Nachweis der Konvergenz; außerdem war sie oft nützlich, um ohne großen Aufwand neue FOURIER-Reihen aus bekannten herzuleiten.

Hier im nichtperiodischen Fall ist sie einfacher und anschaulicher zu verstehen als im periodischen Fall: $f(\star g)(t)$ ist einfach das gewichtete Mittel der Funktionswerte von f in der Umgebung von t, wobei g die Gewichtsfunktionen ist. Am einfachsten ist es, wenn man sich g als eine Funktion vorstellt, die im Punkt Null ein Maximum hat und dann nach beiden Seiten monoton abfällt; dann kann man sich $f \star g$ als eine „verschmierte" (oder auch geglättete) Version von f vorstellen. Indem man für $g(t)$ GAUSSsche Glockenkurven nimmt, kann man beispielsweise unscharfe (oder weichgezeichnete) Photographien simulieren – je größer der Parameter σ, desto unschärfer ist das Resultat.

Für die formale Definition lassen wir allerdings beliebige Funktionen f und g zu; später werden wir sogar Faltungen von Funktionen mit Distributionen betrachten.

Definition: Für zwei Funktionen $f, g\colon \mathbb{R} \to \mathbb{C}$ heißt

$$f \star g\colon \begin{cases} \mathbb{R} \to \mathbb{C} \\ t \mapsto \displaystyle\int_{-\infty}^{\infty} f(t-s)g(s)\,ds \end{cases},$$

falls dieses Integral existiert, *Faltung* von f mit g.

Lemma: Für $f, g \in \mathrm{L}^2(\mathbb{R}, \mathbb{C})$ existiert die Faltung $f \star g$.

Beweis: Mit f liegt für jedes $t \in \mathbb{R}$ auch die Funktion $s \mapsto f(t-s)$ in $\mathrm{L}^2(\mathbb{R}, \mathbb{C})$; die Abschätzungen aus §8*a)* zeigen daher die Existenz des Integrals $f \star g(t)$. ∎

Ebenfalls in völliger Analogie zum periodischen Fall gilt

Lemma: Falls die FOURIER-Transformationen von f, g und von $h(t) = (f \star g)(t)$ als Funktionen existieren, ist $\widehat{h}(\omega) = \widehat{f}(\omega) \cdot \widehat{g}(\omega)$.

Beweis: Nach dem Satz von FUBINI ist

$$\begin{aligned}
\widehat{h}(\omega) &= \int_{-\infty}^{\infty} h(t)e^{-i\omega t}\,dt = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(t-s)\,g(s)\,ds \right) e^{-i\omega t}\,dt \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(t-s)\,g(s)e^{-i\omega t}\,dt \right) ds \\
&\underset{u=t-s}{=} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(u)\,g(s)e^{-i\omega(u+s)}\,du \right) ds \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(u)e^{-i\omega u}\,du \right) g(s)e^{-i\omega s}\,ds \\
&= \left(\int_{-\infty}^{\infty} f(u)e^{-i\omega u}\,du \right) \cdot \left(\int_{-\infty}^{\infty} g(s)e^{-i\omega s}\,ds \right) = \widehat{f}(\omega) \cdot \widehat{g}(\omega)\,,
\end{aligned}$$

wie behauptet. ∎

Als erste Anwendung hiervon können wir die FOURIER-Transformierte eines Produkts durch die FOURIER-Transformierten der Faktoren ausdrücken:

Korollar: $\widehat{fg}(\omega) = \dfrac{1}{2\pi}(\widehat{f} \star \widehat{g})(\omega)\,.$

Beweis: Wir wenden das gerade bewiesenen Lemma an auf die FOURIER-Transformierten von f und g; dann ist

$$\widehat{\widehat{f} \star \widehat{g}}(t) = \widehat{\widehat{f}}(t) \cdot \widehat{\widehat{g}}(t)\,.$$

Wie wir wissen, unterscheiden sich FOURIER-Transformation und inverse FOURIER-Transformation durch den Faktor $1/2\pi$ vor der inversen Transformation und das Vorzeichen des Argument, d.h.

$$\widehat{\widehat{f}}(t) = 2\pi \cdot f(-t), \quad \widehat{\widehat{g}}(t) = 2\pi \cdot g(-t) \quad \text{und} \quad \widehat{\widehat{f} \star \widehat{g}}(t) = 4\pi^2 \cdot f(-t)g(-t)\,.$$

Aus dem gleichen Grund ist $\widehat{\widehat{f \star g}}(t) = 2\pi \cdot (\widehat{f} \star \widehat{g})(-t)$, also

$$2\pi \cdot (\widehat{f} \star \widehat{g})(-t) = 4\pi^2 \cdot \widehat{fg}(-t) \quad \text{oder} \quad \widehat{fg}(-t) = \frac{1}{2\pi}(\widehat{f} \star \widehat{g})(-t)\,.$$

Dies gilt für alle reellen Zahlen t, deshalb können wir das Minuszeichen links und rechts auch weglassen und haben dann die Behauptung des Korollars. ■

Wie im periodischen Fall folgt auch, daß die Faltung (abgesehen von eventuell vorhandenen Unstetigkeitsstellen) kommutativ und assoziativ ist:

$$f \star g = g \star f \quad \text{und} \quad f \star (g \star h) = (f \star g) \star h\,,$$

denn die FOURIER-Transformationen der beiden Seiten sind jeweils gleich nach dem Kommutativgesetz und Assoziativgesetz für die Multiplikation komplexer Zahlen.

Eine weitere interessante Konsequenz dieses Lemmas ist, daß sich Faltungen gelegentlich rückgängig machen lassen: $f \star g$ ist durch seine FOURIER-Transformation $\widehat{f} \cdot \widehat{g}$ (fast überall) bestimmt; falls $g(\omega)$ keine Nullstellen hat, kann man die Multiplikation mit $g(\omega)$ durch eine Division rückgängig machen. Eine Grundidee zum Rückgängigmachen der Faltung wäre also die folgende: Ist $h(t)$ die inverse FOURIER-Transformation von $1/\widehat{g}(\omega)$, so hat $(f \star g) \star h$ FOURIER-Transformierte

$$\widehat{f}(\omega) \cdot \widehat{g}(\omega) \cdot \widehat{h}(\omega) = \widehat{f}(\omega) \cdot \widehat{g}(\omega) \cdot \frac{1}{\widehat{g}(\omega)} = \widehat{f}(\omega)\,,$$

$(f \star g) \star h$ stimmt also fast überall mit f überein.

Leider ist die Sache aber doch nicht ganz so einfach, denn die Existenz von h ist alles andere als klar: Für eine stark abfallende Funktion $g(\omega)$ ist $1/g(\omega)$ „stark ansteigend“, und natürlich gibt es auch Probleme mit den Nullstellen von g. Die Mathematik kennt jedoch eine ganze Reihe von Regularisierungstechniken, mit denen man solche Probleme umgehen kann. Insbesondere kann man für praktische Zwecke sowohl den Frequenzbereich, über den integriert wird, als auch den Zeit- oder Ortsbereich oft abschneiden, so daß nur ein Integral über ein endliches Intervall betrachtet werden muß.

Die Formel, die wir gerade benutzt haben, gelten, wenn man solche Techniken benutzt, natürlich nicht mehr exakt, aber doch oft mit einer Genauigkeit, die für praktische Zwecke völlig ausreicht. So konnte beispielsweise die NASA die Bilder des falsch fokusierten HUBBLE-Teleskops durch digitale Nachbehandlung so deutlich verbessern, daß die Bildqualität auch vor der Reparatur nicht viel schlechter war als bei einem korrekt fokusierten Teleskop.

Eine neuere Anwendung ist die sogenannte *brennpunktfreie Optik,* die von CMD Optics in Boulder, Colorado entwickelt wurde. Dort benutzt man eine (von Zeiss speziell zu diesem Zweck konstruierte) Linse ohne Brennpunkt; parallele einfallende Strahlen gehen also *nie* durch denselben Punkt der Bildebene, so daß grundsätzlich jedes Bild unscharf ist. Diese Unschärfe wird durch digitale Nachbearbeitung in der oben skizzierten Weise so gut es geht kompensiert.

Zweck dieser auf den ersten Blick unsinnigen Vorgehensweise ist die Erhöhung der Tiefenschärfe: Ein klassisches optisches System bildet, insbesondere wenn es mit wenig Licht auskommen muß und daher eine große Blende braucht, nur in einem sehr kleinen Entfernungsbereich scharf ab. Die brennpunktfreie Linse bildet natürlich überhaupt nirgends scharf ab, aber das Gesamtsystem aus Linse und digitaler Nachbearbeitung liefert scharfe Bilder aus einem deutlich größeren Entfernungsbereich als dies mit konventioneller Optik möglich ist.

Besonders einfach sind Faltungen mit δ-Funktionen zu berechnen: Für $\eta(t) = \delta(t - t_0)$ zeigt die Substitutionsregel mit $u = t - t_0 - s$, daß

$$\eta \star f = \int_{-\infty}^{\infty} \delta(t - t_0 - s) g(s)\, ds = \int_{-\infty}^{\infty} \delta(u) g(t - t_0 - u)\, du = g(t - t_0)$$

ist, Faltung mit $\delta(t - t_0)$ verschiebt also einfach das Argument um t_0. Insbesondere ist $\delta \star f = f$.

Im Falle einer Funktion, die außerhalb eines gewissen Intervalls null (oder praktisch null) ist, läßt sich durch Faltung mit einer Summe von δ-Funktionen der Graph an verschiedene Stellen verschieben; Abbildung 25 zeigt dies für die Faltung eines (fett eingezeichneten) Dreieck-

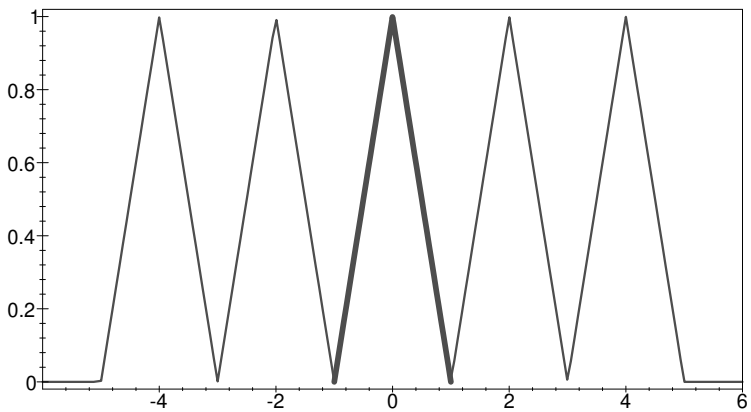

Abb. 25: Faltung eines Dreiecksimpuls mit einer Summe von δ-Distributionen

simpulses auf $[-1,\ 1]$ und die Distribution

$$\delta(t-4)+\delta(t-2)+\delta(t)+\delta(t+2)+\delta(t+4)\,.$$

Zum Schluß wollen wir uns noch kurz überlegen, wie sich Faltungen bezüglich der LAPLACE-Transformation verhalten.

Die LAPLACE-Transformierte einer Funktion f an der Stelle $s = r + i\omega$ ist die FOURIER-Transformierte von

$$f_r\colon \begin{cases} \mathbb{R} \to \mathbb{C} \\ t \mapsto \begin{cases} 0 & \text{für } t < 0 \\ f(t)e^{-rt} & \text{für } t \geq 0 \end{cases} \end{cases}$$

an der Stelle ω. Für zwei Funktionen f, g haben $(fg)_r$ und $f_r g_r$ offensichtlich wenig miteinander zu tun; wir können also keine schöne Formel für $\mathcal{L}\{f(t)g(t)\}(s)$ in Abhängigkeit von $\mathcal{L}\{f(t)\}(s)$ und $\mathcal{L}\{g(t)\}(s)$ erwarten.

Anders sieht es aus für das Produkt $\mathcal{L}\{f(t)\}(s) \cdot \mathcal{L}\{f(t)g(t)\}(s) = \widehat{f}_r(\omega)\widehat{g}_r(\omega)$: Wie wir gerade gesehen haben, ist die die FOURIER-

Transformierte von

$$f_r \star g_r(t) = \int_{-\infty}^{\infty} f_r(t-s)g_r(s)\,ds = \int_0^{\infty} f_r(t-s)g_r(s)\,ds\,,$$

da $g_r(s)$ für negative s verschwindet. Für negative t ist $f_r(t-s)$ im gesamten Integrationsbereich null, also verschwindet das Integral. Für positive Werte von s wissen wir nur, daß $f_r(t-s)$ für $s > t$ verschwindet; hier ist also

$$\begin{aligned} f_r \star g_r(t) &= \int_0^t f_r(t-s)g_r(s)\,ds = \int_0^t f(t-s)e^{-r(t-s)}\,g(s)e^{-rs}\,ds \\ &= \int_0^t f(t-s)g(s)e^{-rt}\,ds = e^{-rt}\int_0^t f(t-s)g(s)\,ds\,, \end{aligned}$$

da der Faktor e^{-rt} nicht von der Integrationsvariablen abhängt. Mit der Funktion

$$h(t) \underset{\text{def}}{=} f * g(t) \underset{\text{def}}{=} \int_0^t f(t-s)g(s)\,ds$$

ist somit $f_r \star g_r = h_r$; mit LAPLACE-Transformationen ausgedrückt ist das die Regel

$$\mathcal{L}\{f * g(t)\}(s) = \mathcal{L}\{f(t)\}(s) \cdot \mathcal{L}\{g(t)\}(s)\,.$$

Man beachte, daß wir hier ein etwas anderes Faltungsprodukt benutzen müssen als bisher; es stimmt nur dann mit dem üblichen überein, wenn sowohl $f(t)$ als auch $g(t)$ für negative Argumente verschwinden.

h) Der Abtastsatz von Nyquist

Egal ob es um die automatische Erfassung von Meßwerten geht oder um die Aufzeichnung von Musik: Die digitale Darstellung analoger Daten ist wesentlicher Bestandteil der Informationsverarbeitung. Nun ist aber eine beliebige Funktion $f\colon \mathbb{R} \to \mathbb{R}$ sicherlich nicht durch ihre Funktionswerte an endlich vielen Stellen oder auch an ein einer diskreten

Menge von Stellen bestimmt: Auch wenn wir wissen, daß $f(t) = 0$ ist für jedes ganzzahlige Vielfache von 0,001, wissen wir noch nicht, daß f die Nullfunktion ist: Auch die Funktionen $f(t) = \sin(1000\pi t)$ und $f(t) = -3\sin(5000\pi t)$ haben diese Eigenschaft. Auch bei von null verschiedenen Abtastwerten tritt dieses Problem auf: Beispielsweise stimmen auch die Funktionen $f(t) = \cos(500\pi t)$ und $g(t) = \cos(1500\pi t)$ für alle ganzzahligen Vielfachen von 0,001 überein, aber sie nehmen hier abwechselnd die Werte $1, 0, -1, 0$ an; siehe Abbildung 26. Da der Frequenzunterschied zwischen den beiden Schwingungen fast dem zwischen Baß und Sopran entspricht, ist klar, daß man die beiden Schwingungen zumindest auf einer Musik-CD nicht miteinander verwechseln darf.

