhängiger Zufallsvariablen = Faltung der zwei Verteilungsdichten).

Definition 5.8. Die Faltung zweier Funktionen $f, g \in L^1$ ist

$$f*g(x) = \int_{\mathbb{R}} f(t)g(x-t)dt$$

Die Faltung beschreibt häufig Ereignisse in der Natur. Z.B. ist der Schatten eines Gegenstands, der von einer Lichtquelle beleuchtet wird, die Faltung der Form des Gegenstands mit der Form der Lichtquelle. Unscharfe Fotos entstehen durch falsche Fokussierung, also bildlich: eine falsche Linse. Das unscharfe Foto ist dann die Faltung des scharfen Fotos mit der Form der Linse. Dazu muss man natürlich mehrdimensional arbeiten.

Beispiel 5.9. Sei $f=g$ die Rechtecksfunktion $f(x) = 1_{[-1,1]}$. Dann ist

$$f*f(x) = \int_{\mathbb{R}} 1_{[-1,1]}(t)1_{[-1,1]}(x-t)dt = \int_{-1}^{1} 1_{[-1,1]}(x-t)dt = \int_{-1}^{1} 1_{[-1,1]}(t-x)dt = \int_{-1}^{1} 1_{[-1+x,1+x]}(t)dt.$$

Die Variable t läuft also nur von -1 bis 1: Interessant also nur $t \in [-1,1]$. Was passiert für $x < -2$? Dann ist $x-t < -2+1 = -1$, also $1_{[-1,1]}(x-t) = 0$.
Was passiert für $x > 2$? Dann ist $x-t > 2-1 = 1$, also $1_{[-1,1]}(x-t) = 0$. Interessant nur $x \in [-2,2]$.
Sei zunächst $x \in [-2,0]$. Dann ist $-1+x \leq -1$, also in diesem Fall

$$\int_{-1}^{1} 1_{[-1+x,1+x]}(t)dt = \int_{-1}^{1} 1_{[-1,1+x]}(t)dt = \int_{-1}^{1+x} 1(t)dt = t\Big|_{t=-1}^{1+x} = 2+x.$$

Ganz analog ist für $x \in [0,2]$

$$\int_{-1}^{1} 1_{[-1+x,1+x]}(t)dt = 2-x.$$

Damit ist $f * f$ eine Dreiecksfunktion auf $[-2,2]$ mit einer Spitze der Höhe 2 bei 0.

Die Faltung hat nette Eigenschaften:

Lemma 5.10. *1.* $f * g = g * f$ *(Symmetrie)*

2. $f * (g * h) = (f * g) * h$ *(Assoziativität)*

3. $f * (g + h) = f * g + f * h$ *(Distributivität)*

Proof. Das ist einfach:

$$f * g(x) = \int_{\mathbb{R}} f(t)g(x-t)dt = \int_{\mathbb{R}} f(x-y)g(y)dy = g * f(x);$$

und alle Funktionen sind in L^1, also *absolut integrierbar*, dann darf man die Integrale vertauschen (Satz von Fubini):

$$\begin{aligned} f * (g * h)(x) &= \int_{\mathbb{R}} f(t)(g * h)(x-t)dt = \int_{\mathbb{R}} f(t) \int_{\mathbb{R}} g(s)h(x-t-s)dsdt = \int_{\mathbb{R}} \int_{\mathbb{R}} f(t)g(s)h(x-t-s)dtds \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} f(t)g(r-t)h(x-r)dtdr = \int_{\mathbb{R}} \int_{\mathbb{R}} f(t)g(r-t)dt\, h(x-r)dr = \\ &\int_{\mathbb{R}} (f * g)(r)h(x-r)dr = (f * g) * h \end{aligned}$$

Die Distributivität ist noch einfacher, das schreiben wir gar nicht erst auf. □

Satz 5.11 (Faltungssatz). *Für* $f, g \in L^1$ *gilt:* $\widehat{f * g} = \widehat{f} \cdot \widehat{g}$.

Das Schöne ist: Damit kann man Faltung (im Zeit-Raum bzw x-Raum) in Multiplikation (im Frequenzraum bzw k-Raum) übersetzen, und umgekehrt. Das ist z.B. Grundlage für schnelle Multiplikation (das wird ausführlich behandelt in Abschnitt 6.4, 40): Naives Multiplizieren zweier Binärzahlen der Länge n braucht $O(n^2)$ Operationen. FT auf endlichen Mengen braucht nur $O(n \log n)$ (wenn man's richtig macht), und falten auch nur $O(n \log n)$ (wenn man's richtig macht). Also statt $f \cdot g$ berechne

$$\widehat{\check{f} * \check{g}} = f \cdot g, \tag{5.2}$$

(der Faltungssatz gilt entsprechend für IFT) in $O(n \log n)$ Schritten.

Proof.

$$\begin{aligned} \widehat{f * g}(k) &= \int_{\mathbb{R}} f * g(x)e^{-ikx}dx = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x-t)g(t)\, dt\, e^{-ikx}dx = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x-t)e^{-ikx}\, dx\, g(t)dt \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} f(y)e^{-ik(y+t)}\, dy\, g(t)dt = \int_{\mathbb{R}} \int_{\mathbb{R}} f(y)e^{-iky}\, dy\, g(t)e^{-ikt}dt \\ &= \int_{\mathbb{R}} f(y)e^{-iky}\, dy \int_{\mathbb{R}} g(t)e^{-ikt}dt = \widehat{f}(k) \cdot \widehat{g}(k) \end{aligned}$$

□

Generelle Anmerkung:

Bemerkung 5.12. Praktisch alles in diesem Kapitel geht genauso für Funktionen $f : \mathbb{R}^d \to \mathbb{C}$. Die FT ist dann

$$\widehat{f}(\mathbf{k}) = \int_{\mathbb{R}^d} f(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} d\mathbf{x}.$$

Dabei sind $\mathbf{k} = (k_1, \ldots, k_d)$ und $\mathbf{x} = (x_1, \ldots, x_d)$ Vektoren in $\mathbb{R}^d$, das Integral ist also ein Mehrfachintegral $\int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{R}} \cdots\cdots dx_3 dx_2 dx_1$, und $\mathbf{k} \cdot \mathbf{x}$ ist das Standardskalarprodukt in $\mathbb{R}^d$, also

$$\mathbf{k} \cdot \mathbf{x} = k_1 x_1 + \cdots + k_d x_d$$

Damit übertragen sich alle (soweit ich weiß) Ergebnisse hier auf den mehrdimensionalen Fall. Auf die Vorfaktoren 2π muss man allerdings aufpassen, daraus wird üblicherweise $(2\pi)^d$.

Bemerkung 5.13. Vergleicht man die Kap. 2-4 mit Kap. 5, so fallen Parallelen auf: Es gibt z.B. beidesmal eine Parsevalsche Gleichung und einen Satz von Riemann-Lebesgue. Das ist kein Zufall. Begibt man sich auf eine noch abstraktere Warte als wir es hier tun, so kann man FT definieren für lokalkompakte abelsche topologische Gruppen (LKAG).

> Topologische Gr. heißt: Gruppe mit einer Regel, was offene Mengen sein sollen. Die Regel muss konsistent sein unter verschiedenen Bedingungen: Vereinigung zweier offenen Mengen ist wieder offen etc. Alle Hilberträume sind top Gruppen, denn sie haben eine Metrik ("Abstand"), und alle metrischen Gruppen sind top Gruppen. Damit kann man dann erklären, was kompakte Mengen sind. "Lokalkompakt" heißt dann einfach, dass jedes Element der Gruppe eine kompakte Umgebung besitzt.

Für LKAG ist immer eine *duale* Gruppe erklärt. So ist die duale Gruppe von $\mathbb{R}$ gerade (isomorph zu) $\mathbb{R}$. Aber auch $[-\pi, \pi]$ ist eine lokalkomp abelsche Gruppe, wenn man die beiden Endpunkte gleichsetzt ("aneinander klebt"). Die heißt (eindim) Torus. Und die duale Gruppe davon ist gerade (isomorph zu) $\mathbb{Z}$.
Treibt man also FT allgemein auf LKAG, so lebt die FT von f auf der dualen Gruppe. Bei $\mathbb{R}$ ist das dieselbe: $\mathbb{R}$. Beim Torus ist es aber $\mathbb{Z}$. Und die FR, gelesen als $(\ldots b_3, b_2, b_1, a_0, a_1, a_2 \ldots)$ lebt auf $\mathbb{Z}$! Wir könnten also alles in einem einheitlichen Rahmen betrachten. Das ist für uns aber wenig hilfreich, des Verständnisses und der Anwendungen wegen.

5.1 Poissons Summenformel (PSF)

In diesem Abschnitt bringen wir FT und FR zusammen; oder in anderen Worten: FT auf $\mathbb{R}$ (Kap. 5) und FT auf dem Torus (Kap. 2-4; wir hatten dort FR behandelt, aber von einer abstrakten Warte aus ist das die FT auf dem Torus, vgl obige Bemerkung). Wir bleiben aber, im Ggs zur letzten Bemerkung, recht konkret.

Lemma 5.14. *Die komplexe FR von $f \in L^1$ ist $\sum_{k\in\mathbb{Z}} c_k e^{ixk}$, wobei*

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ikx} dx.$$

Proof. Vergleichen wir die c_k mit den alten Formeln für a_n, b_n (vor Definition 2.1) und vergleichen den k-ten Term der FR:

$$\begin{aligned} c_k + c_{-k} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)(\cos(kx) + i\sin(kx))dx + \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)(\cos(-kx) + i\sin(-kx))dx \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)(\cos(kx) + \cos(-kx))dx + i\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)(\sin(kx) + \sin(-kx))dx \\ &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(kx)dx + i\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\cdot 0 dx \\ &= a_k + 0 \end{aligned}$$

und analog $i(c_k - c_{-k}) = b_k$. Damit stimmen die FR überein: Betrachtet man den k-ten Summanden der "alten" FR, so entspricht der dem k-ten plus dem $-k$-ten der "neuen" FR. Damit ergibt sich

$$\begin{aligned} c_k e^{ikx} + c_{-k}e^{-ikx} &= c_k(\cos(kx) + i\sin(kx)) + c_{-k}(\cos(-kx) + i\sin(-kx)) \\ &= c_k(\cos(kx) + i\sin(kx)) + c_{-k}(\cos(kx) - i\sin(kx)) \\ &= (c_k + c_{-k})\cos(kx) + i(c_k - c_{-k})\sin(kx) \end{aligned}$$

□

Unter bestimmten Bedingungen gilt $\sum_{n\in\mathbb{Z}} f(n) = \sum_{n\in\mathbb{Z}} \widehat{f}(2\pi n)$. Das sind diese:

Satz 5.15 (PSF). *Sei $f \in L^1$ beschränkt, stetig und stückweise stetig diff-bar. Außerdem seien $x^2 f(x)$ und $x^2 f'(x)$ beschränkt. Dann gilt*

$$\sum_{n\in\mathbb{Z}} f(n) = \sum_{n\in\mathbb{Z}} \widehat{f}(2\pi n)$$

Proof. Weil $x^2 f(x)$ beschränkt ist, gilt folgendes: Sei $x \in \mathbb{R}$ festgelegt. Dann gibt's eine Konstante C so dass für alle $n \in \mathbb{Z}$: $(x+n)^2|f(x+n)| < C$. Dann ist $|f(x+n)| \le \frac{C}{(x+n)^2}$. Also konvergiert die Reihe $g(x) = \sum_{n\in\mathbb{Z}} f(x+n)$ gleichmäßig, daher ist g eine stetige Funktion.

Genauso ist $h(x) = \sum_{n\in\mathbb{Z}} f'(x+n)$ eine stetige Funktion. Nach dem Hauptsatz der Analysis ist $f(x) = \int_0^x f'(t)dt + a$ für ein $a \in \mathbb{R}$. Damit ist

$$\begin{aligned} \int_0^x h(t)dt &= \int_0^x \sum_{n\in\mathbb{Z}} f'(t+n)dt = \sum_{n\in\mathbb{Z}} \int_0^x f'(t+n)dt \\ &= \sum_{n\in\mathbb{Z}} \int_n^{n+x} f'(t)dt = \sum_{n\in\mathbb{Z}} f(n+x) - f(n) = g(x) - g(0). \end{aligned}$$

Hier durften wir Summe und Integral vertauschen, da die Reihe gleichmäßig konvergiert.
Damit ist also g das Integral einer stetigen (und integrierbaren) Funktion, also stetig und diff-bar. Nach Satz 3.6 konvergiert dann seine FR punktweise gegen g. D.h. insbesondere

$$\sum_{n\in\mathbb{Z}} f(n) = g(0) = \sum_{m\in\mathbb{Z}} c_m,$$

wobei g allerdings 1-periodisch und nicht 2π-periodisch war. Betrachten wir also die FR zu $g(\frac{x}{2\pi})$, dann gilt

$$\begin{aligned} c_m =& \frac{1}{2\pi}\int_{-\pi}^{\pi} g(\frac{x}{2\pi})e^{-imx}dx = \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{n\in\mathbb{Z}} f(\frac{x}{2\pi}+n)e^{-imx}dx = \sum_{n\in\mathbb{Z}}\frac{1}{2\pi}\int_{-\pi}^{\pi} f(\frac{x}{2\pi}+n)e^{-imx}dx \\ &= \sum_{n\in\mathbb{Z}}\frac{1}{2\pi}\int_{(2n-1)\pi}^{(2n+1)\pi} f(\frac{y}{2\pi})e^{-im(y-2\pi n)}dy \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} f(\frac{y}{2\pi})e^{-im(y-2\pi n)}dy = \frac{1}{2\pi}\int_{\mathbb{R}} f(\frac{y}{2\pi})e^{-imy}dy = \int_{\mathbb{R}} f(z)e^{-im2\pi z}dz = \widehat{f}(2\pi m) \end{aligned}$$

(dabei ist $2\pi z = y$.) Also ist $\sum_{n\in\mathbb{Z}} f(n) = g(0) = \sum_{n\in\mathbb{Z}} \widehat{f}(\pi n)$. □

Die PSF ist ein starkes Werkzeug zum Berechnen der Werte bestimmter unendlicher Reihen, vgl Aufgaben 29 und 31, Blatt 8.