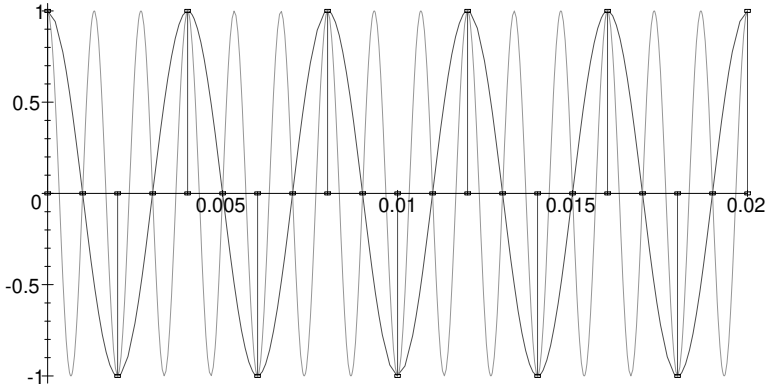

Abb. 26: Abtastung zweier Schwingungen

Die Probleme bei den obigen Beispielen beruhen offensichtlich darauf, daß es zu jedem gegebenen Signal auch höherfrequente Signale gibt, die an vorgegebenen Abtastpunkten mit ihm übereinstimmen; eine eindeutige Rekonstruktion ist höchstens dann möglich, wenn man eine Grenze festlegt, oberhalb derer Frequenzen nicht mehr berücksichtigt werden sollen. Der Abtastsatz von NYQUIST sagt, daß dann in der Tat eine Rekonstruktion möglich ist, und er sagt auch, wo die Grenze liegen soll, oberhalb derer man die Frequenzen abschneiden muß: Die Abtastfrequenz muß mehr als doppelt so hoch sein als die höchste im Signal vorkommende Frequenz.

Die genaue Formulierung des Satzes ist etwas technischer; insbesondere müssen wir berücksichtigen, daß die Kreisfrequenz ω, mit der wir immer arbeiten, etwas anderes ist, als die Frequenz: Eine reine Schwingung mit einer Frequenz von 1000 Hz ist nicht gegeben durch eine Funktion wie $\sin 1000t$, sondern – bei in Sekunden gemessener Zeit – durch $\sin 2000\pi t$. Entsprechend kommt auch jetzt bei der Formulierung des Abtastsatzes von NYQUIST ein Faktor 2π ins Spiel:

Satz: $f \in L^2(\mathbb{R}, \mathbb{C})$ habe die Eigenschaft, daß $\widehat{f}(\omega)$ außerhalb eines Intervalls der Länge Ω verschwinde. Dann ist f eindeutig bestimmt durch die Werte $f(2k\pi/\Omega)$ mit $k \in \mathbb{Z}$.

Beweis: (ω_1, ω_2) sei ein Intervall der Länge Ω derart, daß $\widehat{f}(\omega)$ außerhalb dieses Intervalls verschwindet. Dann ist bis auf eine Nullfunktion

$$f(t) = \overset{\times}{\widehat{f}}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{f}(\omega)e^{i\omega t}\,d\omega = \frac{1}{2\pi}\int_{\omega_1}^{\omega_2} \widehat{f}(\omega)e^{i\omega t}\,d\omega\,.$$

Indem wir f durch die rechte Seite ersetzen (was nichts wesentliches ändert) können wir annehmen, daß diese Gleichung wirklich gilt. Also ist insbesondere

$$f\left(\frac{2k\pi}{\Omega}\right) = \frac{1}{2\pi}\int_{\omega_1}^{\omega_2} \widehat{f}(\omega)e^{2k\pi i\omega/\Omega}\,d\omega\,. \qquad (*)$$

Nun betrachten wir jene Funktion $g(\omega)$, die auf dem Intervall $[\omega_1, \omega_2)$ mit $\widehat{f}(\omega)$ übereinstimmt und die periodisch mit Periode Ω in ω auf $\mathbb{R}$ fortgesetzt ist. Für diese Funktion ist natürlich auch

$$f\left(\frac{2k\pi}{\Omega}\right) = \frac{1}{2\pi}\int_{\omega_1}^{\omega_2} g(\omega)e^{2k\pi i\omega/\Omega}\,d\omega\,,$$

denn im Integrationsintervall stimmen $\widehat{f}$ und g überein.

g als periodische Funktion in ω mit Periode Ω hat eine Darstellung als FOURIER-*Reihe*

$$\sum_{k=-\infty}^{\infty} c_k e^{ik\lambda\omega} \quad \text{mit} \quad \lambda = \frac{2\pi}{\Omega}\,;$$

der k-te FOURIER-Koeffizient ist

$$\begin{aligned} c_k &= \frac{1}{\Omega}\int_{\omega_1}^{\omega_2} g(\omega)e^{-ik\lambda\omega}\,d\omega = \frac{1}{\Omega}\int_{\omega_1}^{\omega_2} g(\omega)e^{-2\pi i\omega k/\Omega}\,d\omega \\ &= \frac{2\pi}{\Omega} f\left(\frac{-2k\pi}{\Omega}\right)\,, \end{aligned}$$

wobei das letzte Gleichheitszeichen wegen $(*)$ gilt.

Durch die Werte $f\left(\frac{2k\pi}{\Omega}\right)$ sind also alle FOURIER-Koeffizienten von g bestimmt, damit auch (fast überall) die Funktion $g(\omega)$, und damit auch die Funktion $\widehat{f}(\omega)$, die im Intervall $[\omega_1, \omega_2)$ mit $g(\omega)$ übereinstimmt und außerhalb (außer eventuell im Punkte ω_2) verschwindet. Damit ist auch $f(t) = \check{\widehat{f}}(t)$ fast überall durch diese Werte bestimmt. ■

HARRY NYQUIST (1889–1976) wurde in Schweden geboren, arbeitete aber ab Anfang der zwanziger Jahre bei den Bell Labaratories; das Bild zeigt ihn um 1960 mit seinen dortigen Kollegen JOHN PIERCE *(links)* und RUDOLF KOMPFNER *(Mitte)*. Seine Arbeit von 1924 über die Übertragungsgeschwindigkeit von Telegraphen gilt als eine der Begründungen der Informationstheorie. Den Abtastsatz, den CAUCHY bereits 1841 postuliert hatte, bewies er 1928. Weitere wichtige Arbeiten befassen sich mit der quantitativen Erforschung des thermischen Rauschens und der Stabilität von Verstärkern.

Bei praktischen Anwendungen dieses Satzes wird $f(t)$ im allgemeinen

eine reelle Funktion sein; dann verschwindet

$$\begin{aligned} \widehat{f}(-\omega) &= \int_{-\infty}^{\infty} f(t)e^{-i(-\omega)t}\,dt = \int_{-\infty}^{\infty} f(t)e^{i\omega t}\,dt = \int_{-\infty}^{\infty} \overline{f(t)e^{-i\omega t}}\,dt \\ &= \overline{\int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,dt} = \overline{\widehat{f}(\omega)} \end{aligned}$$

genau dann, wenn auch $\widehat{f}(\omega)$ verschwindet. Daher wird in diesem Fall alles einfacher, wenn man das Intervall, außerhalb dessen $\widehat{f}(\omega)$ verschwindet, symmetrisch zum Nullpunkt wählen kann, also von der Form $(-\omega_0, \omega_0)$. Die zur Kreisfrequenz $\omega_0 = 2\pi\nu_0$ gehörende Frequenz ν_0 wird in diesem Zusammenhang oft als *Bandbreite* bezeichnet. Hier ist $\Omega = 2\omega_0$, zur Rekonstruktion der Funktion f brauchen wir also die Funktionswerte

$$f\left(\frac{2k\pi}{\Omega}\right) = f\left(\frac{k\pi}{\omega_0}\right) \quad \text{mit} \quad k \in \mathbb{Z}\,.$$

Ein Signal der Bandbreite ν_0 muß also mit einer Frequenz von mindestens $2\nu_0$ abgetastet werden, damit man es eindeutig rekonstruieren kann.

Bekanntestes Beispiel hierfür sind Musik-CDs: Praktisch niemand kann Töne mit Frequenzen von mehr als 20kHz hören; für Aufnahmen auf CD wird 44 100 Mal pro Sekunde der Schalldruck gemessen und gespeichert, für Signale die nicht allzu weit oberhalb von 20kHz abgeschnitten werden, ist also eine perfekte Rekonstruktion möglich.

Auch in der Computergraphik spielt der Satz von NYQUIST eine wichtige Rolle, denn Pixelgraphik ist schließlich nichts anderes als die (zweidimensionale) diskrete Abtastung eines kontinuierlichen Bilds. Falls das Bild zu hochfrequente Anteile enthält, entstehen sogenannte *alias-Effekte,* da das Auge diese Anteile anhand des Pixelbilds als niedrigerfrequente Strukturen mit gleichen Abtastwerten interpretiert. Vor der Abtastung muß das Bild daher tiefpaßgefiltert werden; da die Funktion $\frac{\sin ax}{ax}$, die FOURIER-Transformierte des Rechteckimpulses, einigermaßen schnell abfällt, wendet man dazu meist das Lemma aus dem letzten

Abschnitt an und faltet mit einer geeigneten solchen Funktion. Falls das Ursprungsbild auch schon als (höher aufgelöste) Pixelgraphik gegeben war, wird die Faltung hier einfach zu einer Summation über nicht garzu viele Nachbarpixel, was sehr effizient durchgeführt werden kann.

§9: Ausblick: Mehrdimensionale Fourier-Theorie

a) Faltungen und Fourier-Integrale

In völliger Analogie zum eindimensionalen Faltungsintegral läßt sich auch ein n-dimensionales definieren: Für zwei Funktionen $f, g\colon \mathbb{R}^n \to \mathbb{C}$ definieren wie die Faltung als

$$f \star g = \int \underset{\mathbb{R}^n}{\cdots} \int f(x_1 - y_1, \ldots, x_n - y_n) g(y_1, \ldots, y_n)\, dy_1 \ldots dy_n$$

– sofern dieses Integral existiert.

Auch die anschauliche Interpretation ist dieselbe wie im eindimensionalen Fall: Wenn wir f als eine Gewichtsfunktion auffassen, ist $f \star g$ ein gewichtetes Mittel über Werte von g; für

$$f(x_1, \ldots, x_n) = \frac{1}{\pi^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^n x_i^2}$$

etwa, die n-dimensionale GAUSS-Funktion, entspricht das im Fall $n = 2$ einem je nach Größe von σ mehr oder weniger defokusierten Bild.

Durch mehrdimensionale Faltungen mit δ-Distributionen lassen sich Verschiebungungen realisieren: Beispielsweise wäre, wenn der Satz von FUBINI in einer solchen Situation anwendbar wäre,

$$\iint\limits_{\mathbb{R}^2} \delta(x-a)\delta(y-b) f(x,y)\, dx\, dy = f(x-a, y-b)\,,$$

und genau so *definieren* wir die Interpretation der *a priori* sinnlosen linken Seite.

(Man beachte, daß Ausdrücke wir $\delta(x-a)\delta(x-b)$ oder $\delta(x)^2$ weiterhin sinnlos bleiben, egal ob sie unter einem oder mehreren Integralzeichen stehen.)

Ist also $f\colon \mathbb{R}^2 \to \mathbb{R}$ eine Funktion, die (z.B. durch Grauwerte) ein Bild definiert und die außerhalb des Bereichs $0 \le x, y \le 1$ verschwindet, so ist mit der Distribution

$$\eta(x,y) = \sum_{k=1}^N \sum_{\ell=1}^M \delta(x-k)\delta(y-\ell)$$

die Faltung $\eta \star f$ ein Bilderbogen aus NM Exemplaren dieses Bildes. Abbildung 27 zeigt dies für den Graph einer zweidimensionalen Normalverteilung.

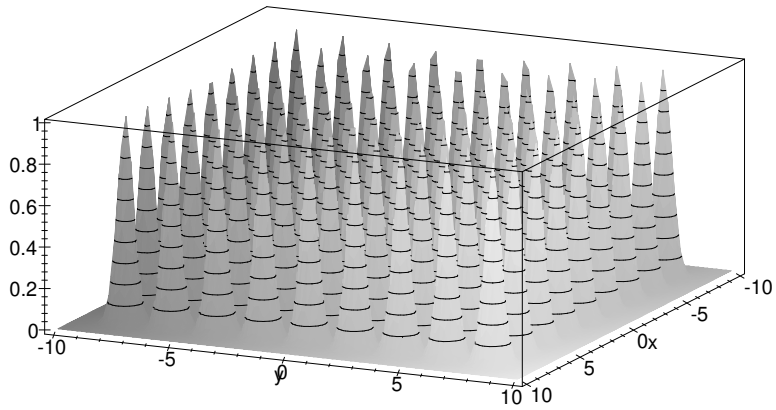

Abb. 27: Eine zweidimensionale Faltung

Auch die FOURIER-Transformation läßt sich in völliger Analogie zum eindimensionale Fall auf beliebige Dimensionen verallgemeinern: Für $f\colon \mathbb{R}^n \to \mathbb{C}$ definieren wir

$$\widehat{f}\colon \begin{cases} \mathbb{R}^n \to \mathbb{C} \\ (\omega_1, \ldots, \omega_n) \mapsto \int \underset{\mathbb{R}^n}{\cdots} \int f(x_1, \ldots, x_n) e^{-i\sum_{k=1}^n \omega_i x_i}\, dx_1 \cdots dx_n \end{cases}.$$

Da es nur eine Zeit gibt, läßt sich dies nicht als Zerlegung eines zeitlichen Signals in seine Frequenzen interpretieren; die x_i sollte man sich hier als *räumliche* Koordinaten vorstellen. Beispiele dazu folgen im nächsten

Abschnitt, wo wir eine Anwendung solcher räumlicher FOURIER-Transformationen betrachten.

Wenigsten kurz sei noch angedeutet, wie man auch die mehrdimensionale FOURIER-Theorie über stark abfallende Funktionen mehrerer Veränderlicher exakt begründen kann:

Eine Funktion $\varphi\colon \mathbb{R}^n \to \mathbb{C}$ heißt *stark abfallend,* wenn *alle* Ausdrücke der Form

$$x_1^{e_1} \cdots x_n^{e_n} \frac{\partial^{r_1+\cdots+r_n}}{\partial x_1^{r_1} \cdots \partial x_n^{r_n}} \varphi(x_1, \ldots, x_n)$$

auf ganz $\mathbb{R}^n$ beschränkt sind. Der Vektorraum aller dieser Funktionen ist der SCHWARTZ-Raum $\mathcal{S}(\mathbb{R}^n)$.

Für Funktionen aus diesem Raum ist wieder alles relativ problemlos; zur Verallgemeinerungen auf interessantere Funktionen führt auch hier der Umweg über Distributionen $T\colon \mathcal{S}(\mathbb{R}^n) \to \mathbb{C}$, die in der naheliegenden Weise als Verallgemeinerungen eindimensionaler Distributionen definiert werden. Beispielsweise kann man dem gerade *ad hoc* betrachteten Produkt $\delta(x-a)\delta(y-b)$ über die Distribution

$$\Delta_{a,b}\colon \begin{cases} \mathbb{R}^2 \to \mathbb{C} \\ \varphi \mapsto \varphi(a,b) \end{cases}$$

einen präzisen Sinn geben – solange es in einem sinnvollen Kontext unter zwei Integralzeichen steht.

b) Fraunhofer-Beugung

Wenn Licht auf Strukturen trifft, in Vergleich zu deren Größe seine Wellenlänge nicht mehr vernachlässigbar klein ist, lassen sich die Gesetze der geometrischen Optik bekanntlich nicht mehr anwenden; man beobachtet dann Beugungsphänomene.

Beugung ist ein sehr komplexes Gebiet; für ein Beispiel im Rahmen einer Vorlesung über Höhere Mathematik müssen wir uns auf den allereinfachsten Fall beschränken. Wir gehen daher aus von einem Lichtstrahl, der aus sehr großer Entfernung kommt oder der zumindest (z.B. dank

einer Linse, aus deren Brennpunkt er kommt) so aussieht, und beobachten auch die Beugungsfigur in großer Entfernung. Diese Situation bezeichnet man als FRAUNHOFER-Beugung.

JOSEPH VON FRAUNHOFER (1787–1826) wurde in Straubing als elftes und letztes Kind eines Glasermeisters geboren; er machte auch selbst eine Lehre als Glasschleifer und Spiegelmacher. Daneben besuchte er die Feierabendschule, wo er zumindest primitive Grundkenntnisse im Rechnen erwarb. 1806 kam er an das optische Institut von UTZSCHNEIDER, der ihm Bücher über Optik und Mathematik besorgte. FRAUNHOFER entwikkelte Präzisionsmaschinen zur Fertigung optischer Instrumente von bis dahin nicht gekannter Qualität und erfand auch das optische Gitter. Durch seine Versuche zur Lichtbeugung bewies er die Wellennatur des Lichts. 1824 wurde er zum Professor ernannt; er berichtete unter anderem in öffentlichen Sonntagsvorlesungen über seine Arbeit. Im gleichen Jahr wurde er vom bayrischen König LUDWIG I. in den Adelsstand erhoben. Zwei Jahre später starb er an Tuberkulose.

Zur mathematischen Behandlung der optischen Beugung brauchen wir zunächst ein physikalisches Modell für Lichtwellen. Für eine physikalisch korrekte Beschreibung müssen wir Licht als zeitlich veränderliches elektromagnetisches Feld betrachten, d.h. wir brauchen zwei räumlich und zeitlich variable Vektorfelder $\vec{E}(x,y,z;t)$ und $\vec{B}(x,y,z;t)$, die den MAXWELLschen Gleichungen genügen. Glücklicherweise muß man in der Optik aber nur selten so weit gehen: Zwar hängt die Beugung an einem Spalt theoretisch durchaus von der Leitfähigkeit des verwendeten Materials ab, aber diese Abhängigkeit ist so gering, daß man sie für alle praktischen Zwecke vernachlässigen kann.

Man arbeitet daher in der Wellenoptik gerne mit einer sogenannten *skalaren Welle,* über deren physikalische Bedeutung man sich keine sonderlichen Gedanken macht. Aus rechnerischen Gründen betrachtet man sie als komplexwertige Funktion; falls man sich unbedingt etwas darunter vorstellen will, kann man beispielsweise den Realteil dieser Funktion als die x-Komponente des elektrischen Felds interpretieren, muß dann aber beachten, daß eine skalare Welle im Gegensatz zu einem elektrischen Feld *keine* Wechselwirkung mit Materie egal welcher Leitfähigkeit zeigt – das ist eine der Idealisierungen hinter dem Konzept

der skalaren Welle. Wichtig für uns ist nur, daß die Intensität der Welle (also z.B. die Intensität der Beugungslinien, die wir auf einem Schirm beobachten) gleich dem Betragsquadrat der Wellenfunktion sein soll.

Eine Welle hat eine räumliche wie auch zeitliche Periodizität. Zeitlich periodische Vorgänge kennen wir bereits: Das sind Schwingungen, die mathematisch durch Funktionen der Art

$$f(t) = A_0 e^{i\omega t} \quad \text{oder etwas allgemeiner} \quad f(t) = A_0 e^{i(\omega t+\varphi)}$$

beschrieben werden, wobei die Phasenverschiebung φ dafür sorgt, daß wir auch Schwingungen behandeln können, die ihre maximale Auslenkung nicht zur Zeit $t = 0$ erreichen. Wir wollen dies jedoch im folgenden ignorieren und mit der einfacheren ersten Funktion arbeiten.

Für räumlich periodische Vorgänge haben wir entsprechend zur Periode T einer Schwingung eine Wellenlänge λ; das Analogon zur Kreisfrequenz $\omega = 2\pi/T$ bezeichnen wir als

$$\textit{Wellenzahl} \quad k = \frac{2\pi}{\lambda}\,.$$

Ein eindimensionaler periodischer Vorgang kann somit beschrieben werden durch eine Funktion $g(x) = A_0 e^{ikx}$.

Im Mehrdimensionalen müssen wir die Wellenzahl k ersetzen durch einen Vektor $\vec{k}$ der Länge k, den *Wellenzahlvektor,* und betrachten die Funktion $g(\boldsymbol{x}) = A_0 e^{i\vec{k}\cdot\vec{x}}$. (Die Wellenlänge betrachten wir weiterhin nur als Skalar.)

Eine Welle soll zeitlich *und* räumlich periodisch sein; dies leistet die Funktion

$$\psi(\boldsymbol{x}, t) = A_0 e^{i(\omega t - \vec{k}\cdot\vec{x})}$$

oder natürlich auch die entsprechende Funktion mit einem Pluszeichen im Exponenten. Der Grund, warum wir das Minuszeichen bevorzugen, ist folgender:

Im eindimensionalen Fall ist $\psi(x,t)A_0 e^{i(\omega t-\vec{k}\cdot\vec{x})} = A_0 e^{ik\left(\frac{\omega}{k}t - x\right)}$, die Funktion $\psi(x,t)$ hängt also nur ab von $x - \frac{\omega}{k}t$. Dies können wir auch so interpretieren, daß

$$v = \frac{\omega}{k} = \frac{\lambda}{T} = \frac{\lambda\omega}{2\pi}$$

die Ausbreitungsgeschwindigkeit der Welle ist; denn eine Änderung der Zeit um Δt hat denselben Effekt wie eine Änderung des Orts um $v \cdot \Delta t$.

Im Falle mehrerer räumlicher Dimensionen ist alles grundsätzlich genauso, nur die Schreibweise ist etwas komplizierter: Ist $\vec{k}_0$ ein Einheitsvektor in Richtung von $\vec{k}$, d.h. $\vec{k} = k \cdot \vec{k}_0$, so ist

$$\psi(\boldsymbol{x}, t) = A_0 e^{ik\left(\frac{\omega}{k}t - \vec{k}_0\cdot\vec{x}\right)}\ ;$$

dabei ist $\vec{k}_0 \cdot \vec{x}$ die $\vec{x}$-Komponente in Richtung von $\vec{k}$. Damit ist $\vec{k}_0$ die *Richtung* des Geschwindigkeitsvektors; der Wellenzahlvektor zeigt also in Richtung der Ausbreitungsgeschwindigkeit der Welle, und der Betrag v des Geschwindigkeitsvektors ist durch obige Formel gegeben.

Die Annahme einer konstanten Amplitude A_0 in obigen Formeln ist nur in seltenen Fällen realistisch: Licht kommt meist aus einer (zumindest in erster Näherung) punktförmigen Lichtquelle, und seine Intensität nimmt mit dem Quadrat der Entfernung ab. Da die Intensität das Betragsquadrat der Wellenfunktion sein soll, müssen wir eine solche Kugelwelle also in der Form

$$\psi(\boldsymbol{x}, t) = \frac{A_0}{|\vec{x}|} e^{i(\omega t - \vec{k}\cdot\vec{x})}$$

ansetzen, sofern die Lichtquelle im Nullpunkt des Koordinatensystems sitzt.

Im Falle einer weit entfernten Lichtquelle, wie wir sie bei der FRAUNHOFER-Beugung annehmen und auch von der Sonne her kennen, spielt allerdings die Ortsabhängigkeit der Amplitude praktisch keine Rolle, so daß wir keinen nennenswerten Fehler machen, wenn wir sie als konstant annehmen. In diesem Fall sprechen wir von einer *ebenen* Welle.

Ausgangspunkt für die Berechnung von Beugungsbilder ist das HUYGENSsche Prinzip: Jeder Punkt des Hindernisses ist Quelle einer Kugelwelle, deren Amplitude gleich der Amplitude der einfallenden Welle mal der Durchlässigkeitsfunktion α des Hindernisses im betrachteten Punkt ist. Letztere gibt an, welcher Teil des Lichts durchgelassen wird; sie ist also eins an den Stellen, an denen alles Licht durchkommt, und null dort, wo nichts durchkommt. An Stellen, an denen ein Teil des Lichts durchgelassen wird, kann sie auch Zwischenwerte annehmen.

CHRISTIAAN HUYGENS (1629–1695) kam aus einer niederländischen Diplomatenfamilie. Dadurch und später auch durch seine Arbeit hatte er Kontakte zu führenden europäischen Wissenschaftlern wie DESCARTES und PASCAL. Nach seinem Studium der Mathematik und Jura arbeitete er teilweise auch selbst als Diplomat, interessierte sich aber bald vor allem für Astronomie und den Bau der dazu notwendigen Instrumente. Er entwickelte eine neue Methode zum Schleifen von Linsen und erhielt ein Patent für die erste Pendeluhr. Trotz des französisch-niederländischen Kriegs arbeitete er einen großen Teil seines Lebens an der *Académie Royale des Sciences* in Paris, wo beispielsweise LEIBNIZ viel Mathematik bei ihm lernte. HUYGENS war ein scharfer Kritiker sowohl von NEWTONS Theorie des Lichts als auch seiner Gravitationstheorie, die er für absurd und nutzlos hielt. Gegen Ende seines Lebens beschäftigte er sich mit der Möglichkeit außerirdischen Lebens.

Bei der FRAUNHOFER-Beugung betrachten wir auch die gebeugten Wellen nur aus sehr großer Entfernung und können daher statt von Kugelwellen von ebenen Wellen ausgehen. Außerdem können wir die Zeitabhängigkeit der Welle ignorieren, denn die Frequenzen, mit denen das sichtbare Licht schwingt, liegen um Größenordnungen jenseits sowohl unserer Reaktionszeit als auch der unserer Meßinstrumente, so daß wir nur die Amplituden messen können. Schreiben wir die einfallende Welle als

$$\psi(\boldsymbol{x}, t) = A_0 e^{-i\vec{k}\vec{x}} \cdot e^{i\omega t} ,$$

so ist der zweite Faktor eine zeitabhängige Phasenvariation, die wir bei der Berechnung des räumlichen Intensitätsverteilung des Beugungsbilds ignorieren können. Es reicht also, den Faktor $A_0 e^{-i\vec{k}\vec{x}}$ zu betrachten.

Eine weitere Konsequenz des (auch im Vergleich zur Größe des Hindernisses) weit entfernten Betrachtungspunkts ist, daß wir das Hindernis vom Schirm aus praktisch nur als Punkt sehen; was an einer gegebenen Stelle des Schirms ankommt, hängt also im wesentlichen nur ab vom Winkel θ oder (im Zweidimensionalen) den Winkeln θ und φ, unter dem (oder denen) die Strahlen von diesem „Punkt“ ausgehen.

Um die Intensität des Beugungsbilds in einem gegebenen Punkt zu berechnen, müssen wir also alle vom Hindernis in einem festen Winkel

ausgehenden Strahlen aufsummieren, und *hierbei* müssen wir auch die Phasen berücksichtigen, da diese Strahlen miteinander interferieren. Abbildung 28 zeigt, wie sich die Laufwege zweier benachbarter Strahlen unterscheiden, und diese Differenzen können wir nicht vernachlässigen, da sie in der Größenordnung des Hindernisses und damit auch der Wellenlänge des Lichts liegen.

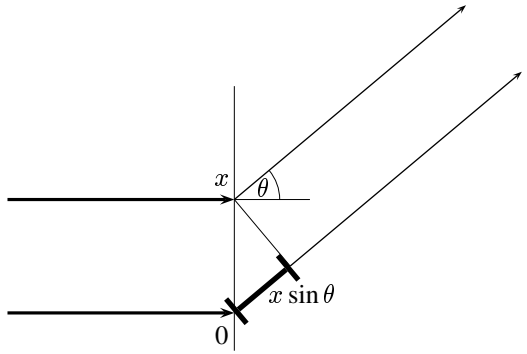

Abb. 28: Laufwegunterschied zweier paralleler Strahlen

Betrachten wir zunächst den (in Abbildung 28 dargestellten) eindimensionalen Fall. Verglichen mit dem Strahl, der von einem (irgendwie gewählten) Nullpunkt des Hindernisses ausgeht, hat der Strahl mit Ausgangspunkt in Entfernung x einen Laufwegunterschied von $x \sin\theta$; dies entspricht einem Phasenfaktor von $e^{-ikx\sin\theta}$. Wählen wir also die Phase im Nullpunkt als Referenz (die wir in den zu ignorierenden Phasenfaktor der einfallenden Welle hineinziehen können), ist die Summe aller unter dem Winkel θ abgehenden Strahlen gleich

$$\int_{-\infty}^{\infty} \alpha(x) e^{-ikx\sin\theta}\, dx \,;$$

das ist gleich der FOURIER-Transformierten von $\alpha(x)$, ausgewertet im Punkt $u = k \sin\theta$.

Bei einem zweidimensionalen Hindernis müssen entsprechend zwei Winkelvariablen θ und ϕ berücksichtigt werden, und auch die Durchlässigkeitsfunktion α hängt von zwei Variablen x, y ab; außerdem müssen wir nun vom Wellenzahlvektor sowohl die x- als auch die y-Komponente berücksichtigen. Wir erhalten daher als Summe aller Strahlen unter den beiden gegeben Winkeln das Integral

$$\iint\limits_{\mathbb{R}^2} \alpha(x,y) e^{-i(k_1 x \sin\theta + k_2 y \sin\phi)}\, dx\, dy\,,$$

d.h. die zweidimensionale FOURIER-Transformierte von α, ausgewertet im Punkt $(u,\ v) = (k_1 \sin\theta,\ k_2 \sin\varphi)$.

Zur Vereinfachung der Schreibweise drückt man das Beugungsbild meist einfach in der Variablen u *bzw.* den Variablen u und v aus statt in den Winkelvariablen; dann ist das Beugungsbild eines Hindernisses mit Durchlässigkeitsfunktion α einfach die FOURIER-Transformierte von α.

Die Größen u und v lassen sich zwar als Strecken interpretieren, sind aber *nicht* proportional zu den Strecken, die man auf einem ebenen Schirm messen kann: Deren Längen sind proportional zu $\tan\theta$ und $\tan\varphi$. Für kleine Winkel, auf die man sich bei der FRAUNHOFER-Beugung wegen des großen Abstands zum Schirm notwendigerweise beschränken muß, unterscheiden sich allerdings Sinus, Tangens und Bogenmaß nur wenig, so daß man auch ohne Umrechnung ein gutes Bild des Beugungsmusters erhält.

Als erstes Beispiel wollen wir das Beugungsbild eines eindimensionalen Spalts berechnen. Dieser habe die Breite a; seine Durchlässigkeitsfunktion kann also beispielsweise geschrieben werden als

$$\alpha\colon \begin{cases} \mathbb{R} \to \mathbb{R} \\ x \mapsto \begin{cases} 1 & \text{für } -\frac{a}{2} \le x \le \frac{a}{2} \\ 0 & \text{sonst} \end{cases} \end{cases}$$

mit
$$\widehat{\alpha}(u) = \int\limits_{-\infty}^{\infty} \alpha(x) e^{-uix}\, dx = \int\limits_{-\frac{a}{2}}^{\frac{a}{2}} e^{-iux} dx$$

$$= \frac{e^{-\frac{iau}{2}} - e^{\frac{iau}{2}}}{-iu} = \frac{2\sin\frac{au}{2}}{u} = a\frac{\sin\frac{au}{2}}{\frac{au}{2}} = a \operatorname{sinc}\frac{au}{2}\,.$$

Dies erklärt, warum die Funktion $\operatorname{sinc} x$ auch als *Spaltfunktion* bezeichnet wird.