6 DFT

Angenommen, wir haben nun nicht mehr Funktionen mit Definitionsbereich $\mathbb{R}$ oder $[-\pi,\pi]$, sondern welche mit einem endlichen Definitionsbereich:

$$f : \{0,1,\ldots,N-1\} \to \mathbb{C}$$

So eine Funktion kann man sich einfach als N-Vektor vorstellen: Die Funktion ist komplett beschrieben durch die Angabe all ihrer Funktionswerte, das sind N Stück. Also schreiben wir auch einfach $f = (f_0, f_1, \ldots, f_{N-1})$, mit $f_n = f(n)$. Auch dafür können wir FT definieren, diese heißt *diskrete Fouriertransformation* (DFT).

Definition 6.1. Die DFT von $f = (f_0, f_1, \ldots, f_{N-1})$ ist $d = (d_0, d_1, \ldots, d_{N-1})$ mit

$$d_k = \frac{1}{N}\sum_{j=0}^{N-1} f_j e^{-2\pi i jk/N} \quad (k = 0,1,\ldots,N-1)$$

Schreibweise: $\widehat{f} = d$, oder $DFT(f) = d$. Je nach Kontext schreiben wir statt d auch $\widehat{f} = (\widehat{f}_0, \widehat{f}_1, \ldots, \widehat{f}_{N-1})$. Guckt man sich die Defintion an, so sieht man, dass da eine Matrix-Vektor-Multiplikation steht (mit $\xi = e^{2\pi i/N}$) :

$$d = \begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_{N-1} \end{pmatrix} = \frac{1}{N}\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \xi^{-1} & \xi^{-2} & \cdots & \xi^{-(N-1)} \\ 1 & \xi^{-2} & \xi^{-4} & \cdots & \xi^{-2(N-1)} \\ 1 & \xi^{-3} & \xi^{-6} & \cdots & \xi^{-3(N-1)} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & \xi^{-(N-1)} & \xi^{-2(N-1)} & \cdots & \xi^{-(N-1)^2} \end{pmatrix}\begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_{N-1} \end{pmatrix} = \frac{1}{\sqrt{N}}\overline{V}f,$$

wobei $V = \frac{1}{\sqrt{N}}\left(\xi^{jk}\right)_{0\le j,k\le N-1}$. Den Vorfaktor $\frac{1}{N}$ haben wir in zweimal $\frac{1}{\sqrt{N}}$ aufgespalten, damit gilt:

Satz 6.2. *Die Matrix V ist unitär, d.h. $V\overline{V}^T = E$.*

Dabei bezeichnet E die $N \times N$-Einheitsmatrix. Erinnerung:

Eine Matrix $M \in \mathbb{R}^{N\times N}$ heißt *orthogonal*, falls $MM^T = E$. In $\mathbb{R}^2$ oder $\mathbb{R}^3$ sind das gerade die Matrizen, die eine Drehung oder eine Spiegelung bewirken. Orthogonale Matrizen bewirken abstands- und winkeltreue Abbildungen. Für das Standardskalarprodukt (siehe (5.2)) in $\mathbb{R}^d$ drückt sich das so aus: V ist orthogonal, genau dann wenn $Vx \cdot Vx = x \cdot x$. ("Die Länge von Vx ist gleich der Länge von x")

Eine Matrix $M \in \mathbb{C}^{N\times N}$ heißt *unitär*, falls $M\overline{M}^T = E$. Für relle Matrizen heißt das dasselbe wie orthogonal. Für das Standardskalarprodukt im $\mathbb{C}^d$ drückt sich das so aus: V unitär, genau dann wenn $Vx \cdot \overline{Vx} = x \cdot \overline{x}$.

Proof. Wir zeigen das zeilenweise. Es muss gelten: k-te Zeile mal k-te Zeile = 1, und k-te Zeile mal m-te Zeile = 0 für $m \neq k$. Bezeichnen wir mit V_k die k-te Zeile von V. Dann gilt:

$$V_k(\overline{V_k})^T = \frac{1}{N}\sum_{\ell=0}^{N-1} \xi^{\ell k}\overline{\xi^{\ell k}} = \frac{1}{N}\sum_{\ell=0}^{N-1} \xi^{\ell k}\xi^{-\ell k} = \frac{1}{N}\sum_{\ell=0}^{N-1} \xi^{(\ell-\ell)k} = \frac{1}{N}\sum_{\ell=0}^{N-1} 1 = 1,$$

und für $k \neq m$:

$$V_k(\overline{V_m})^T = \frac{1}{N}\sum_{\ell=0}^{N-1} \xi^{\ell k}\xi^{-\ell m} = \frac{1}{N}\sum_{\ell=0}^{N-1} (\xi^{k-m})^{\ell} = \frac{1}{N}\frac{\xi^{(k-m)N}-1}{\xi^{k-m}-1}.$$

Der Zähler im letzten Term ist $1-1=0$, da ξ eine komplexe N-te Einheitswurzel ist [WIK]. Der Nenner ist ungleich null, da $-N < k-m < N$ und $k-m \neq 0$; aber ξ^n wird nur 1, falls n Vielfaches von N ist. Also ist für $m \neq k$

$$V_k(\overline{V_m})^T = 0$$

und die Behauptung ist bewiesen. □

Bemerkung 6.3. Wegen dieses Satzes ist $V^{-1} = \overline{V}$. Die Inverse von V ist also effizient berechenbar.

Die Umkehrung der DFT, die *inverse DFT*, ist definiert durch

$$IDFT(d) = \check{d} = \sqrt{N}Vd$$

Damit bekommt man (natürlich!) aus jedem $d = \widehat{f}$ das f zurück. Hier sind's ja nur Matrix-Vektor-Multiplikationen, um Existenz von Integralen braucht man sich keine Sorgen zu machen.

Satz 6.4. *Für V wie oben gilt: $V^4 = E$, wobei E die Einheitsmatrix bezeichnet.*

Daraus folgt direkt: $\boxed{\widehat{\widehat{\widehat{f}}} = \frac{1}{N^2} f}$

Also liefert, wie im kontinuierlichen Fall, die viermalige Anwendung der Fouriertransformation das f zurück, bis auf einen Vorfaktor (damals 2π, hier $\frac{1}{N^2}$).

Weiterhin gelten etliche aus dem kontinuierlichen Falle bekannte Beziehungen. Zum Beispiel gelten für $f = (f_0, f_1, \ldots, f_{N-1}) \in \mathbb{C}^N$, $g = (g_0, g_1, \ldots, g_{N-1}) \in \mathbb{C}^N$ und deren DFTs $\widehat{f} = (\widehat{f_0}, \widehat{f_1}, \ldots, \widehat{f_{N-1}})$, $\widehat{g} = (\widehat{g_0}, \widehat{g_1}, \ldots, \widehat{g_{N-1}})$ folgende Aussagen:

1. **Faltungssatz:** $\widehat{f * g} = N\widehat{f} \cdot \widehat{g}$. Dabei ist

$$f * g(n) = \frac{1}{N}\sum_{k=0}^{N-1} f_k g_{n-k} \qquad (0 \leq n \leq N-1)$$

Der Index $n-k$ von g_{n-k} kann negativ werden, also legen wir hier für $1 \leq k \leq N-1$ fest: $g_{-k} = g_{N-k}$.

2. **Satz von Parseval:**
$$\sum_{n=0}^{N-1} |d_n|^2 = \frac{1}{N} \sum_{n=0}^{N-1} |f_n|^2$$

3. **Satz von Plancherel:**
$$\sum_{n=0}^{N-1} \widehat{f}_n \overline{\widehat{g}}_n = \frac{1}{N} \sum_{n=0}^{N-1} f_n \overline{g}_n$$

4. **Darstellung als FR:** Die FR von f beschreibt f:
$$f_n = \sum_{k=0}^{N-1} \widehat{f}_k e^{2\pi i k n/N}$$

Zu den Beweisen der ersten drei Aussagen siehe die Aufgaben 34, 35 (Blatt 9) und 36 (Blatt 10). Bezüglich der letzten stellen wir fest: Es gilt $\widehat{f} = \frac{1}{\sqrt{N}} \overline{V} f$, also

$$f = \sqrt{N} \overline{V}^{-1} \widehat{f} = \sqrt{N} V \widehat{f},$$

das heißt für den n-ten Eintrag $f_n = \sum_{k=0}^{N-1} \widehat{f}_k e^{2\pi i k n/N}$.

6.1 Trigonometrische Interpolation

Wir wissen bereits, wie man mittels der FR periodische Funktionen gut approximieren kann (Kap 2.1). Zum Bestimmen der Fourierkoeffizienten mussten wir aber Integrale berechnen. Nun versuchen wir es mit den diskreten FR von oben, für deren Koeffizienten müssen wir nur endliche Summen berechnen. Außerdem betrachten wir eine etwas andere Variante des Problems: statt Approximation Interpolation.

Interpolation heißt: Zu einem gegebenen Datensatz von Punkten $(x_0, y_0), \ldots, (x_{N-1}, y_{N-1})$ finde eine Funktion p, so dass $p(x_k) = y_k$ gilt für alle $0 \leq k \leq N-1$. Diese Daten könnten z.B. Messdaten sein, oder Punkte einer komplizierten Funktion f. Ist der Abstand zwischen benachbarten x-en (den *Stützstellen*) immer gleich (gilt also $|x_k - x_{k+1}| = h$ für alle $0 \leq k \leq N-2$), so heißen die Stützstellen *äquidistant*.

Wieder wollen wir, dass p ein trigonometrisches Polynom ist (s. Kap 2), also

$$p(x) = \sum_{k=0}^{N-1} a_k \cos(2\pi k x) + b_k \sin(2\pi k x).$$

Satz 6.5. *Sei $f : [0,1] \to \mathbb{R}$ eine Funktion. Setze $f_k = f(\frac{k}{N})$, und es sei $DFT(f_0, f_1, \ldots, f_{N-1}) = (d_0, d_1, \ldots, d_{N-1})$ die DFT von $(f_0, \ldots, f_{N-1})$. Dann gilt für $a_k = \mathrm{Re}(d_k), b_k = -\mathrm{Im}(d_k)$:*

$$p(x) = \sum_{k=0}^{N-1} a_k \cos(2\pi k x) + b_k \sin(2\pi k x)$$

hat die Interpolationseigenschaft, d.h., $p(\frac{k}{N}) = f_k = f(\frac{k}{N})$.

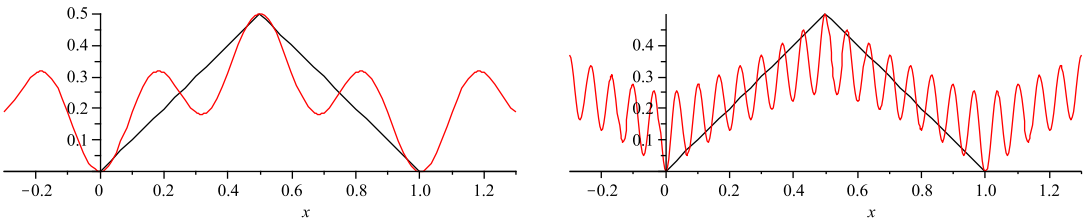


Figure 7: Trigonometrische Interpolation (roter bzw grauer Graph) einer Dreiecksfunktion (schwarzer Graph) mit Hilfe der DFT nach Satz 6.5, für $N = 4$ Stützstellen (links) und $N = 16$ (rechts). Zwischen den Stützstellen oszilliert der Graph stark.

Proof. Laut Def der IDFT ist

$$\begin{aligned} f_j &= \sum_{k=0} d_k(\cos(2\pi j\frac{k}{N}) + i\sin(2\pi j\frac{k}{N})) \\ &= \sum_{k=0} \mathrm{Re}(d_k)(\cos(2\pi j\frac{k}{N}) + i\sin(2\pi j\frac{k}{N})) + i\,\mathrm{Im}(d_k)(\cos(2\pi j\frac{k}{N}) + i\sin(2\pi j\frac{k}{N})), \end{aligned}$$

und da f reell ist, verschwinden die Imaginärteile, und es bleibt stehen

$$\sum_{k=0} \mathrm{Re}(d_k)\cos(2\pi j\frac{k}{N}) - \mathrm{Im}(d_k)\sin(2\pi j\frac{k}{N})$$

□

Es gibt aber ein Problem: Das ist eine perfekte Interpolation — wie der Satz zeigt — aber eine schlechte Approximation.

Beispiel 6.6. Nehmen wir eine Dreiecksfunktion wie in Bild 7 (schwarzer Graph). Approximieren wir wie oben beschrieben für $N = 4$, so erhalten wir für p den Graphen in Bild 7 links (rot bzw grau). Er trifft an den Stützstellen genau die Funktionswerte (muss er nach dem letzten Satz ja), aber dazwischen weicht er stark ab. Versuchen wir nun, die Qualität der Approximation zu verbessern, indem wir die Zahl der Stützstellen erhöhen, z.B. auf $N = 16$, so trifft der Graph des entsprechenden p die korrekten Funktionswerte an den Stützstellen, dazwischen aber oszilliert er nur um so stärker.

Ein anderer Ansatz liefert eine bessere Approximation:

$$r(x) = \sum_{k=-N/2}^{N/2-1} a_k\cos(2\pi kx) + b_k\sin(2\pi kx).$$

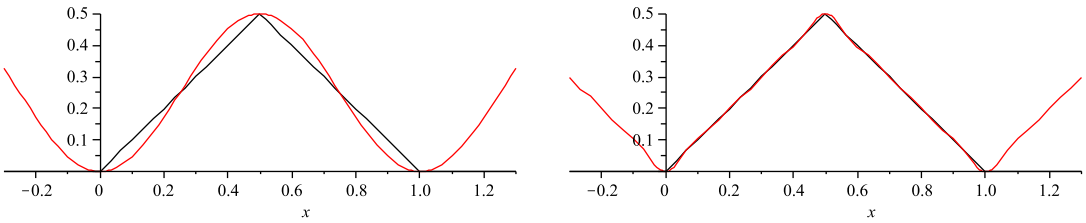


Figure 8: Trigonometrische Interpolation mit Hilfe der DFT nach Satz 6.7. Die Oszillation ist sehr gering.