Die Lichtintensitäten, die man im Beugungsbild beobachtet, sind allerdings *nicht* durch diese Funktion gegeben: $\widehat{\alpha}(u)$ ist die Amplitude einer skalaren Welle; die Intensität ist gleich dem Betragsquadrat davon, bei einer reellen Funktion wie hier also einfach das Quadrat

$$\widehat{\alpha}(u)^2 = 4\frac{\sin^2\frac{au}{2}}{u^2} = a^2 \operatorname{sinc}^2\frac{au}{2}\,.$$

Als nächstes Beispiel betrachten wir Beugung an einem regelmäßigen Strichgitter. Der Abstand zweier Striche sei d und es gebe insgesamt $2N+1$ Striche. Wenn wir in erster Näherung die Breite der Striche vernachlässigen, können wir die Durchlässigkeitsfunktion α als Summe von δ-Distributionen schreiben:

$$\alpha(t) = \sum_{k=-N}^{N} \delta(x - kd)\,.$$

Das Beugungsbild ist somit gegeben durch

$$\begin{aligned}
\widehat{\alpha}(u) &= \int\limits_{-\infty}^{\infty} \alpha(x) e^{-iux}\, dx = \sum_{k=-N}^{N} \int\limits_{-\infty}^{\infty} \delta(x-kd) e^{-iux}\, dx \\
&= \sum_{k=-N}^{N} e^{-iudk} = \sum_{k=-N}^{N} e^{iukd} = e^{-iuNd} \sum_{k=0}^{2N} e^{iukd} \\
&= e^{-iuNd} \frac{1 - e^{iu(2N+1)d}}{1 - e^{iud}} = \frac{e^{-iuNd} - e^{iu(N+1)d}}{1 - e^{iud}} \\
&= \frac{e^{iu(N+\frac{1}{2})d} - e^{-iu(N+\frac{1}{2})d}}{e^{iu\frac{d}{2}} - e^{-iu\frac{d}{2}}} = \frac{\sin u\left(N + \frac{1}{2}\right) d}{\sin\frac{ud}{2}}\,.
\end{aligned}$$

Abbildung 29 zeigt diese Funktion; man sieht ihr das charakteristische Linienmuster an, das man bei der Beugung am Gitter beobachtet.

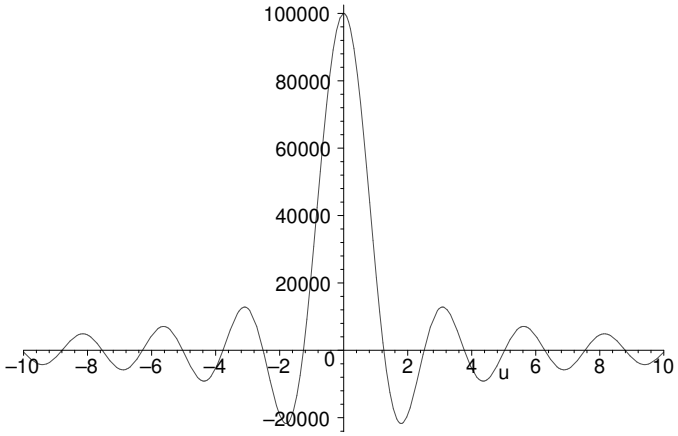

Abb. 29: Beugung am Gitter

Trotzdem wird vielleicht einigen Lesern unwohl sein beim Gedanken an eine Summe von δ-Distributionen als Durchlässigkeitsfunktion. Deshalb wollen wir zur Sicherheit nachrechnen, was sich ändert, wenn wir stattdessen die Striche als Spalte der Breite a annehmen.

Für einen einzelnen solchen Spalt haben wir dann die oben betrachtete Durchlässigkeitsfunktion

$$\alpha_a\colon \begin{cases} \mathbb{R} \to \mathbb{R} \\ x \mapsto \begin{cases} 1 & \text{für } -\frac{a}{2} \le x \le \frac{a}{2} \\ 0 & \text{sonst} \end{cases} \end{cases}$$

eines Spalts der Breite a, und die Durchlässigkeitsfunktion des gesamten Strichgitters ist die Faltung $\alpha \star \alpha_a$ dieser Funktion mit der oben betrachteten Funktion α. Das Beugungsbild ist somit gegeben durch das Produkt des gerade berechneten Beugungsbilds mit dem Beugungsbild eines Spalts, also durch

$$\frac{\sin u\left(N + \frac{1}{2}\right)d}{\sin \frac{ud}{2}} \cdot \frac{\sin \frac{au}{2}}{\frac{u}{2}} \,.$$

Da der Abstand zwischen zwei Spaltmitten gleich d ist, muß die Spaltbreite a echt kleiner als d sein, und die Anzahl N der Striche im Gitter

liegt typischerweise bei mindestens einigen Zehntausend. Somit hat der Sinus im zweiten Term eine Kreisfrequenz, die um einen mindestens fünfstelligen Faktor größer ist als die im ersten; der zweite Term zeigt also erst dann eine nennenswerte Variation, wenn wir im ersten Faktor mehrere Tausend Linien betrachten. In dem Bereich, den wir realistischerweise beobachten können, ist der zweite Term daher für alle praktischen Zwecke konstant. Der Betrag dieser Konstanten ist irrelevant, denn da wir bei der FRAUNHOFER-Beugung das Beugungsbild in „sehr großer“ Entfernung vom Gitter betrachten, können wir sinnvollerweise ohnehin nur von relativen, nicht aber von absoluten Helligkeiten reden.

Als letztes Beispiel zur eindimensionalen Beugung möchte ich eines betrachten, bei dem das Licht nicht als konstante Wellenfront einfällt: Je nach Wahl der Randbedingungen im optischen Resonator entsteht nicht immer ein Strahl, der näherungsweise als ebene Welle betrachtet werden kann (die sogenannte TEM_{00}-Mode); bringt man Hindernisse in den Strahlengang, können auch höhere TEM-Moden angeregt werden (TEM = *transversal elektromagnetisch*). Bei einem dünnen Hindernis wie etwa einem Haar genau in der Mitte des Strahls beispielsweise entsteht die TEM_{01}-Mode, die aus einem linken und einem rechten Halbstrahl besteht, deren Phasen sich um 180° unterscheiden, und die man ansonsten wieder näherungsweise als ebene Wellen betrachten kann. Trifft ein solcher Strahl auf einen Spalt, dessen Mitte mit der Grenze zwischen den beiden Halbstrahlen zusammenfällt, ist also in der linken Hälfte des Spalts die Phase um 180° gegenüber der rechten verschoben; dies können wir formal dadurch beschreiben, daß wir die Durchlässigkeitsfunktion des Spalts multiplizieren mit einer Funktion, die in der linken Hälfte +1 und in der rechten gleich -1 ist. Für einen Spalt der Breite a erhalten wir als Beugungsbild

$$\int_{-\frac{a}{2}}^{0} e^{-iux}\,dx + \int_{0}^{\frac{a}{2}} -e^{-iux}\,dx = \frac{1 - e^{iua/2} - e^{-iua/2} + 1}{-iu}$$

$$= \frac{2i}{u}\left(1 - \cos\frac{ua}{2}\right) \,.$$

Alternativ läßt sich dies auch über die Beziehung

$$e^{iua/2} + e^{-iua/2} - 2 = (e^{iua/4} - e^{-iua/4})^2 = -4\sin^2\frac{ua}{4}$$

als

$$\frac{4i}{u}\sin^2\frac{ua}{4}$$

schreiben. Daß hier imaginäre Größen auftreten, braucht uns natürlich nicht zu stören: Die beobachteten Intensitäten sind bekanntlich die Betragsquadrate der hier berechneten Funktionen, also reell und positiv.

Zum Abschluß möchte ich noch zumindest ein Beispiel eines zweidimensionalen Beugungsbilds betrachten. Leider sind die zugehörigen FOURIER-Integrale schon in so einfachen Fällen wie dem einer scheibenförmigen Blende nicht mehr elementar auswertbar; wir beschränken uns daher auf den extrem einfachen Fall einer rechteckigen Blende. Deren Durchlässigkeitsfunktion ist

$$\alpha\colon \begin{cases} \mathbb{R}^2 \to \mathbb{R} \\ (x,y) \mapsto \begin{cases} 1 & \text{falls } -\frac{a}{2} \le x \le \frac{a}{2} \text{ und } -\frac{b}{2} \le y \le \frac{b}{2} \\ 0 & \text{sonst} \end{cases} \end{cases},$$

die Beugungsfigur ist also gegeben durch

$$\begin{aligned}
\widehat{\alpha}(u,v) &= \iint\limits_{\mathbb{R}^2} \alpha(x,y)e^{-i(ux+vy)}\,dx\,dy = \iint\limits_{\substack{-\frac{a}{2}\le x\le\frac{a}{2}\\ -\frac{b}{2}\le y\le\frac{b}{2}}} e^{-i(ux+vy)}\,dx\,dy \\
&= \int\limits_{-\frac{a}{2}}^{\frac{a}{2}}\int\limits_{-\frac{b}{2}}^{\frac{b}{2}} e^{-iux}e^{-ivy}\,dx\,dy = \int\limits_{-\frac{a}{2}}^{\frac{a}{2}} e^{-iux}\left(\int\limits_{-\frac{b}{2}}^{\frac{b}{2}} e^{-ivy}\,dy\right)dx \\
&= \int\limits_{-\frac{a}{2}}^{\frac{a}{2}} e^{-iux}\cdot\frac{\sin\frac{bv}{2}}{v}\,dx = 4\cdot\frac{\sin\frac{au}{2}}{u}\cdot\frac{\sin\frac{bv}{2}}{v},
\end{aligned}$$

da wir das Rechteck als Normalbereich betrachten können und somit das zweidimensionale Integral über zwei eindimensionale Integrationen berechnen können.

Als Beugungsfigur erhalten wir, nicht gerade überraschenderweise, das Produkt einer vertikalen und einer horizontalen Beugungsfigur eines eindimensionalen Spalts.

Kapitel 4
Differentialgleichungen

Differentialgleichungssysteme sind so ziemlich *das* wichtigste mathematischen Hilfsmittel der Naturwissenschaften und der Technik. Die dahinterstehende Grundidee ist einfach: Man kann zwar nur selten *a priori* sagen, wie sich ein System über einen längeren Zeitraum hinweg entwickeln wird, aber man hat oft aufgrund von Naturgesetzen eine klare Vorstellung über die Zustandsänderung *im nächsten Augenblick,* d.h. also über den Wert der zeitlichen Ableitung der Zustandsgrößen in Abhängigkeit vom gegenwärtigen Zustand des Systems.

§1: Definitionen und erste Beispiele

a) Wurfparabel

Ein einfaches Beispiel hierfür liefert das zweite NEWTONsche Gesetz, wonach die zeitliche Ableitung des Impuls eines Teilchens gleich der auf das Teilchen wirkenden Kraft ist.

Ein in die Luft geworfener Gegenstand bewegt sich unter gewissen Bedingungen näherungsweise auf einer parabelförmigen Bahn. Wir wollen diese etwas vage Aussage präzisieren und mathematisch herleiten.

Es gibt viele Wurftechniken, und nur wenige davon können auf einfache Weise durch ein mathematisches Modell beschrieben werden; wir ignorieren daher den genauen Vorgang des Abwurfs und gehen davon aus, daß der Gegenstand *irgendwie* eine Anfangsgeschwindigkeit $\vec{v}$ erreicht hat im Abwurfpunkt mit Koordinaten (x_0, y_0, z_0); den Zeitpunkt des Abwurfs bezeichnen wir mit t_0.

Als nächstes nehmen wir an, daß wir den Luftwiderstand vernachlässigen können, eine Annahme, die beim Kugelstoßen kaum zu Fehlern führt, die aber beispielsweise für einen Fallschirmspringer (auch mit geschlossenem Fallschirm) oder einen Papierflieger völlig unrealistisch ist. Als nächstes wollen wir auch noch annehmen, daß wir nur relativ geringe Wurfhöhen erreichen, so daß die Erdanziehung als konstant angenommen werden kann.

Die Bewegung des Gegenstandes wird dann durch zwei Naturgesetze bestimmt: Das Gravitationsgesetz beschreibt den Effekt der Erdanziehung, und das zweite NEWTONsche Gesetz sagt uns, wie sich diese Kraft auf die Bewegung des Gegenstands auswirkt. Die Gravitation können wir aufgrund der gemachten Annahmen als konstant annehmen, d.h. auf einen Körper der Masse m wirkt die Kraft gm, wobei $g \approx 9{,}8\, m/s^2$ die Gravitationsbeschleunigung an der Erdoberfläche ist; bei „üblicher" Ausrichtung des Koordinatensystems wirkt sie in Richtung der negativen z-Achse. Diese Gravitationskraft ist nach dem zweiten NEWTONschen Gesetz gleich der Ableitung des Impulses nach der Zeit; wenn wir die Masse m als konstant voraussetzen, ist das also gleich m mal der Ableitung der Geschwindigkeit oder m mal der zweiten Ableitung des Orts. Wir haben somit das Differentialgleichungssystem

$$\ddot{x}(t) = 0, \quad \ddot{y}(t) = 0 \quad \text{und} \quad \ddot{z}(t) = -g\,.$$

Diese Gleichungen sind erfüllt, wann immer $x(t)$ und $y(t)$ lineare Funktionen von t sind und $z(t)$ eine quadratische Funktion mit führendem Koeffizienten $-g$. Die sechs noch fehlenden Koeffizienten dieser drei Polynomfunktionen geben uns die Anfangsbedingungen: Zum Zeitpunkt $t = t_0$ ist

$$x(t_0) = x_0, \quad y(t_0) = y_0 \quad \text{und} \quad z(t_0) = z_0\,,$$

und die Geschwindigkeit ist $\vec{v}$, d.h.

$$\dot{x}(t_0) = v_1, \quad \dot{y}(t_0) = v_2 \quad \text{und} \quad \dot{z}(t_0) = v_3\,.$$

Also ist

$$x(t) = v_1(t - t_0) + x_0, \quad y(t) = v_2(t - t_0) + y_0$$

und

$$z(t) = -g(t - t_0)^2 + v_3(t - t_0) + z_0\,.$$

Diese Gleichungen beschreiben in der Tat fast immer eine Parabel: Falls wir die x-Achse des Koordinatensystems so wählen, daß die Anfangsgeschwindigkeit $\vec{v}$ in der (x, z)-Ebene liegt, ist $v_2 = 0$. Falls auch v_1 verschwindet, falls wir den Gegenstand also senkrecht nach oben (oder gar unten) werfen, sind $x(t) = x_0$ und $y(t) = y_0$ konstant und nur

$$z(t) = -g(t - t_0)^2 + v_3(t - t_0) + z_0$$

hängt von der Zeit ab. Andernfalls können wir durch v_1 dividieren; wir erhalten

$$t - t_0 = \frac{x(t) - x_0}{v_1} \quad \text{und} \quad z(t) = \frac{-g}{v_1}\big(x(t) - x_0\big)^2 + \frac{v_3}{v_1}\big(x(t) - x_0\big) + z_0\,,$$

die Punkte $\big(x(t), z(t)\big)$ liegen also in der Tat auf einer Parabel.

b) Radioaktiver Zerfall

Das gerade durchgerechnete Beispiel war insofern untypisch für Differentialgleichungen, als auf den rechten Seite der Gleichungen nur Konstanten standen; üblicherweise wird man dort Funktionen erwarten, die nicht nur von t abhängen (so daß man sie einfach integrieren kann), sondern auch noch von den gesuchten Funktionen. Beim radioaktiven Zerfall etwa ist die pro (kleiner) Zeiteinheit zerfallende Masse proportional zur noch vorhandenen Masse, es gibt also eine Konstante $\lambda > 0$, die sogenannte Zerfallskonstante, so daß die zum Zeitpunkt t vorhandene Masse $m(t)$ der Gleichung

$$\dot{m}(t) = -\lambda m(t)$$

genügt – zumindest, wenn diese Masse hinreichend groß ist. (Im atomaren Bereich muß man auch statistische Effekte berücksichtigen, aber ab etwa 10^{10} Atomen können die für alle praktischen Fälle vernachlässigt werden.)

Wir kennen bereits eine Funktion, die sich so verhält, wie es die obige Differentialgleichung angibt, nämlich die Exponentialfunktion $e^{-\lambda t}$, und natürlich entspricht auch für jedes konstante Vielfache dieser Funktion die Differentiation einfach der Multiplikation mit $-\lambda$. Das sind dann aber bereits alle Funktionen mit dieser Eigenschaft, denn der Quotient

$$q(t) = \frac{m(t)}{e^{-\lambda t}} = m(t) \cdot e^{\lambda t}$$

einer Lösungsfunktion und der Funktion $e^{-\lambda t}$ hat die Ableitung

$$\dot{q}(t) = \dot{m}(t) \cdot e^{\lambda t} + m(t) \cdot \lambda e^{\lambda t} = -\lambda m(t) \cdot e^{\lambda t} + \lambda m(t) \cdot e^{\lambda t} = 0\,,$$

ist also gleich einer Konstanten c, so daß

$$m(t) = c \cdot e^{-\lambda t}$$

ist. Indem wir $t = 0$ setzen, sehen wir, daß die Konstante $c = m(0)$ gleich der zum Zeitpunkt 0 vorhandenen Masse ist; falls wir stattdessen die Masse $m_0 = m(t_0)$ zu einem anderen Zeitpunkt t_0 kennen, können wir analog zum obigen Beispiel auch schreiben

$$m(t) = m_0 e^{-\lambda(t - t_0)} = (m_0 e^{\lambda t_0}) \cdot e^{-\lambda t}\,.$$

c) Differentialgleichungen und Differentialgleichungssysteme

Wir betrachten ein System, das durch n zeitlich veränderliche Größen $y_1(t), \ldots, y_n(t)$ beschrieben wird; unter einem System von Differentialgleichungen oder kurz einer Differentialgleichung verstehen wir eine Vorschrift, die die zeitlichen Ableitungen $\dot{y}_1(t), \ldots, \dot{y}_n(t)$ aus den Funktionswerten berechnet:

$$\begin{aligned}
\dot{y}_1(t) &= f_1\big(t, y_1(t), \ldots, y_n(t)\big)\\
\dot{y}_2(t) &= f_2\big(t, y_1(t), \ldots, y_n(t)\big)\\
&\;\;\vdots \quad \vdots\\
\dot{y}_n(t) &= f_n\big(t, y_1(t), \ldots, y_n(t)\big)\,.
\end{aligned}$$

Falls die Funktionen f_i nur von $y_1(t), \ldots, y_n(t)$ abhängen und nicht auch noch direkt von der Zeit t spricht man von einem *autonomen* System. Da Naturgesetze nicht von der Zeit abhängen, hat man es in naturwissenschaftlich-technischen Anwendungen meist mit autonomen Systemen zu tun; man kann allerdings auch den Einfluß von Umgebungsgrößen in einem zeitabhängigen Term zusammenfassen und so ein nichtautonomes System erhalten.

Falls wir, wie in den Beispiel aus den vorangegangenen Abschnitten, die Werte der beteiligten Funktionen zu einem festen Zeitpunkt $t = t_0$ kennen, reden wir von einem *Anfangswertproblem.* Solche Probleme treten

typischerweise dann auf, wenn das weitere Verhalten *eines* konkreten Systems vorhergesagt werden soll.

Auch Differentialgleichungen, in denen wie im Beispiel der Wurfparabel höhere Ableitungen vorkommen, lassen sich so interpretieren: Wenn wir dort die drei Komponenten des Geschwindigkeitsvektors als neue Funktionen

$$u(t) = \dot{x}(t), \quad v(t) = \dot{y}(t) \quad \text{und} \quad w(t) = \dot{z}(t)$$

einführen, können wir das System schreiben als

$$\begin{aligned} \dot{x}(t) &= u(t), & \dot{y}(t) &= v(t), & \dot{z}(t) &= w(t), \\ \dot{u}(t) &= 0, & \dot{v}(t) &= 0, & \dot{w}(t) &= -g, \end{aligned}$$

und wir kennen für jede der sechs beteiligten Funktionen ihren Wert an der Stelle $t = t_0$.

Ein für die Informationstechnik wichtiger Spezialfall sind Gleichungen der Form

$$y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \cdots + a_1\dot{y}(t) + a_0y(t) = b(t)\,,$$

die sogenannten linearen Differentialgleichungen n-ter Ordnung mit konstanten Koeffizienten. Auch diese Gleichungen lassen sich leicht auf die obige Form bringen: Wir betrachten n neue Funktionen

$$y_0(t), y_1(t), \ldots, y_{n-1}(t)$$

mit der Idee, daß sich $y_i(t)$ so verhalten soll wie die i-te Ableitung von $y(t)$. Dazu bilden wir das Differentialgleichungssystem

$$\begin{aligned} \dot{y}_0(t) &= y_1(t) \\ \dot{y}_1(t) &= y_2(t) \\ &\vdots \quad \vdots \\ \dot{y}_{n-2}(t) &= y_{n-1}(t) \\ \dot{y}_{n-1}(t) &= b(t) - a_{n-1}y^{(n-1)}(t) - \cdots - a_1\dot{y}(t) - a_0y(t)\,. \end{aligned}$$

Für jede Lösung $y(t)$ der obigen Gleichung ist dann das n-tupel

$$\big(y(t), \dot{y}(t), \ddot{y}(t), \ldots, y^{(n-1)}(t)\big)$$

eine Lösung des Systems, und für jede Lösung

$$\big(y_0(t), y_1(t), y_2(t), \ldots, y_{n-1}(t)\big)$$

des Differentialgleichungssystems ist $y_0(t)$ eine Lösung der obigen Gleichung.

Auch wenn das Differentialgleichungssystem als Anfangswertproblem gegeben ist, läßt sich das leicht in Anfangswerte für die Gleichung höherer Ordnung umschreiben: Hier werden die Werte $y(t_0)$, $\dot{y}(t_0)$ *usw.* bis $y^{(n-1)}(t_0)$ vorgegeben.

Somit beschreiben das System von Differentialgleichungen erster Ordnung und die eine Differentialgleichung höherer Ordnung genau dasselbe Phänomen. Wie wir im vorigen Kapitel gesehen haben, läßt sich die Differentialgleichung höherer Ordnung recht gut mit Hilfe von LAPLACE-Transformationen lösen; in diesem Kapitel werden wir sehen, daß der im letzten Semester entwickelte (und im folgenden noch auszubauende) Apparat der linearen Algebra eine strukturelle Übersicht über die Lösungsmenge des Systems. Erst die Kombination beider Ansätze liefert ein vollständiges Bild.

d) Systeme linearer Differentialgleichungen

Wir betrachten in diesem Abschnitt Systeme von Differentialgleichungen, wie sie zu Beginn dieses Paragraphen definiert wurden, unter der (sehr) einschränkenden Voraussetzungen, daß die rechten Seiten linear in den gesuchten Funktionen $y_1(t), \ldots, y_n(t)$ sind; wir betrachten also ein System

$$\begin{aligned} \dot{y}_1(t) &= a_{11}(t)y_1(t) + a_{12}(t)y_2(t) + \cdots + a_{1n}(t)y_n(t) + b_1(t) \\ \dot{y}_2(t) &= a_{21}(t)y_1(t) + a_{22}(t)y_2(t) + \cdots + a_{2n}(t)y_n(t) + b_2(t) \\ &\vdots \qquad \vdots \\ \dot{y}_n(t) &= a_{n1}(t)y_1(t) + a_{n2}(t)y_2(t) + \cdots + a_{nn}(t)y_n(t) + b_n(t) \end{aligned} \quad . \qquad (*)$$

Eventuell haben wir noch Anfangsbedingungen der Form

$$y_1(t_0) = c_0, \quad y_2(t_0) = c_2, \quad \ldots, \quad y_n(t_0) = c_n \qquad (**)$$

für ein festes $t_0 \in \mathbb{R}$.

Für das im letzten Kapitel betrachtete Beispiel des elektrischen Schwingkreises mit angelegter Wechselspannung etwa haben wir bei dieser Sicht der Dinge die beiden Funktionen $Q(t)$, die Ladung des Kondensators zum Zeitpunkt t, und $I(t) = \dot{Q}(t)$, die resultierende Stromstärke; das Differentialgleichungssystem $(*)$ ist hier also

$$\begin{aligned} \dot{Q}(t) &= I(t) \\ \dot{I}(t) &= -\frac{R}{L} I(t) - \frac{Q(t)}{LC} + A_0 \cos \omega_0 t \end{aligned} .$$

Wir können die Funktionen $y_i(t)$, $b_i(t)$ und die Anfangswerte c_i jeweils zu Vektoren zusammenfassen und die Koeffizientenfunktionen $a_{ij}(t)$ zu einer Matrix: Mit

$$\vec{y}(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{pmatrix}, \qquad \vec{b}(t) = \begin{pmatrix} b_1(t) \\ b_2(t) \\ \vdots \\ b_n(t) \end{pmatrix}, \qquad \vec{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$$

und

$$A(t) = \begin{pmatrix} a_{11}(t) & a_{12}(t) & \dots & a_{1n}(t) \\ a_{21}(t) & a_{22}(t) & \dots & a_{2n}(t) \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}(t) & a_{n2}(t) & \dots & a_{nn}(t) \end{pmatrix},$$

erhalten wir die übersichtlichere Form

$$\dot{\vec{y}}(t) = A(t) \cdot \vec{y}(t) + \vec{b}(t),$$

wobei die Ableitung eines Vektors von Funktionen natürlich der Vektor der abgeleiteten Funktionen sein soll. Falls es Anfangsbedingungen gibt, können sie nun in der kompakte Form $\vec{y}(t_0) = \vec{c}$ geschrieben werden.

In Analogie zu den linearen Gleichungssystemen bezeichnen wir das System $(*)$ als *homogen,* wenn $\vec{b}(t)$ der Nullvektor ist, wenn also alle Funktionen $b_i(t)$ verschwinden; andernfalls bezeichnen wir es als *inhomogen.* Das homogene System zu einem gegebenen inhomogenen System soll einfach dasjenige System sein, in dem alle $b_i(t)$ durch null ersetzt wurden.

Analog zum Fall linearer Gleichungssystemen gilt auch hier

Lemma: *a)* Die Menge aller Lösungen eines *homogenen* Differentialgleichungssystems der Form $(*)$ ist ein $\mathbb{R}$-Vektorraum.
b) Ist das System nicht homogen und ist $\vec{y}(t)$ eine feste Lösung, so läßt sich jede andere Lösung $\vec{z}(t)$ schreiben als $\vec{z}(t) = \vec{y}(t) + \vec{x}(t)$ mit einer Lösung $\vec{x}(t)$ des zugehörigen homogenen Systems; die Lösungsmenge ist also ein affiner Raum.

Beweis: a) Wir müssen zeigen, daß für zwei Lösungen $\dot{\vec{x}}(t)$ und $\dot{\vec{y}}(t)$ eines homogenen Systems auch jede Linearkombination $\lambda\dot{\vec{x}}(t) + \mu\dot{\vec{y}}(t)$ mit $\lambda, \mu \in \mathbb{R}$ wieder eine Lösung ist. Das ist aber klar, denn wenn

$$\dot{\vec{x}}(t) = A(t) \cdot \vec{x}(t) \quad \text{und} \quad \dot{\vec{y}}(t) = A(t) \cdot \vec{y}(t)$$

ist, gilt auch

$$\begin{aligned} \frac{d}{dt}\big(\lambda\vec{x}(t) + \mu\vec{y}(t)\big) &= \lambda\dot{\vec{x}}(t) + \mu\dot{\vec{y}}(t) = \lambda A(t) \cdot \vec{x}(t) + \mu A(t) \cdot \vec{y}(t) \\ &= A(t) \cdot \big(\lambda\vec{x}(t) + \mu\vec{y}(t)\big) . \end{aligned}$$

b) Sind $\vec{y}(t)$ und $\vec{z}(t)$ zwei Lösungen von $(*)$, so ist

$$\dot{\vec{y}}(t) = A(t) \cdot \vec{y}(t) + \vec{b}(t) \quad \text{und} \quad \dot{\vec{z}}(t) = A(t) \cdot \vec{z}(t) + \vec{b}(t) ;$$

die Differenz $x(t) = z(t) - y(t)$ hat somit die Ableitung

$$\begin{aligned} \dot{\vec{x}}(t) = \dot{\vec{z}}(t) - \dot{\vec{y}}(t) &= \Big(A(t) \cdot \vec{z}(t) + \vec{b}(t)\Big) - \Big(A(t) \cdot \vec{z}(t) + \vec{b}(t)\Big) \\ &= A(t) \cdot \vec{z}(t) - A(t) \cdot y(t) = A(t) \cdot \Big(\vec{z}(t) - \vec{y}(t)\Big) = A(t) \cdot \vec{x}(t) \end{aligned}$$

und $x(t)$ löst also in der Tat das zugehörige homogene System. ■

Um die Lösungsmenge des Differentialgleichungssystems $(*)$ zu verstehen, müssen wir nach diesem Lemma zwei Teilaufgaben lösen:

1.) Wir müssen den Vektorraum der Lösungen des homogenen Systems bestimmen.
2.) Wir müssen uns wenigstens eine Lösung des inhomogenen Systems verschaffen – oder zumindest wissen, daß eine existiert.

Um im einfachsten Fall zu sehen, wie so etwas funktionieren könnte, betrachten wir ein „System“ aus genau einer Gleichung

$$\dot{y}(t) = a(t) \cdot y(t) + b(t) ;$$

dabei nehmen wir an, daß y eine differenzierbare Funktion von $\mathbb{R}$ nach $\mathbb{R}$ sei und $a, b\colon \mathbb{R} \to \mathbb{R}$ stetige Funktionen.

Wir beginnen mit der Lösung des homogenen Systems

$$\dot{y}(t) = a(t) \cdot y(t)\,.$$

Unter der Annahme, daß wir das dürfen, dividieren wir durch $y(t)$ und erhalten

$$\frac{\dot{y}(t)}{y(t)} = a(t)\,.$$

Der Quotient links ist bekanntlich die logarithmische Ableitung von $y(t)$; falls dies nicht mehr bekannt sein sollte, zeigt eine einfache Anwendung der Kettenregel, daß in einem Intervall, in dem $y(t)$ positiv ist,

$$\frac{d}{dt} \ln y(t) = \frac{\dot{y}(t)}{y(t)} = a(t)$$

ist. In einem Intervall, in dem $y(t)$ negativ ist, gilt entsprechend

$$\frac{d}{dt} \ln\bigl(-y(t)\bigr) = \frac{-\dot{y}(t)}{-y(t)} = \frac{\dot{y}(t)}{y(t)} = a(t)\,,$$

und allgemein haben wir somit

$$\frac{d}{dt} \ln |y(t)| = \frac{\dot{y}(t)}{y(t)} = a(t)$$

in jedem Intervall, in dem $y(t)$ nirgends verschwindet.

Integration beider Seiten führt auf

$$\ln |y(t)| = \int a(t)\,dt + C \quad \text{oder} \quad y(t) = e^{\int a(t)\,dt + C} = e^{C} \cdot e^{\int a(t)\,dt}$$

oder

$$y(t) = \pm e^{C} \cdot e^{\int a(t)\,dt}\,,$$

wobei das Vorzeichen wegen der Stetigkeit von y im gesamten Intervall konstant ist, da die Exponentialfunktion nie null wird.

Damit ist in diesem Fall das erste Problem auf eine einfache Integration zurückgeführt.

Bleibt noch die Frage, was passiert, wenn $y(t)$ an irgendeinem Punkt t_0 eine Nullstelle haben sollte. Wir wollen uns überlegen, daß $y(t)$ dann auch für jedes $t > t_0$ verschwinden muß.

Falls nicht, gibt es einen Punkt $t_1 > t_0$, so daß $y(t_1) \neq 0$ ist. Wegen der Stetigkeit von $y(t)$ ist die Funktion dann auch in einer Umgebung von t_1 von Null verschieden, d.h. dort können wir die obigen Argumente anwenden und sehen, daß $y(t)$ dort die Form $e^{h(t)}$ hat mit irgendeiner Funktion h. Da y als differenzierbare Funktion insbesondere überall stetig sein muß und $e^{h(t)}$ nirgends verschwindet, ist das nicht möglich. Genauso überlegt man sich, daß $y(t)$ für jedes $t < t_0$ verschwinden muß, $y(t)$ ist also gleich der Nullfunktion. Diese ist somit die einzige Lösung, die noch zusätzlich betrachtet werden muß. Insbesondere folgt daraus auch, daß eine Lösungsfunktion, die in irgendeinem Punkt positiv *bzw.* negativ ist, überall positiv *bzw.* negativ sein muß, denn eine stetige Funktion kann ihr Vorzeichen nur wechseln, wenn sie in irgendeinem Punkt null wird; wie wir gerade gesehen haben, ist das genau dann der Fall, wenn sie überall verschwindet.