Satz 6.7. *Sei* $f:[0,1]\to\mathbb{R}$ *eine Funktion. Setze* $f_k = f(\frac{k}{N})$*, und es sei* $(d_{-N/2}, d_{-N/2+1}, \ldots, d_{N/2-2}, d_{N/2-1})$ *die DFT von* $(f_0, -f_1, f_2, -f_3, \ldots, (-1)^{N-1} f_{N-1})$*. Dann gilt für* $a_k = \mathrm{Re}(d_k), b_k = -\mathrm{Im}(d_k)$*:*

$$r(x) = \sum_{k=-N/2}^{N/2-1} a_k \cos(2\pi kx) + b_k \sin(2\pi kx)$$

hat die Interpolationseigenschaft, d.h., $r(\frac{k}{N}) = f_k = f(\frac{k}{N})$.

Proof. Es ist

$$r(x) = \sum_{k=-N/2}^{N/2-1} d_k e^{2\pi ikx} = \sum_{k=0}^{N-1} d_{k-N/2} e^{2\pi i(k-N/2)x} = \sum_{k=0}^{N-1} d_{k-N/2} e^{2\pi ikx} e^{-\pi iNx}$$

und das ist für $x = x_n = \frac{n}{N}$ gleich

$$(-1)^n \sum_{k=0}^{N-1} d_{k-N/2} e^{2\pi ikx} = (-1)^n \big((-1)^n f_n\big) = f_n.$$

Mit Satz 6.5 folgt die Behauptung. □

Dieses trigonometrische Polynom liefert auch eine bessere Approximation, wie man z.B. in Bild 8 sieht. Das lässt sich auch rigoros nachweisen, vgl [PLA], Kap 3. Da steht i.Wes. eine Fehlerabschätzung für $\|p-f\|_2$, wo der Fehlerterm in unserem Bsp $O(1)$ ist (also i.Wes. konstant) ist, und für $\|r-f\|_2$ ist der Fehlerterm $O(N^{-m})$, falls f m-mal stetig diff-bar ist.

Bisher war unser f auf $[0,1]$ definiert. Daher können wir f zu einer geraden Funktion fortsetzen, indem wir f auf $[-1,0]$ definieren durch $f(x) = f(-x)$. Dann werden in obigem Ansatz alle Sinuskoeffizienten zu Null, nur die Cosinusterme überleben, und wir erhalten die (oder besser eine) **Diskrete Cosinus-Transformation (DCT)**. Wir haben aber noch weitere Variationsmöglichkeiten: Wir können

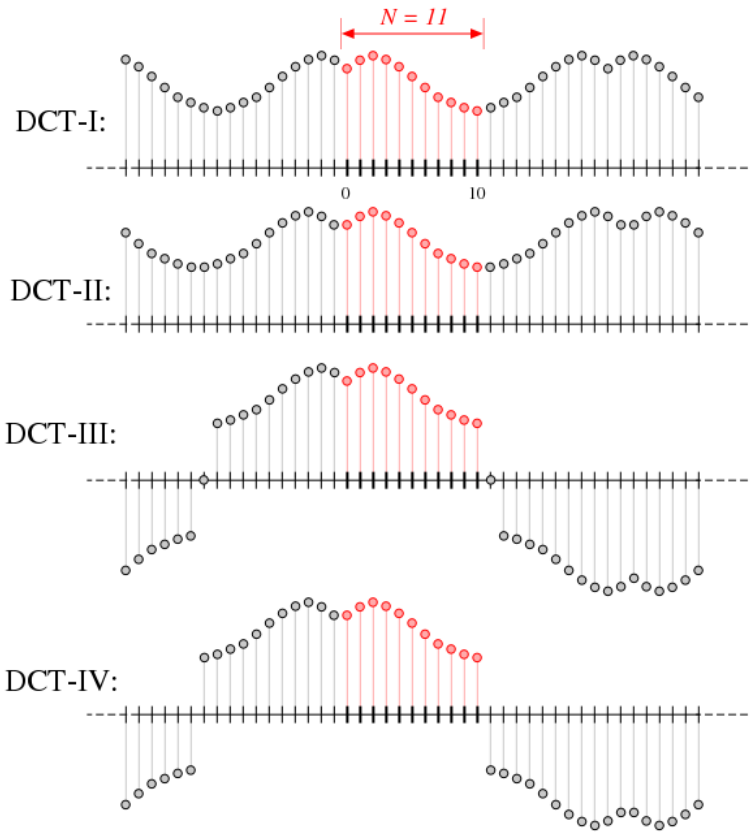


Figure 9: Vier Möglichkeiten zur Definition der DCT. (Quelle: [WIK])

die $f_0, f_1, f_2, \ldots f_{N-1}$ als Funktionswerte an den Stellen $0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}$ interpretieren. Das ist etwas unsymmetrisch: der linke Rand ist drin, der rechte nicht. Oder wir machen es symmetrischer, indem wir die Werte um $\frac{1}{2N}$ verschieben: $f_0, f_1, f_2, \ldots, f_{N-1}$ sind dann die Funktionswerte an den Stellen $\frac{1}{2N}, \frac{1}{N}+\frac{1}{2N}, \frac{2}{N}+\frac{1}{2N}, \ldots, \frac{N-1}{N}+\frac{1}{2N}$. Das liefert zwei Wahlmöglichkeiten. (Bzw vier, da wir die Option "Rand drin" für den rechten und den linken Rand unabhängig treffen können. Aber meist macht man es rechts und links gleich.)

Wir können auch noch festlegen, ob die Funktion nach rechts gerade oder ungerade fortgesetzt wird, siehe Bild 9. Das liefert also vier Ansätze, die DCT zu berechnen.

$$\text{DCT I:} \qquad a_k = \frac{1}{2}\left(f_0 + (-1)^k f_{N-1}\right) + \sum_{n=1}^{N-2} f_n \cos\left(\frac{\pi}{N-1} nk\right) \qquad k = 0, \ldots, N-1. \tag{6.1}$$

$$\text{DCT II:} \qquad a_k = \sum_{n=0}^{N-1} f_n \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right) \qquad k = 0, \ldots, N-1. \tag{6.2}$$

$$\text{DCT III:} \qquad a_k = \frac{1}{2} f_0 + \sum_{n=1}^{N-1} f_n \cos\left(\frac{\pi}{N} n\left(k+\frac{1}{2}\right)\right) \qquad k = 0, \ldots, N-1. \tag{6.3}$$

$$\text{DCT IV:} \qquad a_k = \sum_{n=0}^{N-1} f_n \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)\right) \qquad k = 0,\ldots,N-1. \tag{6.4}$$

Dabei ist häufig DCT II gemeint, wenn von "der DCT" die Rede ist.

6.2 Bildkompression

Die Idee zur Bildkompression ist diese: Wegen des Satzes von Riemann-Lebesgue erwarten wir, dass $a_k \to 0$ für $k \to \infty$. Nehmen wir also eine Bildzeile und interpretieren die Grauwerte des k-ten Pixels als Funktionswert f_k. Auf das so erhaltene f wenden wir DFT an. Von $DFT(f)$ speichern wir nur die ersten 20 % (oder so) der Werte. (Also nur 20% des Speicherbedarfs). Zum Angucken berechnen wir die IDFT, und hoffen, dass (da die letzten 80% nahe an Null sind, also fast Null, also nix ausmachen sollten) das so erhaltene Bild sehr nah am ursprünglichen Bild ist. Das klappt.

Eine erste Verbesserung dieser Idee finden wir in [GG]. Haben wir also ein Schwarzweißbild, können wir es zeilenweise bearbeiten. Jede Zeile hat N Pixel, und z.B. 256 mögliche Grauwerte.

(Genaueres in [GG], Kap. 13. Zu Run-length encoding und Huffman encoding siehe [WIK]. Das schreibe ich jetzt nicht noch mal hier hin. Die Idee in [GG] sowie run-length-encoding sollte man in der Prüfung erklären können.)

Die Vorgehensweise von jpeg ist nun diese: Zum Komprimieren:

- Bearbeite jeden der drei Farbwerte einzeln (RGB bzw. YCbCr)
- Unterteile das Bild in 8×8 Felder (evtl am Rand mit Nullen füllen)
- DCT für jedes einzelne Feld, erst zeilen- dann spaltenweise
- Quantisieren der DCT (s.u., dies ist die einzige Stelle, wo Information verloren geht)
- Run-length encoding, dann Huffman encoding

Ergebnis ist je eine Integer-Sequenz für jedes 8×8-Feld. Deren Länge liegt deutlich unter 256 Byte. Diese Sequenzen werden als `jpeg`-file gespeichert. Beim Angucken eines jpg-files werden die umgekehrten Schritte ausgeführt: Huffman-encoding rückwärts, run-length-encoding rückwärts (beides verlustfrei), IDCT, zusammensetzen der 8×8-Felder und der drei Farbwerte zum Gesamtbild.

Quantisieren heißt hier: Die Einträge der Matrix werden auf die nächste ganze Zahl gerundet, aber bezüglich einer Quantisierungsmatrix Q: Der Eintrag an der Stelle k,m wird durch $Q_{k,m}$ geteilt und dann gerundet. Unten ein Beispiel: Das erste ist die DCT für ein Feld. Der Eintrag oben links, also-415, wird durch 16 geteilt (-25,9375) und dann gerundet (-26). Die Matrix Q ist so gewählt, dass links oben, wo die Koeffizienten mit niedrigem Index stehen, relativ genau gerundet wird, und unten rechts, wo die Koeffizienten mit hohem Index stehen, recht grob gerundet wird. Hier eine typische 8x8-Matrix, wie sie nach der DCT entstehen könnte.

$$DCT: \begin{bmatrix} -415 & -30 & -61 & 27 & 56 & -20 & -2 & 0 \\ 4 & -22 & -61 & 10 & 13 & -7 & -9 & 5 \\ -47 & 7 & 77 & -25 & -29 & 10 & 5 & -6 \\ -49 & 12 & 34 & -15 & -10 & 6 & 2 & 2 \\ 12 & -7 & -13 & -4 & -2 & 2 & -3 & 3 \\ -8 & 3 & 2 & -6 & -2 & 1 & 4 & 2 \\ -1 & 0 & 0 & -2 & -1 & -3 & 4 & -1 \\ 0 & 0 & -1 & -4 & -1 & 0 & 1 & 2 \end{bmatrix}$$

Deren Einträge runden wir bzgl der Quantisierungsmatrix Q:

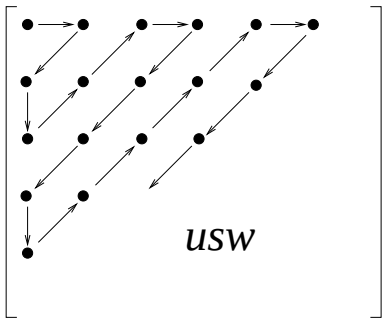
$$Q = \begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix}$$

Das ergibt in diesem Beispiel:

$$\begin{bmatrix} -26 & -3 & -6 & 2 & 2 & -1 & 0 & 0 \\ 0 & -2 & -4 & 1 & 1 & 0 & 0 & 0 \\ -3 & 1 & 5 & -1 & -1 & 0 & 0 & 0 \\ -4 & 1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Wenn der Rechner jemanden fragt, wie hoch der Kompressionsgrad sein soll, so wählt man damit eine gröbere (hohe Kompression) bzw feinere Matrix Q aus.

Nach dem Quantisieren steht da eine Matrix, die tendenziell unten rechts nur Nullen enthält. Auf diese Matrix wenden wir run-length encoding an, indem wir sie diagonal durchlaufen, nach dem unten dargestellten Schema. Auf die so erhaltene Sequenz wenden wir Huffman-encoding an.



Das Ergebnis kommt mit bedeutend weniger Speicherplatz aus als das ursprüngliche Bild.

6.3 FFT

Offenbar ist es wünschenswert, die DFT oder DCT effizient zu berechnen. Das geht so.

Satz 6.8. *Aus den DFTs von* $(f_0, f_1, \ldots, f_{N-1})$ *und* $(f_N, f_{N+1}, \ldots, f_{2N-1})$ *erhalten wir die DFT von* $DFT(f_0, f_N, f_1, f_{N+1}, f_2, f_{N+2}, \ldots, f_{N-1}, f_{2N-1})$ *(Länge 2N) so: Mit*

$$(d_0, d_1, \ldots, d_{N-1}) = \frac{1}{2}\big(DFT(f_0, f_1, \ldots, f_{N-1}) + e^{-\pi ik/N} DFT(f_N, f_{N+1}, \ldots, f_{2N-1})\big)$$

$$(d_N, d_{N+1}, \ldots, d_{2N-1}) = \frac{1}{2}\big(DFT(f_0, f_1, \ldots, f_{N-1}) - e^{-\pi ik/N} DFT(f_N, f_{N+1}, \ldots, f_{2N-1})\big)$$

ist dann

$$DFT(f_0, f_N, f_1, f_{N+1}, f_2, f_{N+2}, \ldots, f_{N-1}, f_{2N-1}) = (d_0, d_1, \ldots, d_{2N-1})$$

.