Somit hat jede Lösung die Form

$$y(t) = a e^{\int a(t)\,dt} \quad \text{mit einem } a \in \mathbb{R}$$

und umgekehrt ist auch jede dieser Funktionen eine Lösung. Insbesondere ist die Lösungsmenge ein eindimensionalen Vektorraum.

Es wäre schön, wenn wir im mehrdimensionalen Fall genauso vorgehen könnten: In Analogie zu

$$\dot{y}(t) = a(t) \cdot y(t) \Longrightarrow y(t) = c e^{\int a(t)\,dt} \quad \text{mit} \quad c \in \mathbb{R}$$

könnte vielleicht gelten

$$\dot{\vec{y}}(t) = A(t) \cdot \vec{y}(t) \Longrightarrow y(t) = e^{\int A(t)\,dt} \cdot \vec{c} \quad \text{mit} \quad \vec{c} \in \mathbb{R}^n \quad ?$$

Das Problem dabei ist nur, daß wir hier nicht die geringste Ahnung haben, was die rechte Seite bedeuten soll; unser nächstes Ziel wird sein, ihr eine Bedeutung zu geben und uns dann zu überlegen, ob *bzw.* unter welchen Bedingungen die obige Formel korrekt ist.

e) Die Matrixexponentialfunktion

Wir orientieren uns wieder am Eindimensionalen: Für eine reelle oder komplexe Zahl x ist

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!} ,$$

also setzen wir analog für eine reelle oder komplexe $n \times n$-Matrix X

$$e^X \underset{\text{def}}{=} \sum_{i=0}^{\infty} \frac{1}{i!} \cdot X^i .$$

Damit ist klar, daß e^X eine $n \times n$-Matrix sein soll, und das erklärt auch, warum oben der Konstantenvektor $\vec{y}_0$ auf der rechten Seite steht. Was wir uns noch überlegen müssen, ist die Konvergenz der Reihe.

Dazu müssen wir die Größe der Einträge in den Matrizen X^i abschätzen: Sind allgemein A, B zwei $n \times n$-Matrizen und sind die Beträge aller Einträge von A kleiner oder gleich a und die von B kleiner oder gleich b, so kann es in AB offensichtlich keinen Eintrag geben, dessen Betrag größer ist als nab: Schließlich ist jeder Eintrag in der Produktmatrix eine Summe von n Summanden, deren jeder Produkt je eines Eintrags von A und von B ist.

Um diese Formel leichter anwenden zu können, machen wir sie mutwillig schlechter und begnügen uns damit, daß jeder Eintrag von AB höchstens den Betrag $(na) \cdot (nb)$ hat.

Ist nun x der Betrag des größten Eintrags in der Matrix X, so folgt induktiv sofort, daß in X^i höchstens Zahlen bis zum Betrag $(nx)^i$ stehen können; in der endlichen Teilsumme

$$\sum_{i=0}^{M} \frac{1}{i!} X^i$$

hat daher jeder Eintrag einen Betrag kleiner

$$\sum_{i=0}^{M} \frac{(nx)^i}{i!} .$$

Letztere Summe konvergiert für $M \to \infty$ gegen e^{nx}, und damit muß auch die Matrixsumme absolut konvergieren, denn die Reihe für e^{nx} ist konvergente Majorante des Betrags eines jeden Eintrags. Insbesondere hat jeder Eintrag von e^X höchstens den Betrag e^{nx}.

Damit wissen wir also, daß die Matrix e^X für jede $n \times n$-Matrix X existiert; somit ist die Funktion $t \mapsto e^{At}$ wohldefiniert.

f) Eigenschaften der Matrixexponentialfunktion

Wir können natürlich nicht erwarten, daß die Matrixexponentialfunktion alle schönen Eigenschaften der gewöhnlichen Exponentialfunktion erbt. Beispielsweise ist nur schwer vorstellbar, daß für beliebige Matrizen A und B gelten sollte $e^{A+B} = e^A \cdot e^B$: Da für zwei Matrizen A und B stets $A + B = B + A$ ist, müßte dann auch $e^A \cdot e^B = e^B \cdot e^A$ sein, was zumindest unwahrscheinlich aussieht. In der Tat ist etwa für

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{und} \quad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

sowohl A^2 als auch B^2 gleich der Nullmatrix, d.h.

$$e^A = E + A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} , \quad e^B = E + B = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

und

$$e^A \cdot e^B = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} .$$

Das Quadrat von $C = A + B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ ist aber gleich der Einheitsmatrix und daher ist

$$\begin{aligned} e^{A+B} &= E + C + \frac{1}{2!}E + \frac{1}{3!}C + \frac{1}{4!}E + \frac{1}{5!}C + \cdots \\ &= \left(\sum_{i=0}^{\infty} \frac{1}{(2i)!} \right) \cdot E + \left(\sum_{i=0}^{\infty} \frac{1}{(2i+1)!} \right) \cdot C \\ &= \cosh 1 \cdot E + \sinh 1 \cdot C = \begin{pmatrix} \cosh 1 & \sinh 1 \\ \sinh 1 & \cosh 1 \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} e + e^{-1} & e - e^{-1} \\ e - e^{-1} & e + e^{-1} \end{pmatrix} . \end{aligned}$$

Allgemeiner ist

$$e^{At} = E + tA = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \quad \text{und} \quad e^{Bt} = E + tB = \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}$$

und

$$e^{Ct} = \sum_{i=0}^{\infty} \frac{1}{i!} C^i = \left(\sum_{i=0}^{\infty} \frac{1}{(2i)!} \right) \cdot E + \left(\sum_{i=0}^{\infty} \frac{1}{(2i+1)!} \right) \cdot C = \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix} .$$

Zum Glück gilt aber wenigstens

Lemma: Für zwei Matrizen $A, B \in \mathbb{C}^{n \times n}$ mit $AB = BA$ ist

$$e^{A+B} = e^A \cdot e^B = e^B \cdot e^A .$$

Insbesondere ist für $s, t \in \mathbb{R}$

$$e^{A(s+t)} = e^{As} \cdot e^{At} .$$

Beweis: Für zwei reelle Zahlen x, y ist $e^{x+y} = e^x e^y = e^y e^x$; damit gilt dieselbe Formel auch für zwei reellwertige Variablen x und y. Wenn wir in allen Potenzreihen alle x- und y-Potenzen oberhalb der N-ten ignorieren, sagt die Gleichung aus, daß drei Polynome in x und y *als Polynome* identisch sind.

Beim Rechnen mit Polynomen in x und y verwendet man keine speziellen Eigenschaften dieser Variablen *außer, daß sie kommutieren.* Damit kann man in so eine Polynomidentität auch kommutierende Matrizen A und B einsetzen: Beispielsweise führt die Polynomidentität

$$(x+y)^2 = x^2 + 2xy + y^2$$

zur Identität

$$(A+B)^2 = A^2 + 2AB + B^2 ,$$

die wegen der für beliebige Matrizen gültigen Gleichung

$$(A+B)^2 = A^2 + AB + BA + B^2$$

für kommutierende Matrizen in der Tat erfüllt ist.

Damit ist für kommutierende Matrizen A, B speziell stets

$$e^{A+B} = e^A \cdot e^B = e^B \cdot e^A .$$

Da zwei skalare Vielfache derselben Matrix stets miteinander kommutieren, folgt damit auch die letzte Aussage des Lemmas. ■

Die für uns wichtigste Anwendung hiervon ist

Satz: Für jede $n \times n$-Matrix $A \in \mathbb{C}^{n \times n}$ ist die Funktion

$$\begin{cases} \mathbb{R} \to \mathbb{C}^{n \times n} \\ t \mapsto e^{At} \end{cases}$$

stetig differenzierbar mit Ableitung $t \mapsto A \cdot e^{At} = e^{At} \cdot A$.

Beweis: Die Ableitung ist definiert als

$$\lim_{h \to 0} \frac{e^{A(t+h)} - e^{At}}{h} .$$

Da die Matrizen At und Ah miteinander vertauschbar sind, ist nach dem gerade bewiesenen Lemma $e^{A(t+h)} = e^{At} \cdot e^{Ah}$, also

$$\frac{e^{A(t+h)} - e^{At}}{h} = e^{At} \cdot \frac{e^{Ah} - E}{h} .$$

Dabei ist

$$\frac{e^{Ah} - E}{h} = \frac{1}{h} \cdot \sum_{i=1}^{\infty} \frac{(Ah)^i}{i!} = A + A^2 h \sum_{i=2}^{\infty} \frac{(Ah)^{i-2}}{i!} = A + A^2 h \sum_{i=0}^{\infty} \frac{(Ah)^i}{(i+2)!} .$$

Nach der obigen Diskussion ist jeder Eintrag der Matrix in der rechtsstehenden Summenmatrix höchstens gleich

$$\sum_{i=0}^{\infty} \frac{(ah)^i}{(i+2)!} = \frac{e^{ah} - (1+ah)}{a^2 h^2} ,$$

bleibt also insbesondere beschränkt. Dies gilt auch für $h \to 0$, denn wie zweimalige Anwendung der DE L'HOSPITALschen Regel oder TAYLOR-Entwicklung zeigen, ist der Grenzwert dann $\frac{1}{2}$. Damit existiert

$$\lim_{h \to 0} \sum_{i=1}^{\infty} \frac{(Ah)^i}{(i+2)!} ,$$

und somit ist $\frac{d}{dt}e^{At} = \lim_{h\to 0}\frac{e^{A(t+h)} - e^{At}}{h} = e^{At}\cdot A$, da der Vorfaktor A^2h gegen Null geht. Dies ist auch gleich $A\cdot e^{At}$, denn da A mit jeder seiner Potenzen vertauschbar ist, ist es auch mit jeder endlichen Teilsumme der Reihe von e^{At} vertauschbar, also auch mit e^{At} selbst. ■

Am Ende des vorigen Abschnitts hatten wir gehofft, daß vielleicht auch für jede matrixwertige Funktion $A(t)$ gelten könnte, daß

$$\frac{d}{dt}e^{A(t)} = \dot{A}(t)e^{A(t)}$$

ist; dies war offensichtlich zu optimistisch: Da

$$e^{A(t+h)} = e^{A(t)+h\dot{A}(t)+o(h)}$$

ist, bräuchten wir für einen Beweis nach obigem Vorbild, daß $A(t)$ und $\dot{A}(t)$ miteinander kommutieren; dies ist aber im allgemeinen nicht der Fall. Für

$$A(t) = \begin{pmatrix} 1 & t \\ 1 & 1 \end{pmatrix}$$

beispielsweise ist

$$\dot{A}(t) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

also

$$A(t)\cdot\dot{A}(t) = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \text{aber} \quad \dot{A}(t)\cdot A(t) = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}.$$

Dies ist natürlich kein Beweis dafür, daß die Ableitung von $e^{A(t)}$ ungleich $\dot{A}(t)\cdot e^{A(t)}$ ist, aber im vorliegenden Fall ist die Ableitung in der Tat verschieden sowohl von $\dot{A}(t)e^{A(t)}$ als auch von $e^{A(t)}\dot{A}(t)$: Mit den Methoden, die wir im nächsten Abschnitt kennenlernen werden, können wir durch eine (alles andere als angenehme) Rechnung zeigen, daß

$$e^{A(t)} = \begin{pmatrix} \frac{e^{1+\sqrt{t}} + e^{1-\sqrt{t}}}{2} & \frac{\sqrt{t}\left(e^{1+\sqrt{t}} - e^{1-\sqrt{t}}\right)}{2} \\ \frac{e^{1+\sqrt{t}} - e^{1-\sqrt{t}}}{2\sqrt{t}} & \frac{e^{1+\sqrt{t}} + e^{1-\sqrt{t}}}{2} \end{pmatrix},$$

$$\frac{d}{dt}e^{A(t)} = \begin{pmatrix} \frac{\left(e^{2\sqrt{t}}-1\right)e^{1-\sqrt{t}}}{4\sqrt{t}} & \frac{\left(-1+\sqrt{t}+\sqrt{t}e^{2\sqrt{t}}+e^{2\sqrt{t}}\right)e^{1-\sqrt{t}}}{4\sqrt{t}} \\ \frac{\left(1-e^{2\sqrt{t}}+\sqrt{t}+\sqrt{t}e^{2\sqrt{t}}\right)e^{1-\sqrt{t}}}{4t\sqrt{t}} & \frac{\left(e^{2\sqrt{t}}-1\right)e^{1-\sqrt{t}}}{4\sqrt{t}} \end{pmatrix},$$

aber

$$\dot{A}(t)\cdot e^{A(t)} = \begin{pmatrix} \frac{e^{1+\sqrt{t}} - e^{1-\sqrt{t}}}{2\sqrt{t}} & \frac{e^{1+\sqrt{t}} + e^{1-\sqrt{t}}}{2} \\ 0 & 0 \end{pmatrix}$$

und

$$e^{A(t)}\cdot\dot{A}(t) = \begin{pmatrix} 0 & \frac{e^{1+\sqrt{t}} + e^{1-\sqrt{t}}}{2} \\ 0 & \frac{e^{1+\sqrt{t}} - e^{1-\sqrt{t}}}{2\sqrt{t}} \end{pmatrix}$$

ist. Wir müssen uns bei diesem Ansatz also begnügen mit linearen homogenen Differentialgleichungen mit *konstanten* Koeffizienten.

Alles was uns zu deren theoretischer Lösung jetzt noch fehlt sind Rechenregeln für den Umgang mit Ableitungen von Matrixfunktionen; für die praktische Lösung fehlen natürlich auch noch Verfahren zur effizienten Berechnung der Matrixexponentialfunktion.

Zumindest für Summen und Produkte gelten, wenn man von der Nichtkommutativität der Multiplikation absieht, für matrixwertige Funktionen die üblichen Regeln:

Lemma: *a)* $F, G\colon (a,b) \to \mathbb{C}^{n\times n}$ seien zwei matrixwertige Funktionen auf dem offenen Intervall (a,b). Dann ist

$$\frac{d}{dt}\big(F(t)+G(t)\big) = \dot{F}(t)+\dot{G}(t).$$

b) Für $F\colon (a,b) \to \mathbb{C}^{n\times m}$ und $G\colon (a,b) \to \mathbb{C}^{m\times p}$ ist

$$\frac{d}{dt}\big(F(t)G(t)\big) = \dot{F}(t)\cdot G(t) + F(t)\cdot\dot{G}(t).$$

c) Für einen konstanten Vektor $\vec{v} \in \mathbb{C}^n$ ist

$$\frac{d}{dt}\big(F(t)\cdot\vec{v}\big) = \dot{F}(t)\cdot\vec{v}.$$

Beweis: a) Sind $f_{ij}(t)$ und $g_{ij}(t)$ die Komponenten der Matrizen F und G, so sind die Summen $f_{ij}(t)+g_{ij}(t)$ die Komponenten von $F+G$, und deren Ableitung ist die Summe der Ableitungen.

b) Die (i,j)-Komponente von FG ist die Funktion $\sum_{\nu=1}^{m} f_{i\nu}(t)\cdot g_{\nu j}(t)$, und deren Ableitung ist

$$\sum_{\nu=1}^{m}\left(\dot f_{i\nu}(t)\cdot g_{\nu j}(t)+f_{i\nu}(t)\cdot \dot g_{\nu j}(t)\right)\ ,$$

die (i,j)-Komponente von $\dot F(t)\cdot G(t)+F(t)\cdot \dot G(t)$.

c) Ist der Spezialfall $n+m$ und $p=1$ von *b)*, wobei zusätzlich noch G eine konstante Funktion ist, so daß alle Terme mit $\dot G(t)$ verschwinden. ■

Damit haben wir alles zusammen und können zeigen

Satz: Die sämtlichen Lösungen des homogenen Differentialgleichungssystems $\dot y = Ay$ mit $A\in\mathbb{R}^{n\times n}$ sind genau die Funktionen $t\mapsto e^{At}\vec y_0$ mit $\vec y_0\in\mathbb{R}^n$.