Proof. Für den k-ten Eintrag ($0 \le k \le N-1$) gilt:

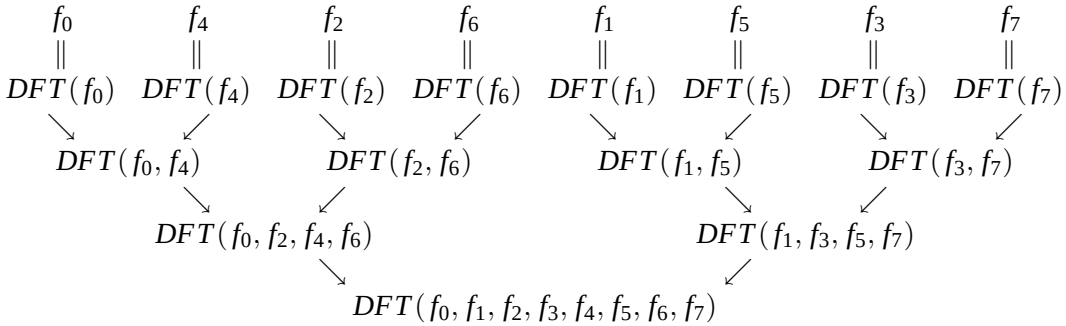
$$\begin{aligned} d_k &= \frac{1}{2N}\Big(\sum_{n=0}^{N-1} f_n e^{-2\pi i(2n)k/2N} + \sum_{n=0}^{N-1} f_{N+n} e^{-2\pi i(2n+1)k/2N}\Big) \\ &= \frac{1}{2N}\Big(\sum_{n=0}^{N-1} f_n e^{-2\pi ink/N} + e^{-\pi ik/N} \sum_{n=0}^{N-1} f_{N+n} e^{-2\pi ink/N}\Big) \\ &= \frac{1}{2}\big(DFT(f_0, f_1, \ldots, f_{N-1}) + e^{-\pi ik/N} DFT(f_N, f_{N+1}, \ldots, f_{2N-1})\big)_k \end{aligned}$$

(1. Gleichung: Def DFT, und Aufteilen in zwei Summen, in einer die geraden $(2n)$, in einer die ungeraden $(2n+1)$.) Für den $k+N$-ten Eintrag ($0 \le k \le N-1$) erhalten wir ganz analog

$$\begin{aligned} d_{k+N} &= \frac{1}{2N}\Big(\sum_{n=0}^{N-1} f_n e^{-2\pi i(2n)(k+N)/2N} + \sum_{n=0}^{N-1} f_{N+n} e^{-2\pi i(2n+1)(k+N)/2N}\Big) \\ &= \frac{1}{2N}\Big(\sum_{n=0}^{N-1} f_n e^{-2\pi in(k+N)/N} + e^{-\pi i(k+N)/N} \sum_{n=0}^{N-1} f_{N+n} e^{-2\pi i(k+N)n/N}\Big) \\ &= \frac{1}{2N}\Big(\sum_{n=0}^{N-1} f_n e^{-2\pi ink/N} e^{-2\pi in} + e^{-\pi i} e^{-i\pi k/N} \sum_{n=0}^{N-1} f_{N+n} e^{-2\pi ikn/N} e^{-2\pi in}\Big) \\ &= \frac{1}{2N}\Big(\sum_{n=0}^{N-1} f_n e^{-2\pi ink/N} - e^{-i\pi k/N} \sum_{n=0}^{N-1} f_{N+n} e^{-2\pi ikn/N}\Big) \\ &= \frac{1}{2}\big(DFT(f_0, f_1, \ldots, f_{N-1}) - e^{-\pi ik/N} DFT(f_N, f_{N+1}, \ldots, f_{2N-1})\big)_k \end{aligned}$$

□

Dieses Ergebnis liefert einen divide-and-conquer-Algorithums zum Berechnen der DFT in $O(n \log n)$ Schritten. (Hier nur für $N = 2^q$.) Dazu muss das Problem aufgeteilt werden in das Berechnen zweier DFTs, die dann wieder aufgeteilt werden in 2 mal 2 usw.; bis jeweils N Stück DFTs der Länge 1 berechnet werden müssen. Die müssen dann in der richtigen Reihenfolge kombiniert werden. Hier das Schema (für $N = 8$):



Das einzige Problem ist nun, wie bringen wir die f_n in die richtige Anfangsreihenfolge? Die richtige Reihenfolge erhalten wir einfach durch Bitumkehr:

$$000 \to 000,\ 001 \to 100,\ 010 \to 010,\ 011 \to 110,\ 100 \to 001,\ usw$$

Also (im Falle $N = 8$) muss an der 0-ten Stelle f_0 stehen, an der ersten Stelle f_4, an der zweiten f_2 usw. Natürlich hängt die Bitumkehr von $N = 2^q$ ab. Für $N = 16$ ergibt sich etwa

$$0000 \to 0000,\ 0001 \to 1000,\ 0010 \to 0100,\ 0011 \to 1100,\ 0100 \to 0010,\ usw$$

Algorithmus (FFT): Sei $N = 2^q$. Sei $g = (g_0, g_1, \ldots, g_{N-1})$ der Vektor, den man durch Umnummerierung mittels Bitumkehr aus $f = (f_0, f_1, \ldots, f_{N-1})$ erhält.

Starte mit den Vektoren der Länge 1: $d^{[0,j]} = (g_j)$ $(j = 0, \ldots, N-1)$. Berechne in Schritt r $(r \geq 1)$ die 2^{q-r} Vektoren der Länge 2^r

$$d^{[r,0]}, d^{[r,1]}, \ldots, d^{[r,2^{q-r}-1]}.$$

aus den Vektoren im $r-1$-ten Schritt gemäß

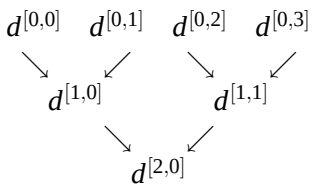
$$d_k^{[r,j]} = \frac{1}{2}\left(d_k^{[r-1,2j]} + e^{-\pi ik/2^{r-1}} d_k^{[r-1,2j+1]}\right)$$
$$d_{2^{r-1}+k}^{[r,j]} = \frac{1}{2}\left(d_k^{[r-1,2j]} - e^{-\pi ik/2^{r-1}} d_k^{[r-1,2j+1]}\right)$$

Dabei läuft (innerste Schleife) $k = 0, 1, \ldots, 2^{r-1} - 1$, und (zweitinnerste Schleife) $j = 0, 1, \ldots, 2^{q-r} - 1$, sowie $r = 1, 2, \ldots, q$.

Beispiel 6.9. Hier ein (sehr einfaches) Beispiel für $N = 4$: Wir berechnen die DFT für den Vektor (Datensatz) $f = (8, -4, -8, 16)^T$. (Das ^{T} bedeutet nur, dass das als Spaltenvektor zu lesen ist). Bitumkehr liefert

$$g_0 = f_0 = 8,\ g_1 = f_2 = -8,\ g_2 = f_1 = -4,\ g_3 = f_3 = 16$$

Das Schema ist



Dabei sind $d^{[1,0]}$ und $d^{[1,1]}$ Vektoren der Länge 2 (also $d^{[1,0]} = (d_0^{[1,0]}, d_1^{[1,0]})$ usw) und $d^{[2,0]}$ ist ein Vektor der Länge 4. Die konkreten Werte:

$$\begin{array}{cccccccc} 8 & & -8 & & -4 & & 16 \\ & \searrow & & \swarrow & & \searrow & & \swarrow \\ & & (0;8) & & & & (6;-10) & \\ & & & \searrow & & \swarrow & & \\ & & & & (3;4+5i;-3;4-5i) & & & \end{array}$$

Dabei berechnet sich z.B. $d^{[1,0]} = (d_0^{[1,0]}, d_1^{[1,0]})$ so:

$$\begin{aligned} d_0^{[1,0]} &= \frac{1}{2}(d_0^{[0,0]} + e^{-\pi i 0} d_0^{[0,1]}) = \frac{1}{2}(8-8) = 0 \\ d_1^{[1,0]} &= \frac{1}{2}(d_0^{[0,0]} - e^{-\pi i 0} d_0^{[0,1]}) = \frac{1}{2}(8-(-8)) = 8, \end{aligned}$$

und $d^{[2,0]} = (d_0^{[2,0]}, d_1^{[2,0]}, d_2^{[2,0]}, d_3^{[2,0]})$ so:

$$\begin{aligned} d_0^{[2,0]} &= \frac{1}{2}(d_0^{[1,0]} + e^{-\pi i 0} d_0^{[1,1]}) = \frac{1}{2}(0+6) = 3 \\ d_1^{[2,0]} &= \frac{1}{2}(d_1^{[1,0]} + e^{-\pi i/2} d_1^{[1,1]}) = \frac{1}{2}(8-i(-10)) = 4+5i, \\ d_2^{[2,0]} &= \frac{1}{2}(d_0^{[1,0]} - e^{-\pi i 0} d_0^{[1,1]}) = \frac{1}{2}(0-6) = -3 \\ d_3^{[2,0]} &= \frac{1}{2}(d_1^{[1,0]} - e^{-\pi i/2} d_1^{[1,1]}) = \frac{1}{2}(8+i(-10)) = 4-5i. \end{aligned}$$

Als Probe die “alte”, die Matrixmethode:

$$DFT\left((8,-4,-8,16)^T\right) = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{pmatrix}\begin{pmatrix} 8 \\ -4 \\ -8 \\ 16 \end{pmatrix} = \begin{pmatrix} 3 \\ 4+5i \\ -3 \\ 4-5i \end{pmatrix}$$

Aufwandsbetrachtungen: Die Matrixmethode braucht $O(N^2)$ Rechenoperationen: Mindestens schon $\frac{1}{2}(N-1)^2$ Multiplikationen, und N^2 Additionen. Andererseits gilt: Der FFT-Algorithmus (für $N = 2^q$) braucht $q = \log N$ Schritte (große Schritte, r läuft), und in jedem Schritt N Additionen und $N/2$ Multiplikationen. Also gilt:

Der FFT- Algorithmus berechnet die DFT von f in $O(N \log N)$ Schritten.

(Das Umordnen durch Bitumkehr dürfen wir vernachlässigen, das ist für große N billig: $O(N)$ oder weniger.) Die folgende Tabelle gibt einen Überblick über das Verhältnis des Aufwands beider Algorithmen (aus [AHKKLS]):

Aufwand			
N	normale DFT	FFT	prozentual
4	28	12	42,857%
16	496	96	19,355 %
64	8128	576	7,087%
256	130816	3072	2,348%
1024	2096128	15360	0,733%

Auch falls N keine Zweierpotenz ist, lassen sich gute Verfahren angeben. Im Wesentlichen geht es nur darum, $N = n\tilde{N}$ in n kleinere Datensätze der Länge $\tilde{N}$ aufzuteilen. Immer wenn N viele kleine Primfaktoren besitzt, geht das gut. Falls N große Primfaktoren besitzt, muss man auf andere Algorithmen zurückgreifen.

6.4 Schnelle Multiplikation

Eine Anwendung der FFT ist schnelle Multiplikation von großen Zahlen. Wir schildern hier ein Spielzeugbeispiel, dass das Prinzip in weiten Teilen verdeutlicht. Der wirkliche Algorithmus ist der *Algorithmus von Schönhage-Strassen*. Der arbeitet (so wie ich das verstehe) in endlichen Zahlringen: $C_N = \{0,1,2,\ldots,N-1\}$ mit Addition und Multiplikation mod N. Auch dort gibt es n-te Einheitswurzeln (Lösungen von $x^n - 1 = 0$). Das werden wir hier nicht vertiefen. Die Grundidee wird durch die hier vorgestellte Methode illustriert.

Schnelle Multiplikation von Polynomen

Wollen wir zwei Polynome vom Grad $N-1$ multiplizieren, etwa $p(x) = x^3 + x^2 + x + 1$ und $q(x) = x^3 + x^2 + 1$, dann brauchen wir naiv N^2 Multiplikationen:

$$\begin{aligned}(x^3+x^2+x+1)(x^3+x^2+1) &= x^6+x^5+x^4+x^3+x^5+x^4+x^3+x^2+x^3+x^2+x+1\\ &= x^6+2x^5+2x^4+3x^3+2x^2+x+1\end{aligned}$$

(hier ein paar weniger, da ein Koeffizient 0 ist). Stellen wir die Polynome als Koeffizientenvektoren p und q dar (also $p_0 + p_1x + p_2x^2 + \cdots \to (p_0, p_1, p_2, \ldots)$, so ist der Koeffizientenvektor des Produkts der Polynome gegeben durch N-mal die Faltung der Vektoren: $Np * q$ (vgl die Def der Faltung auf Seite 30). Wegen evtler Überträge müssen wir den Koeffizientenvektor mit $2N$ Koordinaten darstellen. Für unser Beispiel also

$$8 \cdot (1,1,1,1,0,0,0,0) * (1,0,1,1,0,0,0,0) = (1,1,2,3,2,2,1,0).$$

Jetzt haben wir folgende tolle Idee:

$$\boxed{p * q = IDFT(\widehat{p} \cdot \widehat{q})}$$

In Worten: Die Faltung (das Produkt der Polynome) berechnet sich nach dem Faltungssatz als inverse DFT des Produkts der DFTs von p und q. Die DFT und IDFT können wir effizient berechnen. Was bedeutet das Produkt? Ein Polynom ist gegeben durch seinen Koeffizientenvektor. Jedes Polynom vom Grad $N-1$ ist aber auch eindeutig gegeben durch N Funktionswerte. Die DFT von p ist gerade ($1/N$ mal) der Vektor der Funktionswerte an den (komplexen) Stellen $1, \xi, \xi^2, \ldots \xi^{N-1}$ (s. Aufgabe 41). Also:

$$\widehat{p} = \frac{1}{N}(p(1), p(\xi), p(\xi^2), \ldots, p(\xi^{N-1}))^T$$

Den Koeffizientenvektor von p bekommen wir dann als (N mal) die IDFT von $\widehat{p}$ zurück.