Beweis: Da e^{At} die Ableitung Ae^{At} hat, ist die Ableitung der vektorwertigen Funktion $\vec f(t)=e^{At}\cdot\vec y_0$ nach der zuletzt bewiesenen Formel gleich $A\cdot e^{At}\cdot\vec y_0$, also in der Tat gleich $A\cdot\vec f(t)$.

Nun sei $\vec y\colon(a,b)\to\mathbb{R}^n$ irgendeine differenzierbare vektorwertige Funktion mit der Eigenschaft, daß $\dot{\vec y}(t)=A\cdot\vec y(t)$ ist. Wir betrachten die Funktion $\vec g(t)=e^{-At}\cdot\vec y(t)$. Deren Ableitung ist

$$\dot{\vec g}(t)=-A\cdot e^{-At}\cdot\vec y(t)+e^{-At}\cdot\dot{\vec y}(t)=-A\cdot e^{-At}\cdot\vec y(t)+e^{-At}\cdot A\vec y(t)=\vec 0\,,$$

denn die Matrix A ist mit e^{-At} vertauschbar. Damit haben alle Komponenten von $\vec g$ die Ableitung Null, sind also konstant, und somit ist $\vec g(t)\underset{\text{def}}{=}\vec y_0$ ein konstanter Vektor mit der Eigenschaft, daß

$$\vec y_0=e^{-At}\cdot\vec y(t),\quad d.h.\quad \vec y(t)=e^{At}\cdot\vec y_0\,,$$

wie behauptet. ■

Korollar: Für jeden Vektor $\vec y_0\in\mathbb{R}^n$ und jede reelle Zahl $t_0\in\mathbb{R}$ gibt es genau eine differenzierbare Funktion $\vec y(t)$ mit den Eigenschaften, daß $\dot{\vec y}(t)=A\vec y(t)$ und $\vec y(t_0)=\vec y_0$ ist; dies ist $\vec y(t)=e^{A(t-t_0)}\cdot\vec y_0$. ■

Wir wußten bereits, daß die Lösungen einen Vektorraum bilden; obiges Korollar sagt uns, daß dieser Vektorraum die Dimension n hat und daß für jede reelle Zahl t_0 die Abbildung $\vec y\mapsto\vec y(t_0)$ ein Isomorphismus auf den $\mathbb{R}^n$ ist.

Als erstes Beispiel betrachten wir das Anfangswertproblem

$$\dot x(t)=y(t)\quad\text{und}\quad \dot y(t)=x(t)\quad\text{mit}\quad x(0)=a\quad\text{und}\quad y(0)=b$$

oder

$$\begin{pmatrix}\dot x(t)\\ \dot y(t)\end{pmatrix}=\begin{pmatrix}0&1\\1&0\end{pmatrix}\begin{pmatrix}x(t)\\y(t)\end{pmatrix}\quad\text{mit}\quad\begin{pmatrix}x(0)\\y(0)\end{pmatrix}=\begin{pmatrix}a\\b\end{pmatrix}.$$

Die Koeffizientenmatrix ist gleich der zu Beginn des Abschnitts betrachteten Beispielmatrix C, deren Exponentialfunktion wird dort berechnet haben; die Lösung des Anfangswertproblems ist also

$$e^{Ct}\begin{pmatrix}a\\b\end{pmatrix}=\begin{pmatrix}\cosh t&\sinh t\\\sinh t&\cosh t\end{pmatrix}\begin{pmatrix}a\\b\end{pmatrix}=\begin{pmatrix}a\cosh t+b\sinh t\\a\sinh t+b\cosh t\end{pmatrix}.$$

§2: Eigenwerte, Eigenvektoren und Hauptvektoren

Die Matrixexponentialfunktion ist zwar wohldefiniert, aber eine matrixwertige Potenzreihe ist für allgemeine Matrizen A nicht gerade einfach zu berechnen. Wir brauchen daher alternative Rechenverfahren.

Zumindest ein Fall ist problemlos: Ist nämlich

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

eine Diagonalmatrix, ist offensichtlich

$$e^{Dt}=\begin{pmatrix}e^{d_1t}&0&\dots&0\\0&e^{d_2t}&\dots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\dots&e^{d_nt}\end{pmatrix}$$

wieder eine Diagonalmatrix, wobei die Exponentialfunktion einfach komponentenweise auf die Diagonaleinträge ihres Arguments angewandt wird.

Noch ein weiterer Fall ist relativ unproblematisch: für eine obere (oder untere) Dreiecksmatrix N mit Nullen in der Hauptdiagonalen. Eine solche Matrix definiert eine lineare Abbildung $\mathbb{C}^n \to \mathbb{C}^n$, die den k-ten Einheitsvektor $\vec{e}_k$ in den von $\vec{e}_{k+1}$ bis $\vec{e}_n$ erzeugten Untervektorraum abbildet. Das Quadrat von N bildet ihn entsprechend in den von $\vec{e}_{k+2}$ bis $\vec{e}_n$ erzeugten Untervektorraum ab und so weiter, spätestens N^n ist also die Nullmatrix. Damit wird die Potenzreihe der Exponentialfunktion zu einer endlichen Summe, die, wir wir bereits in zwei Beispielen gesehen haben, zumindest für kleine n leicht berechnet werden kann.

Wir wissen bereits aus dem letzten Semester (*Kap. I, §3j1)*), welche Bedingung eine Basis von $\mathbb{R}^n$ *bzw.* $\mathbb{C}^n$ erfüllen muß, damit eine Matrix A bezüglich dieser Basis Diagonalgestalt hat: Die Basisvektoren müssen allesamt Eigenvektoren von A sein. Aus *Kap. I, §6i)* wissen wir auch, wie man Eigenwerte und ausgehend davon Eigenvektoren mit Hilfe von Determinanten bestimmen kann. In diesem Paragraphen wollen wir wir die entsprechende Theorie noch etwas weiterentwickeln und sehen, daß sich die Berechnung einer beliebigen Matrixexponentialfunktion auf die beiden gerade diskutierten Spezialfälle zurückführen läßt.

Wir arbeiten dabei wieder, wie im ersten Kapitel, über einem beliebigen Körper k, denn auch wenn uns im Augenblick zur Anwendung auf Differentialgleichungen nur die Fälle $k = \mathbb{R}$ und $k = \mathbb{C}$ interessieren, hat die hier entwickelte Theorie doch auch interessante Anwendungen über anderen Körpern: Eigenvektoren über endlichen Körpern spielen beispielsweise bei einigen Problem der Signalverarbeitung eine Rolle.

a) Mehr über Eigenwerte und Eigenvektoren

Zur Bequemlichkeit der Leser sei die Definition noch einmal wiederholt:

Definition: *a)* V sei ein k-Vektorraum. Ein Vektor $\vec{v} \in V \setminus \{\vec{0}\}$ heißt *Eigenvektor* der linearen Abbildung $\varphi\colon V \to V$ zum *Eigenwert* $\lambda \in k$, wenn $\varphi(\vec{v}) = \lambda\vec{v}$ ist.

b) $\lambda \in k$ heißt *Eigenwert* von φ, falls φ einen Eigenvektor zum Eigenwert λ hat.

c) Eigenwerte und Eigenvektoren einer Matrix $A \in k^{n\times n}$ sind die Eigenwerte und Eigenvektoren der linearen Abbildung

$$\varphi\colon \begin{cases} k^n \to k^n \\ \vec{v} \mapsto A\vec{v} \end{cases} .$$

Offensichtlich ist mit einem Vektor $\vec{v}$ auch jedes Vielfache (außer dem nach Definition ausgeschlossenen Nullvektor) ein Eigenvektor zum selben Eigenwert; allgemeiner ist sogar jede Linearkombination (außer $\vec{0}$) von Eigenvektoren zum Eigenwert λ wieder ein Eigenvektor zum Eigenwert λ, d.h. die Eigenvektoren zu einem festen Eigenwert λ bilden zusammen mit dem Nullvektor einen Untervektorraum von V, den sogenannten *Eigenraum* von λ.

Definition: Die Dimension des Eigenraums von λ heißt *geometrische Vielfachheit* des Eigenwerts λ.

Lemma: Sind $\vec{v}_1, \dots, \vec{v}_r \in V$ Eigenvektoren der linearen Abbildung $\varphi\colon V \to V$ zu verschiedenen Eigenwerten $\lambda_1, \dots, \lambda_r$, so sind diese Vektoren linear unabhängig.

Beweis: Angenommen, $\vec{v}_1, \dots, \vec{v}_r$ seien linear abhängig. Dann können wir eine Zahl $2 \le s \le r$ finden, so daß zwar $\vec{v}_1, \dots, \vec{v}_s$ linear abhängig sind, nicht aber $\vec{v}_1, \dots, \vec{v}_{s-1}$. Es gibt daher Skalare $\alpha_i \in k$, so daß

$$\alpha_1\vec{v}_1 + \dots + \alpha_s\vec{v}_s = \vec{0}$$

ist. Wenden wir auf beide Seiten dieser Gleichung die Abbildung φ an und beachten, daß $\varphi(\vec{v}_i) = \lambda_i\vec{v}_i$ ist, folgt, daß auch

$$\alpha_1\lambda_1\vec{v}_1 + \dots + \alpha_s\lambda_s\vec{v}_s = \vec{0}$$

ist. Andererseits können wir obige Gleichung auch einfach mit λ_s multiplizieren mit dem Ergebnis, daß

$$\lambda_s\alpha_1\vec{v}_1 + \dots + \lambda_s\alpha_s\vec{v}_s = \vec{0}\,.$$

Durch Subtraktion der letzten beiden Gleichungen voneinander erhalten wir eine lineare Abhängigkeit

$$\alpha_1(\lambda_s - \lambda_1)\vec{v}_1 + \cdots + \alpha_{s-1}(\lambda_s - \lambda_{s-1})\vec{v}_{s-1} = \vec{0}$$

zwischen $\vec{v}_1, \ldots, \vec{v}_{s-1}$. Da diese Vektoren linear unabhängig sind, müssen alle Koeffizienten verschwinden. Da die Eigenwerte $\lambda_1, \cdots, \lambda_s$ aber allesamt verschieden sind, ist dies nur möglich, wenn α_1 bis α_{s-1} verschwinden. Wegen $\vec{v}_s \neq \vec{0}$ muß dann aber auch α_s verschwinden, im Widerspruch zur angenommenen linearen Unabhängigkeit von $\vec{v}_1, \ldots, \vec{v}_s$.

Also sind die Vektoren $\vec{v}_1, \ldots, \vec{v}_r$ linear unabhängig. ■

Eigenwerte und Eigenvektoren sind auch interessant für Selbstabbildungen eines unendlichdimensionalen Vektorraums: Ist $V = \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ beispielsweise der Vektorraum aller beliebig oft stetig differenzierbarer reeller Funktionen, so sind $\sin \omega t$ und $\cos \omega t$ Eigenvektoren der linearen Abbildung

$$\varphi: \begin{cases} \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}) \to \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}) \\ f \mapsto \ddot{f} \end{cases}$$

zum Eigenwert $-\omega^2$; genauso sind $\sinh \omega t$ und $\cosh \omega t$ Eigenvektoren zum Eigenwert ω^2. Für $\omega = 0$ degenerieren diese beiden Eigenvektoren jeweils zu null und eins, wodurch ein Eigenwert verschwindet; dafür kommt die Identität als neuer Eigenvektor hinzu. Damit ist also jede reelle Zahl Eigenwert von φ mit einer geometrischen Vielfachheit von mindestens zwei. (Wir werden im nächsten Paragraphen sehen, daß die Vielfachheit immer gleich zwei ist.)

Eigenwertprobleme für lineare Abbildungen, die durch Differentialoperatoren gegeben sind, spielen in vielen Anwendungen eine wichtige Rolle; im Hinblick auf solche Anwendungen bezeichnet man die Menge aller Eigenwerte einer linearen Abbildung oder Matrix auch als deren *Spektrum.* Dieses Wort kommt daher, daß z.B. beim (mehrdimensionalen) Differentialoperator, der die Schwingungen des Fells einer Trommel beschreibt, die Eigenwerte gerade die Frequenzen sind, die die Trommel produzieren kann.

Uns interessieren hauptsächlich Eigenwerte und Eigenvektoren in endlichdimensionalen Vektorräumen. Dort können wir konkret mit Matrizen rechnen; ist A die Abbildungsmatrix zu $\varphi: V \to V$, so ist $\varphi(\vec{v}) = \lambda\vec{v}$ äquivalent dazu, daß $A\vec{v} = \lambda\vec{v}$ oder $(A - \lambda E)\vec{v} = \vec{0}$ ist, wobei E wie üblich die Einheitsmatrix bezeichnet.

In letzterer Form ist dies jenes homogene lineare Gleichungssystem für die Komponenten von $\vec{v}$, das wir bereits in *Kap. I, §6i1)* betrachtet haben. Wie jedes homogene lineare Gleichungssystem hat es den Nullvektor als Lösung, der allerdings nach Definition genau aus diesem Grund *nicht* als Eigenvektor betrachtet wird. Weitere Lösungen gibt es genau dann, wenn die Matrix $A - \lambda E$ des Gleichungssystems singulär ist, wenn also $\det(A - \lambda E)$ verschwindet. Somit ist $\lambda \in k$ genau dann ein Eigenwert, wenn $\det(A - \lambda E) = 0$ ist; die zugehörigen Eigenvektoren sind die nichttrivialen Lösungen des homogenen linearen Gleichungssystems $(A - \lambda E)\vec{v} = \vec{0}$.

Damit ist klar, wie man Eigenwerte und Eigenvektoren berechnen kann: Man löse die Gleichung $\det(A - \lambda E) = 0$ und dann für jede Nullstelle λ_i dieser Gleichung das lineare Gleichungssystem $(A - \lambda E)\vec{v} = \vec{0}$. Dieses homogene lineare Gleichungssystem hat *nie* maximalen Rang, da es nach Definition eines Eigenwerts nichttriviale Lösungen geben muß; kommt man also auf ein eindeutig lösbares Gleichungssystem (und damit auf den Nullvektor als einzige Lösung), ist das immer ein Zeichen für einen Rechenfehler.

b) Ein erstes Beispiel

Wir wollen e^A *bzw.* e^{At} berechnen für die bereits in *Kap. I, §6i)* betrachtete Matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 8 & 7 & 6 & 5 \\ 4 & 3 & 2 & 1 \end{pmatrix}.$$

Wie wir dort nachgerechnet haben, ist hier

$$\det(A - \lambda E) = \lambda^2(\lambda + 4)(\lambda - 18)$$

mit Nullstellen $\lambda_1 = \lambda_2 = 0, \lambda_3 = -4$ und $\lambda_4 = 18$. Als Eigenvektoren dazu hatten wir die Vektoren

$$\vec{b}_1 = \lambda \begin{pmatrix} 1 \\ -2 \\ 1 \\ 0 \end{pmatrix}, \ \vec{b}_2 = \begin{pmatrix} 2 \\ -3 \\ 0 \\ 1 \end{pmatrix}, \ \vec{b}_3 = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix} \text{ und } \vec{b}_4 = \begin{pmatrix} 5 \\ 13 \\ 13 \\ 5 \end{pmatrix}$$

gefunden, wobei $\vec{b}_i$ Eigenvektor zum Eigenwert λ_i ist, d.h.

$$\varphi(\vec{b}_1) = 0\,\vec{b}_1, \quad \varphi(\vec{b}_2) = 0\,\vec{b}_2, \quad \varphi(\vec{b}_3) = -4\,\vec{b}_3 \quad \text{und} \quad \varphi(\vec{b}_4) = 18\,\vec{b}_4 .$$

Bezüglich der neuen Basis $(\vec{b}_1, \vec{b}_2, \vec{b}_3, \vec{b}_4)$ hat φ daher die Abbildungsmatrix

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 18 \end{pmatrix},$$

bei der die Eigenwerte von A in der Hauptdiagonalen stehen und alle sonstigen Einträge verschwinden. Bei dieser Matrix haben wir keinerlei Probleme mit der Berechnung der Exponentialfunktion: Offensichtlich ist

$$e^M = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & e^{-4} & 0 \\ 0 & 0 & 0 & e^{18} \end{pmatrix} \quad \text{und} \quad e^{Mt} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & e^{-4t} & 0 \\ 0 & 0 & 0 & e^{18t} \end{pmatrix}.$$

Damit läßt sich auch e^A berechnen: Sind nämlich

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad \vec{e}_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

die vier Einheitsvektoren des $\mathbb{R}^4$, so ist

$$\vec{b}_i = B\vec{e}_i \quad \text{mit} \quad B = \begin{pmatrix} 1 & 2 & -1 & 5 \\ -2 & -3 & -1 & 13 \\ 1 & 0 & 1 & 13 \\ 0 & 1 & 1 & 5 \end{pmatrix},$$

und die Gleichung $A\vec{b}_i = \lambda_i \vec{b}_i$ mit $\lambda_{1/2} = 0, \lambda_3 = -4$ und $\lambda_4 = 18$ wird zu $AB\vec{e}_i = \lambda_i B\vec{b}_i$ oder $B^{-1}AB\,\vec{e}_i = \lambda_i \vec{e}_i$.