Die Funktionswerte des Produkts pq an den Stellen $1, \xi, \ldots, \xi^{N-1}$ ist der Vektor der Funktionswerte

$$\big(pq(1), pq(\xi), \ldots, pq(\xi^{N-1})\big)^T = \big(p(1)q(1), p(\xi)q(\xi), \ldots, p(\xi^{N-1})q(\xi^{N-1})\big)^T.$$

Den können wir aus $\widehat{p}$ und $\widehat{q}$ einfach berechnen: elementweise multiplizieren. Dessen IDFT liefert dann also die Koeffizientendarstellung von pq **(!!!)**. Nun ist pq ein Polynom von höherem Grad $(2N-2)$. Daher brauchen wir mindestens $2N-1$ Funktionswerte. Es bietet sich an, mit Koeffizientenvektoren der Länge $2N$ zu arbeiten. Also:

Algorithmus (schnelle Multiplikation nach Hausfrauenart)

Gegeben zwei Polynome p, q vom Grad $N-1$. Seien p und q ihre Koeffizientenvektoren der Länge $2N$. Berechne $\widehat{p}$ und $\widehat{q}$ mittels FFT. Berechne dann die IDFT des Vektors

$$2N(\widehat{p}_0 \cdot \widehat{q}_0, \widehat{p}_1 \cdot \widehat{q}_1, \ldots, \widehat{p}_{2N-1} \cdot \widehat{q}_{2N-1})$$

Ergebnis ist der Koeffzientenvektor von pq.

Beispiel 6.10. (Etwas anders als das oben) Sei $p(x) = x^3 + x^2 + x + 1$, $q(x) = x^2 + 1$. Also $N = 4$, also arbeiten wir mit Vektoren der Länge 8:

$$p = (1,1,1,1,0,0,0,0), \quad q = (1,0,1,0,0,0,0,0)$$

FFT liefert

$$\widehat{p} = \frac{1}{8}\big(4, 1-(1+\sqrt{2})i, 0, 1+(1+\sqrt{2})i, 0, 1-(1-\sqrt{2})i, 0, 1+(1+\sqrt{2})i\big)^T$$

$$\widehat{q} = \frac{1}{8}\big(2, 1-i, 0, 1+i, 2, 1-i, 0, 1+i\big)^T$$

Damit ist

$$8\widehat{p}\widehat{q} = (1, \frac{1}{8}(1-i)(1-(1+\sqrt{2})), 0, \frac{1}{8}(1+i)(1+(1+\sqrt{2})), 0, \frac{1}{8}(1-i)(1-(1-\sqrt{2})), 0, \frac{1}{8}(1+i)(1+(1+\sqrt{2})))^T$$

und die IDFT dieses Vektors ist $(1,1,2,2,1,1,0,0)^T$. Also ist

$$(x^3 + x^2 + x + 1)(x^2 + 1) = x^5 + x^4 + 2x^3 + 2x^2 + x + 1.$$

Schnelle Multiplikation großer Zahlen

Das ist nun einfach: Statt des Koeffizientenvektors eines Polynoms sei der Vektor nun die Binärdarstellung der Länge $2N$ zweier Zahlen mit der Länge N. (Man überlege sich: wenn $(p_0, p_1, p_2, \ldots)$ die Binärdarstellung ist, und p das Polynom zu diesem Koeffizientenvektor, was ist dann $p(2)$?) Ein Beispiel:

$$15 = (1,1,1,1,0,0,0,0), \quad 5 = (1,0,1,0,0,0,0,0)$$

Ganz analog wie oben wenden wir den Algorithmus an. Dann ist

$$15 \cdot 5 = (1,1,2,2,1,1,0,0)$$

Das ist so keine Binärzahl, aber Abarbeiten der Überträge (von links nach rechts) liefert die korrekte Binärzahl:

$$(1,1,2,2,1,1,0,0) \to (1,1,0,3,1,1,0,0) \to (1,1,0,1,2,1,0,0) \to (1,1,0,1,0,2,0,0)$$
$$\to (1,1,0,1,0,0,1,0) = 75 = 15 \cdot 5$$

Bemerkung 6.11. Das obige Beispiel macht auch klar, warum es besser ist mit Einheitswurzeln in $(C_N, +, \cdot)$ zu arbeiten statt mit komplexen Einheitswurzeln: Die ersteren sind ganzzahlig und reell (integer), die letzteren weder noch (Brüche, Wurzeln, imaginäre Anteile). Das erste ist numerisch einfacher und stabiler.

Ein Vergleich mit Tabelle auf Seite 40 ergibt: Für Zahlen mit 1024 Bit lohnt sich's schon: Statt (i.Wes.) $1024^2 = 1.048.576$ Operationen für "naive" Multiplikation brauchen wir (i.Wes.) nur drei DFTs der Länge 2048, also weniger als 180.000 Operationen. (Grob grechnet: Für DFT der Länge 1024: 15.360, für eine der Länge 2048 dann wohl weniger als 60.000.)

7 FT und DGL

Vorgänge in der Natur lassen sich oft durch Differentialgleichungen (DGL) beschreiben (z.B. die schwingende Saite in Kapitel 1, aber auch Diffusion von Flüssigkeiten, zeitliche Entwicklung von Tierpopulationen ("Räuber-Beute-Modell") oder die Flecken des Leoparden ("Aktivator-Inhibitor-Modell"). DGL sind Gleichungen, in denen Variablen, Funktionen und ihre Ableitungen vorkommen. Z.B.

$$f'' = 2f' - f \qquad \text{oder} \qquad f''' = x^2 f$$

Bei der zweiten dieser Gleichungen muss man spezifizieren, ob x irgendeine Variable ist, oder die, nach der abgeleitet wird. Daher schreibt man i.Allg. nicht f', sondern $\frac{d}{dx}f$, um klarzumachen, dass nach x abgeleitet wird. (Wenn aus dem Kontext klar ist, was gemeint ist, schreiben wir aber weiterhin oft f' statt $\frac{d}{dx}f$.) Dass ist insbesondere wichtig für Funktionen in mehreren Variablen, z.B. die für die schwingende Saite: da darf man t einsetzen (Zeit) und x (Ort). Daher kann man nicht mehr von *der* Ableitung sprechen. Die entsprechenden Ableitungen nach einer dieser Variablen heißen *partielle Ableitungen*. Bezeichnet werden die folgendermaßen. Ist f von der Form $f(t,x)$, so sind

$$\frac{\partial}{\partial t}f \qquad \text{bzw} \qquad \frac{\partial}{\partial x}f$$

die Ableitungen von f nach t bzw nach x. Gleichbedeutend damit ist die Schreibweise $\frac{\partial f}{\partial t}$. Mehrfache Ableitungen schreiben wir als

$$\frac{\partial^3 f}{\partial t^3} \qquad \text{bzw} \qquad \frac{\partial^2 f}{\partial t \partial x} \qquad \text{bzw} \qquad \frac{\partial^7 f}{\partial t^3 \partial x^2 \partial t^2}$$

für: f 3mal nach t abgeleitet, bzw f einmal nach x, dann einmal nach t abgeleitet, bzw f 2mal nach t, dann 2mal nach x, dann 3mal nach t abgeleitet. Wichtig ist: Existieren die zweiten Ableitungen nach jeder einzelnen Koordinate, dann ist die Reihenfolge egal [H1]:

Satz 7.1. *Sei $f : \mathbb{R}^d \to \mathbb{C}$, also f von der Form $f(x_1, x_2, \ldots, x_d)$. Existieren alle zweiten partiellen Ableitungen $\frac{\partial^2 f}{\partial x_m \partial x_n}$ $(n, m = 1, 2, \ldots, d)$, und sind stetig, dann gilt*

$$\frac{\partial^2 f}{\partial x_n \partial x_m} = \frac{\partial^2 f}{\partial x_m \partial x_n}$$

Analoges gilt für höhere Ableitungen. Unsere Funktionen werden diese Bedingung immer erfüllen.

Beispiel 7.2. Betrachte die DGL ganz am Anfang in Kapitel 1 (für $c = 1$):

$$\frac{\partial^2}{\partial t^2}f = \frac{\partial^2}{\partial x^2}f, \tag{7.1}$$

wobei f von der Form $f(t,x)$ ist. Eine Lösung war $f(t,x) = \sin(t)\big(\sin(x) + \cos(x)\big)$. Warum? Hatten wir damals hergeleitet. Aber machen wir die Probe: Dieses f zweimal nach t abgeleitet ist

$$\frac{\partial^2}{\partial t^2}f = -\sin(t)(\sin(x) + \cos(x)),$$

und zweimal nach x abgeleitet ist

$$\frac{\partial^2}{\partial x^2}f = \sin(t)(-\sin(x) + (-\cos(x))) = -\sin(t)(\sin(x) + \cos(x)).$$

Die DGL ist also erfüllt.

DGL, in denen partielle Ableitungen vorkommen, heißen auch PDE (partial differential equation). Oft ist nicht nur die DGL vorgegeben, sondern auch Anfangswerte (wie ja auch bei der schwingenden Saite). Dann spricht man von einem Anfangswertproblem (AWP) (synonym: Randwertproblem, Anfangswertaufgabe). Im Beispiel oben würden wir etwa zusätzlich fordern, dass gilt:

$$f(0,x) = 0 \quad \text{und} \quad f(\pi,x) = 0 \quad (x \in \mathbb{R}) \tag{7.2}$$

Dann sind (7.1) und (7.2) zusammen das AWP.

DGL und PDE sind im Allgemeinen schwierig zu lösen. Es gibt z.B. eine, bzw ein System von PDE, die Navier-Stokes-Gleichungen [WIK], die das Strömungsverhalten von Flüssigkeiten beschreiben. Wenn man beweisen kann, dass dieses PDEsystem eine Lösung auf ganz $\mathbb{R}^3$ hat, gewinnt man eine Million Dollar. Man braucht die Lösung nicht unbedingt hinzuschreiben, ein Existenzbeweis genügt.

Zu DGL gibt es also sehr viel zu erzählen. Hier beschränken wir uns nur auf Lösungsmethoden, die auf FT beruhen. Grundlage ist folgendes Resultat, welches besagt, dass Ableitungen der Funktion f nach x (im "real space") einfach Multplikation der FT der Fktn (im "Fourier-space") mit einer Potenz von ix entsprechen. (Bzw mit ik, wenn man die Variablen im real space und im Fourier-space unterscheiden will.)

Satz 7.3. *Es sei $f : \mathbb{R} \to \mathbb{C},\ f \in L^1$ und diff-bar, und f' sei auch in L^1. Dann gilt*

$$\widehat{(f')}(k) = ik\widehat{f}(k)$$

Das mehrdimensionale Analogon des letzten Satzes ist:

Satz 7.4. *Es sei $f : \mathbb{R}^d \to \mathbb{C},\ f \in L^1$, also f von der Form $f(x_1,x_2,\ldots,x_d) = \cdots$. Dann ist $\widehat{f}$ von der Form $\widehat{f}(k_1,k_2,\ldots,k_d)$. Weiterhin soll die partielle Ableitung $\frac{\partial}{\partial x_m} f$ existieren und auch in L^1 liegen. Dann gilt*

$$\widehat{\frac{\partial f}{\partial x_m}} = ik_m\widehat{f}$$

Der Beweis ist in beiden Fällen der Gleiche. Wir formulieren ihn in der ersten Form (vgl auch Aufgabe 28, Blatt 8).

Proof. Weil f' integrierbar ist, gilt $f(x) - f(0) = \int_0^x f'(t)dt$, und der Limes $\lim_{x\to\infty} f(x)$ existiert, nämlich $\int_0^\infty f'(t)dt + f(0)$. Dann aber muss dieser Limes 0 sein (sonst wäre f nicht integrierbar: es käme beim Integrieren was unendliches raus.) Dito für $x \to -\infty$. Dann hilft partielle Integration:

$$\widehat{(f')}(k) = \int_{\mathbb{R}} f'(x)e^{-ikx}dx = f(x)e^{-ikx}\Big|_{x=-\infty}^{\infty} - (-ik)\int_{\mathbb{R}} f(x)e^{-ikx}dx = 0 - 0 + ik\widehat{f}(k) = ik\widehat{f}(k)$$

□

Korollar 7.5. *Unter den Voraussetzungen von Satz 7.4, und falls die höheren partiellen Ableitungen existieren und in L^1 liegen, gilt*

$$\widehat{\frac{\partial^n f}{\partial x_m^n}} = (ik_m)^n\widehat{f}$$

Das folgt direkt durch mehrmalige Anwendung von Satz 7.4.

Für die folgenden Anwendungen stellen wir uns f oft als Funktion in t und x vor. Dabei ist $t \in \mathbb{R}$ immer eindimensional (Zeit), x darf auch mehrdimensional sein (Ort), also $x \in \mathbb{R}^d$.

Oft werden wir keine *Lösung in geschlossener Form* bekommen, also keine, die sich direkt als Funktion hinschreiben lässt. Oft sind wir zufrieden, wenn wir f irgendwie in Abhängigkeit von g hinschreiben können (z.B. f als Integral über irgendwas mal g, oder irgendwas mal $\widehat{g}$). Dann nämlich können wir erstens hoffen, etwas über das qualitative Verhalten auszusagen (z.B. das Langzeitverhalten), zweitens können wir die Formeln numerisch, also näherungsweise, auswerten.

7.1 Die Wärmegleichung

Die folgende PDE bzw das AWP heißt (eindimensionale) Wärmegleichung. Es beschreibt die zeitliche Verteilung der Wärme in einem Medium, bei gegebener Anfangsverteilung g. Hier ist das Medium eindimensional (also etwa ein langer Draht oder so.)

$$\frac{\partial f}{\partial t} - \frac{\partial^2 f}{\partial x^2} = 0, \quad f(0,x) = g(x) \quad (x \in \mathbb{R})$$

Wir wenden nun FT auf die Gleichungen an, und zwar (TRICK) nur auf die räumliche Variable, also x. Wir erhalten

$$\frac{\partial}{\partial t}\widehat{f} - (ik)^2\widehat{f} = \frac{\partial}{\partial t}\widehat{f} + k^2\widehat{f} = 0, \quad \widehat{f} = \widehat{g} \quad \text{für } t = 0$$

Das fassen wir nun als Gleichung nur noch in t auf, die andere Variable, k, betrachten wir als konstant. Damit ist es eine einfache DGL geworden: Löse

$$\left(\hat{f}\right)' + k^2\widehat{f} = 0, \quad \widehat{f}(0) = \widehat{g}$$

Diese ist nun einfacher, insbesondere kommt nur noch eine erste Ableitung vor. Die Lösung kann man fast schon raten. (Ansonsten DGL-Buch, recht weit vorne; oder Prop 8.1, S 50.) Die einzige Lösung ist $\widehat{f}(t) = \widehat{g}e^{-k^2t}$. Also ist f die inverse FT davon, $f = IFT(\widehat{g} \cdot e^{-k^2t})$. Nach dem Faltungssatz ist dann $f = g * h$, mit $h = IDFT(e^{-k^2t})$. Dann ist (vgl Bsp 5.3, S. 23)

$$h(x) = \frac{1}{\sqrt{4\pi t}} e^{-x^2/4t}$$

Also ist

$$f(x,t) = g * h(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{\mathbb{R}} e^{\frac{-(x-y)^2}{4t}} g(y)dy$$

Damit sind wir zufrieden: Das ist immer noch ein komplizierter Term, aber er gilt für alle g (zumindest alle erlaubten, z.B. $g \in L^1$), liefert also eine allgemeine Lösung. Außerdem lässt sich diese Darstellung — bei gegebenem g — numerisch auswerten.

Auch lässt sich etwas aussagen über das Langzeitverhalten, also $t \to \infty$: Für große t wird der Vorfaktor immer kleiner. Der e-Term im Integral ist i.Wes. eine Glockenkurve. Diese wird mit wachsendem t immer flacher und breiter. Diese wird mit g gefaltet. Egal wie g aussieht (z.B. starke Erhitzung eines kleinen Teils des Drahts), die Temperatur gleicht sich im Laufe der Zeit überall an und geht gegen 0 (wobei 0 der "Grundtemperatur" des Drahtes entspricht, also z.B. Umgebungstemperatur).

7.2 Die mehrdimensionale Wellengleichung

Zunächst machen wir uns genauer klar, was die mehrdim FT bedeutet: Ist $f : \mathbb{R}^d \to \mathbb{C}$, $f \in L^2$, dann ist

$$\widehat{f} : \mathbb{R}^d \to \mathbb{C}, \quad \widehat{f}(k) = \int_{\mathbb{R}^d} e^{-ik\cdot x} f(x) dx$$

Dabei ist $k \cdot x$ das Skalarprodukt: $k \cdot x = (k_1, \ldots, k_d) \cdot (x_1, \ldots, x_d) = k_1 x_1 + \cdots + k_d x_d$. Also steht in der Def oben $\cdots e^{-i(k_1 x_1 + \cdots + k_d x_d)} \cdots$, und somit ein Mehrfachintegral

$$\widehat{f}(k) = \int_{\mathbb{R}} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} e^{-i(k_1 x_1 + \cdots + k_d x_d)} f(x_1, \ldots, x_d) dx_1 \cdots dx_d.$$

Die Verallgemeinerung von Korollar 7.5 auf gemischte partielle Ableitungen sieht dann so aus:

Korollar 7.6. *Unter den Voraussetzungen von Korollar 7.5 gilt:*

$$\widehat{\frac{\partial^n f}{\partial x_1^{n_1} \partial x_2^{n_2} \cdots \partial x_d^{n_d}}} = (ik_1)^{n_1} (ik_2)^{n_2} \cdots (ik_d)^{n_d} \widehat{f},$$

wobei $n = n_1 + \cdots + n_d$.

Das wollen wir gerne etwas kürzer schreiben. Daher setzen wir für $\alpha = (n_1, \ldots, n_d) \in \mathbb{N}_0^d$

$$D^\alpha f = \frac{\partial^n f}{\partial x_1^{n_1} \partial x_2^{n_2} \cdots \partial x_d^{n_d}}$$

Weiterhin sei für $k \in \mathbb{R}^d$ definiert: $k^\alpha = k_1^{n_1} k_2^{n_2} \cdots k_d^{n_d}$. Damit wird aus der Gleichung oben

$$\widehat{D^\alpha f} = (ik)^\alpha \widehat{f}$$

Eine weitere wichtige Notation ist der *Laplaceoperator*:

$$\Delta f = \frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2} + \cdots + \frac{\partial^2 f}{\partial x_d^2}$$

In Worten: Δf ist die Summe aller zweiten partiellen Ableitungen nach *einer* Variablen.

Beispiel 7.7. Sei $f(x) = x_1^3 + x_1^2 - x_1 - x_2^4 + 8x_2^2$. Dann ist

$$\begin{aligned} \frac{\partial^2 f}{\partial x_1^2} &= 6x_1 + 2 \\ \frac{\partial^2 f}{\partial x_2^2} &= -12x_2^2 + 16, \quad \text{also} \\ \Delta f &= 6x_1 - 12x_2^2 + 18 \end{aligned}$$

Recall: Der *Gradient* ∇f einer differenzierbaren Funktion $f : \mathbb{R}^d \to \mathbb{R}$ ist der Vektor der ersten partiellen Ableitungen. Der Gradient der Funktion f aus dem obigen Beispiel ist also $\nabla f(x_1, x_2) = (3x_1^2 + 2x_1 - 1;\ -4x_2^3 + 16x_2)$. Der Gradient an der Stelle $x = (x_1, \ldots, x_d)$ ist ein Vektor, der in die Richtung des steilsten Anstiegs an der Stelle x zeigt. Je steiler der Anstieg, desto länger ist dieser Vektor. Das illustrieren die folgenden Bilder.

Der Laplaceoperator ist die Summe der partiellen Ableitungen der Einträge des Gradienten, also ein Maß für die Änderungsrate der Steigung an der entsprechenden Stelle.

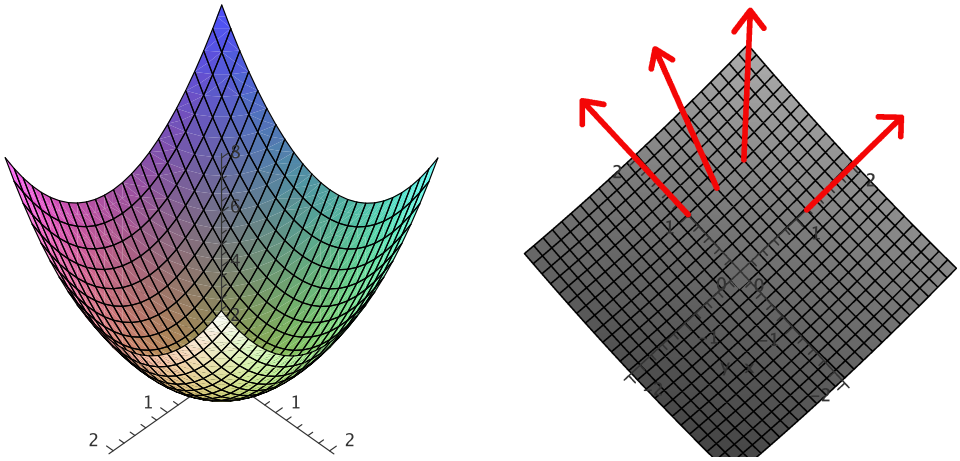


Figure 10: Die Funktion $f(x,y) = x^2 + y^2$ in $[-2;2] \times [-2;2]$ (links), und die Gradienten bei $(1,0), (1,\frac{1}{2}), (1,1)$ und $(0,1)$. (Das rechte Bild ist der Graph "von oben" gesehen.)

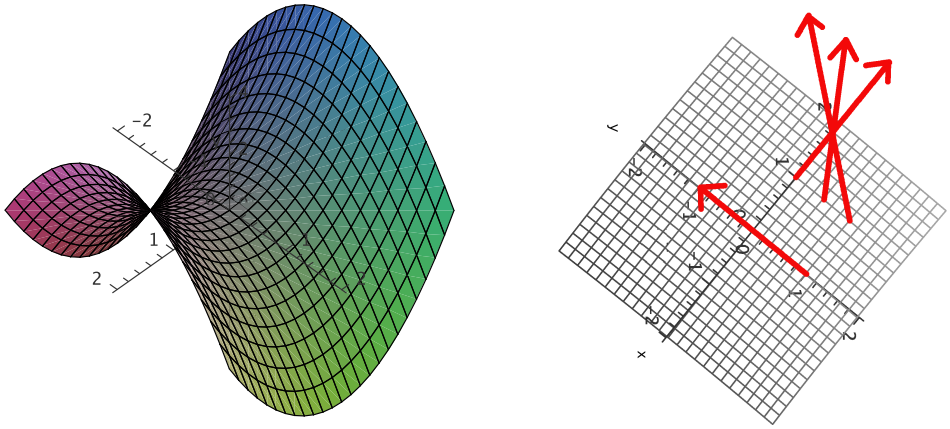


Figure 11: Die Funktion $f(x,y) = x^2 - y^2$ in $[-2;2] \times [-2;2]$ (links), und die Gradienten bei $(1,0), (1,\frac{1}{2}), (1,1)$ und $(0,1)$.

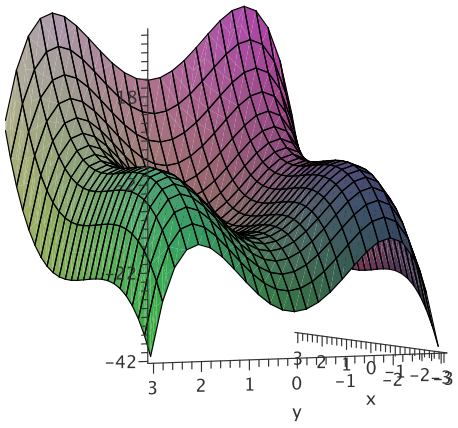


Figure 12: Die Funktion $f(x,y) = x^3 + x^2 - x - y^4 + 8y^2$ in $[-3;3] \times [-3;3]$. Der Gradient wird 0 an den lokalen Maxima bei $(-2,-1)$ und $(2,-1)$ (auf den "Armlehnen" des Sessels), sowie beim lokalen Minimum bei $(1,0)$ (in der "Sitzfläche" des Sessels).

Beispiel 7.8. Für $f : \mathbb{R} \to \mathbb{R},\ f(x) = x^2$ ist die Steigung an der Stelle x genau $f'(x) = x$. Die Steigung nimmt also linear zu, (bei $x = 1$: 1, bei $x = 2$: 2, bei $x = 3$: 3, usw). die Änderungsrate der Steigung ist also konstant. Passt: Hier ist Δf einfach die zweite Ableitung, also $\Delta f(x) = 1$.

Physikalisch drückt der Gradient die Richtung aus, in der ein Kraftgefälle herrscht, und wie hoch das ist. Drückt f etwa den Luftdruck aus, so zeigt der Gradient an der Stelle x in Richtung des höchsten Druckanstiegs, und sein negatives daher in die Richtung des höchsten Druckabfalls. In genau diese Richtung werden dann die Luftteilchen bei x gezogen. Dadurch ändert sich nun die Druckverteilung, also f. Ein Maß für das Ungleichgewicht der Druckverteilung bei x (also für die Änderungsrate des Luftdrucks) ist also Δf. Insbesondere ist, wenn kein Ungleichgewicht vorliegt, $\Delta f = 0$. Das System ist dann in einem Gleichgewichtszustand. Ergo:

Bemerkung 7.9. Gleichgewichte in physikalischen Modellen lassen sich durch Laplaceoperatoren beschreiben. Ein System, beschrieben durch f, ist im Gleichgewicht, falls $\Delta f = 0$.

Nun erinnern wir an die physikalische Tatsache, dass, wenn h den Ort zur Zeit t beschreibt, die erste Ableitung von h nach t die Geschwindigkeit beschreibt, und die zweite Ableitung die Beschleunigung.

Beispiel 7.10. Ein raketengetriebenes Auto, das zur Zeit $t = 0$ am Ort $h(0) = 0$ ist, erfährt durch die Rakete eine konstante Beschleunigung, also $\frac{d^2}{dt^2}h(t) = 1$ für alle $t \geq 0$. (Die Konstante können wir durch geeignete Wahl der Maßeinheit immer zu eins machen.) Daher ist die Geschwindigkeit zur Zeit t genau $\frac{d}{dt}h(t) = t$ (evtl plus einer Konstanten). Und der Ort, an das die Raketenauto zur Zeit t ist, ist $h(t) = \frac{1}{2}t^2$ (evtl plus einer Konstanten, aber wir wissen ja: $h(0) = 0$, also ist die Konstante Null).

Das bringt uns zur Wellengleichung. Die Idee ist, dass die Beschleunigung, die auf ein Teilchen zur Zeit t an der Stelle x wirkt (also $\frac{\partial^2 f}{\partial t^2}(t,x)$) gleich dem Ungleichgewicht bei x ist, also $\Delta f(t,x)$. Gehen wir von einer Anfangsverteilung g aus, und einer Anfangsgeschwindigkeit 0, so erhalten wir das AWP

$$\frac{\partial^2}{\partial t^2} f - \Delta f = 0, \qquad f(0,x) = g(x), \qquad \frac{\partial f}{\partial t}(0,x) = 0.$$

Das ist die *Wellengleichung* (wave equation). Um sie zu lösen, benutzen wir FT in x. Wir wissen (vgl Aufgabe 48, Blatt 13):

$$\widehat{\Delta f} = -|k|^2 \widehat{f}$$

Also wird aus der Gleichung oben durch räumliche FT (FT nur in x)

$$\frac{\partial^2}{\partial t^2} f + |k|^2 f = 0, \qquad \widehat{f}(0,k) = \widehat{g}(k), \qquad \frac{\partial \widehat{f}}{\partial t}(0,k) = 0.$$

Wieder ist das eine einfache DGL (bzw ein AWP) nur in t. Deren allgemeine Lösung ist (vgl Aufgabe 45, Blatt 12, und beachte: $-(i|k|)^2 = |k|^2$)

$$c \cdot e^{i|k|t} + d \cdot e^{-i|k|t}$$

Wegen der Anfangsbedingungen muss gelten

$$c \cdot e^{i|k|0} + d \cdot e^{-i|k|0} = \widehat{g}, \qquad ci|k| \cdot e^{i|k|0} - di|k| \cdot e^{-i|k|0} = 0,$$

also $c + d = \widehat{g}$ und $c - d = 0$, also $c = d = \frac{\widehat{g}}{2}$. Damit ist

$$\widehat{f} = \frac{\widehat{g}}{2}(e^{it|k|} + e^{-it|k|}),$$

und mit der Umkehrformel (vgl Satz 5.4, 23) erhalten wir:

$$f(t,x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\widehat{g}(k)}{2}(e^{i(x\cdot k + t|k|)} + e^{i(x\cdot k - t|k|)})dk.$$

Damit sind wir wieder zufrieden: Wie bei der Wärmegleichung ist das eine allgemeine Lösung in g (gut), nicht in geschlossener Form (schlecht), aber in jedem Fall näherungsweise berechenbar (gut). Anders als bei der Wärmegleichung zeigt die Wellengleichung kein eindeutiges Langzeitverhalten. Das passt zur physikalischen Interpretation: Es könnte ja — da in dem Modell Reibungsverlust vernachlässigt wird — eine Welle sich ewig weiter ausbreiten, oder es könnte ein Wellenmuster sich (zeit-)periodisch immer wiederholen.

Bei der Wellengleichung war die Grundidee:

Beschleunigung bei x = Ungleichgewicht des Systems bei x

Mit unserem jetzigen Kenntnisstand lesen wir die Wärmegleichung (s. oben) ähnlich. Ihre mehrdim Form ist $\frac{\partial}{\partial t} f - \Delta f = 0$. Also:

Geschwindigkeit bei x = Ungleichgewicht des Systems bei x

Mit Fouriermethoden lassen sich einige andere wichtige PDE behandeln. In dem Buch [EV] gibt es ein Kapitel, das heißt: “Four important linear PDE”. Zwei davon sind die (mehrdimensionale) Wärmegleichung und die Wellengleichung. Eine weitere, in der Quantenphysik fundamental wichtige Gleichung ist die *Schrödingergleichung* [EV], [WIK]; auch diese lässt sich mit den oben beschriebenen Methoden behandeln.

8 Elementares zu Funktionalgleichungen

Wie bei DGL sind Funktionalgleichungen (FGL) solche, bei denen die gesuchte Lösung eine Funktion ist. Im Gegensatz zu DGL kommen aber keine Ableitungen vor. Oft ist es so, dass die Funktionswerte an verschiedenen Stellen in Beziehung gesetzt werden. Z.B.

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x)f(y) = f(x+y), \quad f(0) = 1 \tag{8.1}$$

für alle $x,y \in \mathbb{R}$. (Diese FGL heißt *Cauchysche FGL*.) Die Aufgabe ist nun, alle Funktionen zu bestimmen, für die (8.1) gilt.

Erster Versuch: Sicher erfüllen die konstante Funktionen $f(x) = 0$ und $f(x) = 1$ den ersten Teil der FGL (8.1). Wegen dem zweiten Teil bleibt von denen nur die Funktion $f(x) = 1$ übrig.

Zweiter, raffinierterer Versuch: Im Hinblick auf die Potenzgesetze: $a^x a^y = a^{x+y}$ ($a > 0, x,y \in \mathbb{R}$) und $a^0 = 1$ sind Funktionen der Form a^x Lösungen.

Recall: Was heißt a^x eigentlich? Für $x \in \mathbb{N}$ ist das klar: a wird n-mal mit sich selbst malgenommen. Aber was ist $a^{\sqrt{2}}$? Solche Potenzen sind daher über die Exponentialfunktion definiert, es ist

$$a^x := e^{x \ln a}$$

Dabei ist

$$e^x = \exp(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!},$$

und $\ln(x)$ ist die Umkehrfunktion davon (also die eindeutige mit $\exp(\ln(x)) = \ln(\exp(x)) = x$).

Damit ist jetzt klar: $f : \mathbb{R} \to \mathbb{R},\ f(x) = e^{x \ln a}$ ist Lösung der FGL (8.1), für alle $a > 0$. Checken:

$$f(x)f(y) = e^{x \ln a} e^{y \ln a} = e^{(x+y)\ln a} = f(x+y), \quad \text{stimmt,}$$

und $f(0) = e^0 = 1$, stimmt auch. Statt $\ln a$ schreiben wir im Folgenden b. Für a sind nur positive Werte erlaubt, aber $\ln a$ durchläuft ganz $\mathbb{R}$, wenn a ganz $\mathbb{R}^+$ durchläuft. Also ist $b \in \mathbb{R}$ beliebig. Wir fragen nun: Sind das alle Lösungen? Das ist sehr kompliziert, wenn keine weiteren Einschränkungen vorgenommen werden. Z.B. gibt es Lösungen der FGL (8.1), die nirgends stetig sind (vgl Engel-Nagel: "One-Parameter Semigroups for Linear Evolution Equations", oder Hamel 1905). Also fragen wir lieber:

Sind alle stetigen Lösungen der FGL (8.1) von der Form $f(x) = e^{bx}$?

Das werden wir im Folgenden beantworten.

Proposition 8.1. *Sei $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^{bx}$, $b \in \mathbb{R}$. Dann ist f diff-bar und erfüllt die DGL (genauer, das AWP)*

$$\frac{d}{dx} f(x) = b f(x), \quad f(0) = 1. \tag{8.2}$$

Es ist dann $a = \frac{df}{dx}(0)$. Außerdem ist dieses f die einzige diff-bare Lösung des AWP.

Proof. Wir brauchen nur die Eindeutigkeit zu zeigen. Sei g eine weitere diff-bare Lösung von (8.2). Dann ist, für ein festes $x > 0$, die Funktion

$$q : [0,x] \to \mathbb{R}, \quad q(y) := f(y)g(x-y)$$

auch diff-bar, und es ist

$$\begin{aligned}\frac{d}{dy}q(y) &= \Big(\frac{d}{dy}f(y)\Big)g(x-y) + f(y)\frac{d}{dy}\big(g(x-y)\big) \\ &= \frac{df}{dy}(y)g(x-y) - f(y)\frac{dg}{dy}(x-y) = bf(y)g(x-y) - f(y)bg(x-y) = 0.\end{aligned}$$

Erstes "=": Produktregel, zweites "=": Kettenregel, drittes "=": (8.2). Zur Kettenregel: Sei $h(y) = x-y$, dann ist $g(x-y) = g(h(y))$. Das nach y ableiten: $g'(x-y)\cdot(-1)$. Zur Schreibweise: $\frac{d}{dy}\big(g(x-y)\big)$ heißt: Die Ableitung von $g(x-y)$ nach y, während $\frac{dg}{dy}(x-y)$ heißt: g nach y ableiten und dann $(x-y)$ einsetzen.

Die Ableitung von q nach y ist 0, also ist q konstant in y. Daher ist

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

daher ist $f(x) = g(x)$ für alle x, also ist $f = g$. Es kann also außer f keine weitere Lösung von (8.2) geben. □

Nun zeigen wir, dass jede stetige Lösung von (8.1) auch diff-bar ist.

Proposition 8.2. *Sei f eine stetige Lösung von* (8.1). *Dann ist f diff-bar, und es gibt genau ein $b \in R$, so dass* (8.2) *gilt.*

Proof. Weil f stetig ist, existiert

$$F(y) = \int_0^y f(x)dx \qquad (y \geq 0)$$

und ist diff-bar, mit $F' = f$. Daher gilt

$$F'(0) = f(0) = 1.$$

Die Steigung von F bei 0 ist also positiv. $F(0)$ ist 0, daher ist $F(y) \neq 0$ für $y \neq 0$, y nah genug an 0. Dann existiert $\frac{1}{F}(y)$ und ist diffbar (beides für $y \neq 0$, y nah bei 0). Wegen (8.1) gilt dann:

$$\begin{aligned}f(y) &= F^{-1}(y_0)F(y_0)f(y) = F(y_0)^{-1}\int_0^{y_0} f(y)f(x)dy = F(y_0)^{-1}\int_0^{y_0} f(x+y)dy \\ &= F(y_0)^{-1} + \int_x^{x+y_0} f(y)dy = F(y_0)^{-1}\big(F(x+y_0) - F(x)\big).\end{aligned}$$

Der rechte Ausdruck ist diff-bar, also auch f. Die Ableitung ist

$$\begin{aligned}\frac{d}{dy}f(y) &= \lim_{t\searrow 0}\frac{f(y+t)-f(y)}{t} = \lim_{t\searrow 0}\frac{f(t)-f(0)}{t}f(y) \\ &= f'(0)f(y)\end{aligned}$$

Also erfüllt f das AWP (8.2) mit $b := f'(0)$. □

Zusammengenommen besagen die letzten beiden Resultate, dass jede stetige Lösung der FGL (8.1) diff-bar ist und Lösung des AWP (8.2), und die eindeutige Lösung dieses AWP ist $f(x) = e^{bx}$.

Satz 8.3. *Sei f stetige Lösung der FGL* (8.1). *Dann ist $f(x) = e^{bx}$, $b \in \mathbb{R}$.*

8.1 Die Matrixexponentialfunktion

Bisher betrachteten wir Funktionen von $\mathbb{R}^d \to \mathbb{R}$ oder $\mathbb{R}^d \to \mathbb{C}$. Also Def-bereich mehrdimensional, Bildbereich eindimensional. Nun betrachten wir Funktionen von $\mathbb{R}$ nach $\mathbb{R}^{n\times n}$, also `Input:` $x \in \mathbb{R}$ (Zahl), `Output:` $A \in \mathbb{R}^{d\times d}$ (Matrix). die Einträge von A heißen im folgenden immer $a_{k,m}$ (Eintrag in Zeile k, Spalte m).

Aus drei Gründen tun wir das: Für diese gilt ein analoges Resultat wie das oben (zeigen wir gleich); es sind fundamental wichtige Beispiele für Halbgruppen und somit wichtig in der Biomathematik (kommt in weiterführenden Vorlesungen dran); und hilft, Systeme von linearen DGL mit konstanten Koeffizienten zu lösen (kommt später). Dazu brauchen wir erstmal folgende Definition.

Definition 8.4. Sei $A \in \mathbb{R}^{d\times d}$. Dann ist

$$e^A := I + A + \frac{1}{2}A^2 + \frac{1}{6}A^3 + \cdots + \frac{1}{k!}A^k + \cdots$$

Mit $A^0 := I$, wobei I die $d \times d$-Einheitsmatrix bezeichnet, lässt sich das auch schreiben als

$$e^A = \sum_{n=0}^{\infty} \frac{1}{n!}A^n \tag{8.3}$$

Aber ist das sinnvoll? D.h., konvergiert die Reihe überhaupt?

Lemma 8.5. *Mit $\|A\| := \sum_{k,m=1}^{d} |a_{k,m}|$ ist $(\mathbb{R}^{d\times d}, +, \|\cdot\|)$ ein Banachraum.*

(Ohne Beweis. Zur Def von "Banachraum" siehe Bemerkung 4.10, S. 16)

Konvergenz heißt dann: Die Folge $A^{(0)}, A^{(1)}, A^{(2)}, \ldots$ konvergiert gegen die Matrix A, kurz: $A^{(n)} \to A$ $(n \to \infty)$, falls $a_{k,m}^{(n)} \to a_{k,m}$ für alle $1 \le k, m \le d$.

Aus den Normeigenschaften (nichtnegativ, Dreiecksungleichung) folgt:

$$\| \sum_{n=0}^{\infty} A_n \| \le \sum_{n=0}^{\infty} \|A_n\|$$

Für die Reihe in (8.3) folgt (unter Benutzung von $\|A^n\| = \|A\|^n$, ohne Beweis):

$$\| \sum_{n=0}^{\infty} \frac{1}{n!}A^n \| \le \sum_{n=0}^{\infty} \| \frac{1}{n!}A^n \| = \sum_{n=0}^{\infty} \frac{1}{n!}\|A\|^n = e^{\|A\|} < \infty$$

Die Funktion $E : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$, $E(A) = e^A$ heißt *Matrixexponentialfunktion*. Es folgen einige Eigenschaften der Matrixexponentialfunktion:

Lemma 8.6. *Für $A, B \in \mathbb{R}^{d\times d}$ gilt:*

1. $e^{0A} = I$
2. $e^A e^B = e^{A+B}$, *falls* $AB = BA$.
3. $(e^A)^{-1} = e^{-A}$

Proof. Zu 1.: Es ist $A^0 = I$ für *alle* Matrizen, also auch $(0A)^0 = I$. (Bei Zweifeln überlege man sich: Was ist 0^0 sinnvollerweise? Z.B. in $e^0 = 0^0 + 0^1 + \frac{1}{2}0^2 + \cdots$, oder in $0^0 = 0^{1-1} = \frac{0^1}{0^1}$?)

Zu 2.: Wir wissen, dass die Reihe absolut konvergiert, also ist die Summationsreihenfolge egal. Dann folgt

$$\begin{aligned} e^A e^B &= (I + A + \frac{1}{2}A^2 + \frac{1}{6}A^3 + \cdots)(I + B + \frac{1}{2}B^2 + \frac{1}{6}B^3 + \cdots) \\ &= I \quad + \quad A + B \quad + \quad \frac{1}{2}A^2 + AB + \frac{1}{2}B^2 \quad + \quad \frac{1}{6}A^3 + \frac{1}{2}A^2B + \frac{1}{2}AB^2 + \frac{1}{6}B^3 \quad + \cdots \\ &= I + (A+B) + \frac{1}{2}(A+B)^2 + \frac{1}{6}(A+B)^3 + \cdots = e^{A+B}. \end{aligned}$$

Dabei braucht man für das vorletzte "=" die Eigenschaft $AB = BA$, um die binomische Formel benutzen zu dürfen.

Zu 3: Folgt direkt aus 2 und 1. □

Wie sieht e^A nun konkret aus?

Beispiel 8.7. Für Diagonalmatrizen $D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_n \end{pmatrix}$ ist

$$D^k = \begin{pmatrix} d_1^k & 0 & \cdots & 0 \\ 0 & d_2^k & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_n^k \end{pmatrix}.$$

Also ist

$$e^D = \begin{pmatrix} e^{d_1} & 0 & \cdots & 0 \\ 0 & e^{d_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & e^{d_n} \end{pmatrix}$$

Beispiel 8.8. Für $A = \left(\begin{smallmatrix} 2 & 1 \\ 0 & 2 \end{smallmatrix}\right)$ ist e^A was? Erster Versuch: Es ist ja

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = D + N,$$

und es ist $DN = ND = \left(\begin{smallmatrix} 0 & 2 \\ 0 & 0 \end{smallmatrix}\right)$ (nachrechnen). Also können wir Lemma 8.6 2. benutzen. Aus dem letzten Bsp wissen wir bereits, dass $e^D = \left(\begin{smallmatrix} e^2 & 0 \\ 0 & e^2 \end{smallmatrix}\right)$ ist. Und wegen $N^2 = 0$, also auch $N^k = 0$, ist $e^N = I + N + 0 + 0 + \cdots = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. Also ist

$$e^A = e^{D+N} = e^D e^N = \begin{pmatrix} e^2 & 0 \\ 0 & e^2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} e^2 & e^2 \\ 0 & e^2 \end{pmatrix} = e^2 \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

(und der erste Versuch hat geklappt!)

Eine wichtige Grundlage für alles weitere ist nun folgendes Resultat:

Satz 8.9. *Die (namenlose) Funktion $\mathbb{R} \to \mathbb{R}^{d\times d}$, $t \mapsto e^{tA}$ ist diff-bar, genauer:*

$$\frac{d}{dt}e^{tA} = \lim_{h\to 0} \frac{e^{(t+h)A} - e^{tA}}{h} \text{ existiert, und ist gleich } Ae^{tA}.$$

Proof. Es gilt

$$\begin{aligned}\frac{e^{(t+h)A} - e^{tA}}{h} &= \frac{e^{hA}e^{tA} - e^{tA}}{h} = \frac{e^{hA} - I}{h}e^{tA} = \big(-\frac{1}{h}I + \frac{1}{h}I + \frac{h}{h}A + \frac{h^2}{2h}A^2 + \frac{h^3}{6h}A^3 + \cdots\big)e^{tA} \\ &= A\big(I + \frac{h}{2}A + \frac{h^2}{6}A^2 + \cdots\big)e^{tA},\end{aligned}$$

und für $h \to 0$ gehen alle Summanden in der Klammer gegen 0, außer das I. Die Summe ist absolut konvergent, stehen bleibt also Ae^{tA}. □

Nun können wir die folgende Frage beantworten: Für welche matrixwertigen Funktionen $f : \mathbb{R} \to \mathbb{R}^{d\times d}$ gilt das Analogon von (8.1), also

$$f : \mathbb{R} \to \mathbb{R}^{d\times d}, \quad f(x)f(y) = f(x+y), \quad f(0) = I\,? \tag{8.4}$$

Wieder fragen wir nach allen stetigen Lösungen. Mit Lemma 8.6 ist klar:

Funktionen der Form $f(x) = e^{xA}$ sind Lösungen von (8.4).

A ist dabei eine beliebige Matrix in $\mathbb{R}^{d\times d}$. Denn wegen $xAyA = xyA^2 = yAxA$ gilt:

$$e^{xA}e^{yA} = e^{xA+yA} = e^{(x+y)A}.$$

Genau wie im eindimensionalen Fall kann man nun zeigen:

- Falls f stetige Lösung von (8.4) ist, ist f auch diff-bar (vgl Prop 8.2).
- Jede diff-bare Lösung von (8.4) ist auch Lösung des (mehrdim) AWPs

  $$\frac{d}{dx}f(x) = Af(x), \quad f(0) = I,$$

  für ein geeignetes A (vgl Prop 8.2).
- Die eindeutige Lösung des AWP ist von der Form $f(x) = e^{xA}$ (vgl Prop 8.1).

(Die Beweise gehen fast analog zum eindimensionalen Fall; stehen auch in [EN]). Damit erhalten wir das analoge Resultat zu Satz 8.3:

Satz 8.10. *Sei f stetige Lösung von* (8.4). *Dann ist $f(x) = e^{xA}$, mit $A \in \mathbb{R}^{d\times d}$.*

Die Matrixexponentialfunktion, sowie die Tatsache, dass sie die Funktionalgleichung (8.4) löst, gibt ihr eine fundamental wichtige Rolle in der Theorie der Evolutionsgleichungen. Genauer spielen dort oft bestimmte *Halbgruppen* eine Rolle. Ein Beispiel für eine solche ist eben $\{e^{tA} \,|\, t \geq 0\}$ mit Matrizenmultiplikation. Die Matrix A heißt dann *Erzeuger* (engl generator) dieser Halbgruppe.

8.2 Lineare DGL mit konstanten Koeffizienten

Wie erwähnt lassen mittels DGL bzw AWP viele natürliche Vorgänge beschreiben. Nehmen wir ein ganz **naives Räuber-Beute-Modell**: In einem See leben zur Zeit $t = 0$ ein Paar Ungeheuerfische und 5 Paare Vegetarierfische. Die jeweilige Zahl der Paare zur Zeit t nennen wir $u(t)$ bzw $v(t)$. Je mehr Fische des gleichen Typs es gibt, desto höher soll die Geburtenrate sein. Außerdem sinkt die Geburtenrate der Vegetarierfische, je mehr Ungeheuerfische es gibt; und umgekehrt wirkt sich die Zahl der Vegetarierfische auf die Geburtenrate der Ungeheuerfische positiv aus. Das führt zum AWP

$$\frac{d}{dt}u(t) = 2u(t) + v(t), \quad \frac{d}{dt}v(t) = -u(t) + 2v(t), \quad u(0) = 1,\ v(0) = 5 \tag{8.5}$$

Das Schöne ist nun: Das können wir fast sofort lösen. Schreiben wir das als Matrix-Vektor-Gleichung, dann steht da nämlich:

$$\begin{pmatrix} u'(t) \\ v'(t) \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} u(t) \\ v(t) \end{pmatrix}$$

Setzen wir nun $f(t) = e^{tA} \cdot \left(\begin{smallmatrix} 1 \\ 5 \end{smallmatrix}\right)$ mit $A = \left(\begin{smallmatrix} 2 & 1 \\ -1 & 2 \end{smallmatrix}\right)$. Das ist eine Funktion von $\mathbb{R} \to \mathbb{R}^2$, also von der Form $f(t) = \left(\begin{smallmatrix} u(t) \\ v(t) \end{smallmatrix}\right)$. Dann ist

$$f(0) = \begin{pmatrix} u(0) \\ v(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 5 \end{pmatrix}$$

und

$$\begin{pmatrix} u'(t) \\ v'(t) \end{pmatrix} = \frac{d}{dt} f(t) = \frac{d}{dt}\left(e^{tA} \begin{pmatrix} 1 \\ 5 \end{pmatrix}\right) = \left(\frac{d}{dt} e^{tA}\right) \begin{pmatrix} 1 \\ 5 \end{pmatrix} = A e^{tA} \begin{pmatrix} 1 \\ 5 \end{pmatrix} = A f(t) = A \begin{pmatrix} u(t) \\ v(t) \end{pmatrix}$$

Also ist dieses f Lösung des AWP. Jetzt müssen wir nur noch verstehen, was dieses f konkret bedeutet. Dazu zerlegen wir A wieder in eine Summe, hier:

$$A = \begin{pmatrix} 2 & 1 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = D + C$$

Wieder gilt $DC = CD$ (nachrechnen). e^{tD} kennen wir bereits im Prinzip aus Bsp 8.8, es ist $e^{tD} = e^{2t} I$. Für e^{tC} gilt:

$$e^{tC} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + t \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} + \frac{t^2}{2!} \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} + \frac{t^3}{3!} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} + \frac{t^4}{4!} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{t^5}{5!} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} + \cdots$$

Als erstes stellen wir fest, dass nur 0te, der zweite, der vierte... also die $2n$ten Summanden etwas zu den Einträgen oben links und unten rechts beitragen, und nur die $2n+1$sten Summanden zu den anderen Einträgen. Sodann erinnern wir uns an die Potenzreihen des Sinus und des Cosinus:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!},$$
$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}.$$

Damit folgt nun: $e^{tC} = \left(\begin{smallmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{smallmatrix}\right)$. Unser gesuchtes e^{tA} ist damit

$$e^{tA} = e^{t(D+C)} = e^{tD+tC} = e^{tD} e^{tC} = \begin{pmatrix} e^{2t} & 0 \\ 0 & e^{2t} \end{pmatrix} \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix} = e^{2t} \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix}.$$

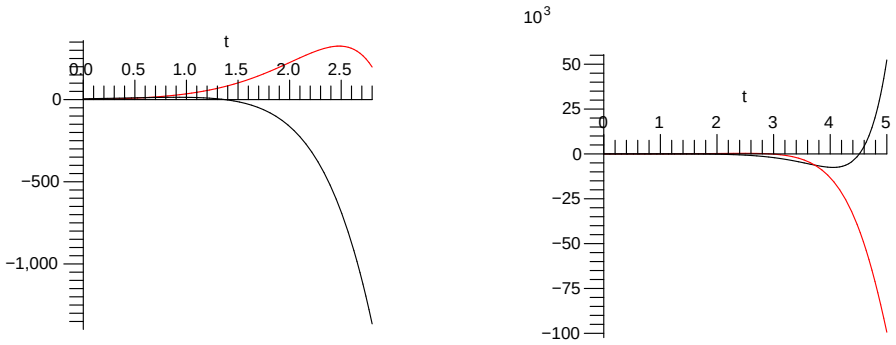


Figure 13: Die Lösungen unseres naiven Räuber-Beute-Modells. In rot (bzw grau): Räuber, in schwarz: Beute.

Um nun unser gesuchtes f zu erhalten, müssen wir noch mit dem Anfangswertvektor $\binom{1}{5}$ multiplizieren und erhalten:

$$e^{2t}\begin{pmatrix}\cos(t) & \sin(t)\\ -\sin(t) & \cos(t)\end{pmatrix}\begin{pmatrix}1\\5\end{pmatrix} = e^{2t}\begin{pmatrix}\cos(t)+5\sin(t)\\ 5\cos(t)-\sin(t)\end{pmatrix}$$

Also ist

$$u(t) = e^{2t}\big(\cos(t)+5\sin(t)\big), \quad v(t) = e^{2t}\big(5\cos(t)-\sin(t)\big).$$

Nun können wir eine Probe machen: Z.B. ist

$$u'(t) = 2e^{2t}\big(\cos(t)+5\sin(t)\big) + e^{2t}\big(-\sin(t)+5\cos(t)\big) = 2u(t)+v(t),$$

wie es sein soll. Allerdings zeigt diese Lösung ein seltsames Verhalten: In Bild 13 sind die Funktionen u (rot bzw grau) und v (schwarz) gezeigt. Nach unserem Modell ist es so, dass es zur Zeit $t = 2$ etwa 200 Paare von Ungeheuerfischen gibt, aber -200 Vegetarierfischpaare! Und zur Zeit $t = 5$ gibt es plötzlich über 50.000 Vegetarierfischpaare, aber -100.000 Ungeheuerfischpaare! Wir können dieses Modell also getrost wegschmeißen.

Bemerkung 8.11. Es gibt PDE-Modelle, die ganz ausgezeichnet das Räuber-Beute-Verhältnis in Populationen modellieren, etwa die *Lotka-Volterra-Gleichungen*. Dessen Lösungen stimmen nicht nur ganz prächtig mit echten Beobachtungen überein, sondern erlauben auch (qualitativ) präzise Vorhersagen.

Wir formulieren zum Abschluss noch das allgemeine Resultat, auf dem die Überlegungen in diesem Abschnitt beruhen.

Satz 8.12. *Sei* $A \in \mathbb{R}^{d\times d}$, $f(t) = \big(f_1(t), f_2(t), \ldots, f_d(t)\big)^T$. *Das AWP*

$$\frac{d}{dt}f(t) = Af(t), \quad f(0) = (v_1,\ldots,v_d)^T$$

hat die eindeutige Lösung $f(t) = e^{tA}(v_1,\ldots,v_d)^T$.

Um an konkrete Lösungen zu kommen, muss man das e^{tA} in "richtige" Funktionen übersetzen. Das ist etwas aufwendig, es gibt aber Standardmethoden, die in jedem Fall schöne Darstellungen der Lösung liefern.

Im Rahmen dieser Vorlesung können wir dieses letzte Thema — Systeme linearer Differentialgleichungen mit konstanten Koeffizienten — nur kurz anschneiden. Die weiteren Details stehen z.B. sehr schön in [H4], Kap 49-51.

References

[AHKKLS] T. Arens, F. Hettlich, Ch. Karpfinger, U. Kockelkorn, K. Lichtenegger, H. Stachel: *Mathematik* (Spektrum Verlag 2008)

[D] A. Deitmar: *A First Course in Harmonic Analysis*, (Springer 2005)

[EN] K.J. Engel, R. Nagel: *One Parameter Semigroups for Linear Evolution Equations*, (Springer 2000)

[EV] L.C. Evans: *Partial Differential Equations*, (AMS 1998)

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

[H1] H. Heuser: *Lehrbuch der Analysis I*, (Teubner, versch. Auflagen)

[H2] H. Heuser: *Lehrbuch der Analysis II*, (Teubner, versch. Auflagen)

[H3] H. Heuser: *Funktionalanalysis*, (Teubner, versch. Auflagen)

[H4] H. Heuser: *Gewöhnliche Differentialgleichungen*, (Teubner, versch. Auflagen)

[PIN] M. Pinsky: *Introduction to Fourier Analysis and Wavelets*, Brooks/Cole (2002)

[PLA] R. Plato: *Numerische Mathematik kompakt* (Vieweg und Teubner 2006)

[WIK] Online: `http://en.wikipedia.org`