Also ist $B^{-1}AB = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 18 \end{pmatrix} = M$ und $A = BMB^{-1}$.

Wir wollen uns überlegen, daß sich eine solche Relation auch in Potenzen sowie in die Exponentialfunktion hineinziehen läßt:

Lemma: Ist $B \in k^{n\times n}$ eine invertierbare Matrix, $M \in k^{n\times n}$ irgendeine Matrix und m eine ganze Zahl, so ist

$$\left(BMB^{-1}\right)^m = BM^mB \quad \text{und} \quad e^{BMB^{-1}} = Be^MB^{-1} .$$

Zum *Beweis* betrachten wir zunächst eine natürliche Zahl $m \in \mathbb{N}$; für diese ist

$$\begin{aligned} \left(BMB^{-1}\right)^m &= \underbrace{BMB^{-1}BMB^{-1}\cdots BMB^{-1}}_{m \text{ mal}} \\ &= B\cdot M\cdot E\cdot M\cdots M\cdot E\cdot MB^{-1} = BM^mB^{-1}, \end{aligned}$$

da BB^{-1} die Einheitsmatrix ist. Für $m = 0$ gibt es ebenfalls keine Probleme, da die nullte Potenz *jeder* $n \times n$-Matrix die Einheitsmatrix ist, und für negative m schließlich ist BM^mB^{-1} invers zu $BM^{-m}B^{-1}$, denn

$$BM^mB^{-1}\cdot BM^{-m}B^{-1} = B\cdot M^m\cdot M^{-m}B^{-1} = B\cdot B^{-1} = E .$$

Also ist

$$BM^mB^{-1} = \left(BM^{-m}B^{-1}\right)^{-1} = \left(\left(BMB^{-1}\right)^{-m}\right)^{-1} = \left(BMB^{-1}\right)^m ,$$

da $-m$ eine natürliche Zahl ist, für die wir die Formel bereits bewiesen haben. Schließlich ist auch

$$e^{BMB^{-1}} = \sum_{m=0}^{\infty} \frac{1}{m!}\left(BMB^{-1}\right)^m = \sum_{m=0}^{\infty} \frac{1}{m!} BM^mB^{-1} = Be^MB^{-1} ,$$

wie behauptet. ∎

In unserem Fall erhalten wir

$$e^A = B \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & e^{-4} & 0 \\ 0 & 0 & 0 & e^{18} \end{pmatrix} B^{-1},$$

was sich zumindest im Prinzip ausrechnen läßt – auch wenn das Ergebnis

$$\frac{1}{72} \begin{pmatrix} 27e^{-4}+35+10e^{18} & 10e^{18}-19+9e^{-4} & 10e^{18}-1-9e^{-4} & 10e^{18}+17-27e^{-4} \\ -53+27e^{-4}+26e^{18} & 9e^{-4}+26e^{18}+37 & 26e^{18}-17-9e^{-4} & 26e^{18}+1-27e^{-4} \\ 26e^{18}+1-27e^{-4} & 26e^{18}-17-9e^{-4} & 9e^{-4}+26e^{18}+37 & -53+27e^{-4}+26e^{18} \\ 10e^{18}+17-27e^{-4} & 10e^{18}-1-9e^{-4} & 10e^{18}-19+9e^{-4} & 27e^{-4}+35+10e^{18} \end{pmatrix}$$

alles andere als angenehm ist. Dies zeigt wieder einmal, wieviel man sich ersparen kann, wenn man *vor* Beginn einer Rechnung eine gute Basis *bzw.* ein gutes Koordinatensystem wählt.

c) Das charakteristische Polynom und seine Nullstellen

Es ist kein Zufall, daß im obigen Beispiel die Gleichung $\det(A-\lambda E) = 0$ auf ein Polynom vierten Grades führte: Ist $A = (a_{ij})$ eine $n \times n$-Matrix, so hat die Matrix $A - \lambda E$ in der Diagonalen die Einträge $a_{ii} - \lambda$, ansonsten stimmen alle Einträge mit denen von A überein. Berechnet man daher $\det(A - \lambda E)$ gemäß der definierenden Formel, so gibt es genau ein Produkt, in dem n mit λ behaftete Faktoren vorkommen, nämlich das Produkt

$$(a_{11} - \lambda) \cdots (a_{nn} - \lambda) = (-1)^n \lambda^n + \textit{Terme niedrigerer Ordnung}$$

der Diagonaleinträge. Die restlichen Produkte, die zur Determinante aufsummiert werden, enthalten zwischen null und $n-1$ mit λ behaftete Faktoren, die Summe ist also ein Polynom vom Grad n mit höchstem Term $(-1)^n \lambda^n$.

Definition: Das Polynom $\det(A-\lambda E)$ heißt *charakteristisches Polynom* der Matrix A.

Demgemäß sind also die Eigenwerte von A gleich den Nullstellen des charakteristischen Polynoms von A, und wir sollten uns wenigsten kurz überlegen, wie man die Nullstellen eines solchen Polynoms bestimmen kann.

Zur Bestimmung der Eigenwerte muß man somit die Nullstellen des charakteristischen Polynoms finden. Für ein Polynom vom Grad höchstens zwei (oder aber ein Polynom, das man als Produkt solcher Polynome schreiben kann) ist das nicht schwer: Nullstellen eines linearen Polynoms erhält man durch eine einfache Division, solche eines quadratischen durch quadratische Ergänzung: Da für $a \neq 0$

$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 + c - \frac{b^2}{4a}$$

ist, hat das linksstehende Polynom die Nullstellen

$$x_{1/2} = -\frac{b}{2a} \pm \sqrt{\frac{b^2}{4a} - c} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Bei Polynomen höherer Grade, die man nicht auf einfache Weise über binomische Formeln oder ähnliches in kleinere Faktoren zerlegen kann, ist es oft einen Versuch wert, einige der Lösungen zu *erraten,* um so den Grad des Polynoms zu reduzieren.

Bei Polynomen mit ganzzahligen (und eventuell auch rationalen) Nullstellen, ist dazu der Wurzelsatz von VIÈTE ein vielversprechender Ansatzpunkt: Angenommen, das Polynom n-ten Grades

$$f(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_2x^2 + a_1x + a_0$$

mit höchstem Koeffizient eins habe die Nullstellen $z_1, \ldots, z_n$. Dann ist

$$f(x) = (x - z_1)(x - z_2) \cdot \cdots \cdot (x - z_n).$$

Dies läßt sich ausmultiplizieren und liefert dann einen Zusammenhang zwischen Nullstellen und Koeffizienten: Beispielsweise ist

$$a_0 = (-1)^n z_1 z_2 \cdots z_n \quad \text{und} \quad a_{n-1} = -(z_1 + z_2 + \cdots + z_n),$$

und genauso zeigt man auch daß der allgemeine Koeffizient a_i die Summe aller möglicher Produkte aus $n-i$ Nullstellen z_j ist, multipliziert mit $(-1)^{n-i}$. Diese Aussage bezeichnet man als den Wurzelsatz von VIÈTE.

FRANÇOIS VIÈTE (1540–1603) studierte Jura an der Universität Poitiers, danach arbeitete er als Hauslehrer. 1573, ein Jahr nach dem Massaker an den Hugenotten, berief ihn CHARLES IX (obwohl VIÈTE Hugenotte war) in die Regierung der Bretagne; unter HENRI III wurde er geheimer Staatsrat. 1584 wurde er auf Druck der katholischen Liga vom Hofe verbannt und beschäftigte sich fünf Jahre lang mit Mathematik. Unter HENRI IV arbeitete er wieder am Hof und knackte u.a. verschlüsselte Botschaften an den spanischen König PHILIP II. In seinem Buch *In artem analyticam isagoge* rechnete er als erster systematisch mit symbolischen Größen.

Für das Erraten von Nullstellen einfacher Polynome, bei denen man (aus inhaltlichen Gründen oder aber weil so etwas in Übungs- und Klausuraufgaben fast die Regel ist) ganzzahlige Lösungen erwartet, ist vor allem die erstgenannte Beziehung wichtig: In der Form

$$(-1)^n a_0 = z_1 z_2 \cdots z_n$$

gibt sie das Produkt aller Nullstellen. Falls die a_i alle ganzzahlig sind, lohnt es sich also, die Teiler von a_0 zu testen. Ist beispielsweise

$$f(x) = x^4 + 14x^3 - 52x^2 - 14x + 51\,,$$

so ist

$$a_0 = 51 = 3 \cdot 17\,.$$

Da das Produkt aller Nullstellen gleich diesem Wert sein muß, kommen – falls *alle* Nullstellen ganzzahlig sind – für diese nur die Werte $\pm 1, \pm 3$ und ± 17 in Frage. Da das Produkt aller vier Nullstellen gleich 51 ist, gibt es jeweils genau eine Nullstelle vom Betrag 3 *bzw.* 17, sowie zwei Nullstellen vom Betrag eins. Welche Vorzeichen wirklich auftreten, läßt sich durch Einsetzen feststellen oder aber auch dadurch, daß nach VIÈTE die *Summe* aller Nullstellen gleich -14 sein muß. Das ist offenbar nur möglich, wenn sowohl +1 als auch -1 Nullstellen sind, sowie -17 und +3. In der Tat zeigt Einsetzen, daß dies auch tatsächlich Nullstellen sind. (Das Einsetzen ist notwendig, da wir nicht sicher sein können, daß wirklich alle Nullstellen ganzzahlig sind.)

In diesem extrem einfachen (und konstruierten) Fall führt also die Primfaktorzerlegung direkt zur Lösung; in komplizierteren Fällen, wenn a_0

mehr Primfaktoren hat, muß man zunächst alle Kombinationsmöglichkeiten, die zum Produkt a_0 führen können, in Betracht ziehen und davon dann durch Einsetzen potentieller Nullstellen alle bis auf die tatsächlichen Nullstellen eliminieren.

Beim Polynom

$$f(x) = x^6 + 27x^5 - 318x^4 - 5400x^3 - 10176x^2 + 27648x + 32768$$

etwa ist $a_0 = 32768 = 2^{15}$; hier wissen wir also nur, daß – sofern alle Nullstellen ganzzahlig sind – jede Nullstelle die Form $\pm 2^i$ haben muß, wobei die Summe aller Exponenten gleich 15 sein muß und die Anzahl der negativen Vorzeichen gerade. Einsetzen zeigt, daß

$$-1, \quad 2, \quad -4, \quad -8, \quad 16, \quad -32$$

die Nullstellen sind.

Man beachte, daß diese Vorgehensweise nur funktioniert, wenn das Polynom höchsten Koeffizienten eins hat; andernfalls ist das Produkt der Nullstellen gleich dem Quotienten aus konstantem Koeffizienten und führendem Koeffizienten mal $(-1)^{\text{Grad}}$.

Falls man nicht sicher sein kann, daß alle Nullstellen ganzzahlig sind, gibt es immer noch eine ganze Reihe von Methoden, um Nullstellen *exakt* zu berechnen: Beispielsweise kennt die Computeralgebra Algorithmen, um ein Polynom (soweit dies möglich ist) in ein Produkt von Polynomen kleineren Grades zu zerlegen mit Koeffizienten aus einem vorgegebenen Körper, der (in einem hier nicht präzisierten) Sinne nicht *zu weit* vom Körper der rationalen Zahlen *bzw.* einem endlichen Körper entfernt ist, und es gibt auch, seit der ersten Hälfte des sechzehnten Jahrhunderts, allgemeine Formeln zur Lösung von Gleichungen dritten und vierten Grades. Diese Formeln spielen wegen ihrer Komplexität und numerischen Instabilität in der Praxis keine sonderlich große Rolle und sollen daher hier nur im Kleindruck behandelt werden:

Für die kubische Gleichung

$$ax^3 + bx^2 + cx + d = 0$$

wenden wir zunächst einen ähnlichen Trick an wie die quadratische Ergänzung beim Fall der quadratischen Gleichungen: Durch die Substitution

$$z = x + \frac{b}{3a}$$

wird die Gleichung zu

$$az^3 + \left(c - \frac{b^2}{3a}\right) z + \frac{2b^3}{27a^2} + d\,,$$

was wir auch kurz als

$$z^3 + pz + q = 0$$

schreiben können. Zur Lösung dieser Gleichung ersetzen wir z durch die Summe

$$z = u + v$$

zweier Variablen und erhalten

$$u^3 + 3u^2v + 3uv^2 + v^3 = u^3 + v^3 + 3uv(u+v) + p(u+v) + q = 0\,.$$

Da die Zerlegung von z in eine Summe äußerst willkürlich ist, können wir hoffen, daß diese Gleichung für die beiden Variablen u und v auch Lösungen hat, wenn wir zusätzliche Bedingungen stellen: Die obige Gleichung für z wird beispielsweise sicherlich dann gelöst, wenn

$$u^3 + v^3 = -q \quad \text{und} \quad 3uv = -p$$

ist. Dann ist

$$u^3 + v^3 = -q \quad \text{und} \quad u^3 \cdot v^3 = -\frac{p}{3}^3\,,$$

wir kennen also Summe und Produkt von u^3 und v^3.

Sind aber Summe und Produkt zweier Zahlen r und s bekannt, so können wir leicht die Zahlen selbst bestimmen: Aus

$$r + s = c \quad \text{und} \quad rs = d$$

folgt, daß

$$r(c-r) = -r^2 + cr = d \quad \text{oder} \quad r^2 - cr + d = 0$$

ist; wir müssen also einfach eine quadratische Gleichung lösen und erhalten

$$r = \frac{c}{2} \pm \sqrt{\frac{c^2}{4} - d}\,.$$

Da die Summe dieser beiden Lösungen gleich c ist, muß also die eine gleich r und die andere gleich s sein.

Auf die kubische Gleichung angewandt heißt das, daß u^3 und v^3 die beiden Zahlen

$$-\frac{q}{2} \pm \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}$$

sind, also

$$u = \sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} \quad \text{und} \quad v = \sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}}$$

oder umgekehrt.

Uns interessiert nur die Summe der beiden Zahlen, also

$$z = \sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}}\,.$$

Damit sind wir fast fertig. Das verbleibende Problem ist, daß hier formal eine Lösung steht, wohingegen wir für eine kubische Gleichung *drei* Lösungen erwarten. Dieses Problem verkehrt sich sofort in sein Gegenteil, wenn wir beachten, daß genauso, wie die Quadratwurzel nur bis aufs Vorzeichen bestimmt ist, die Kubikwurzel nur bis dritte Einheitswurzeln bestimmt ist: Da die drei komplexen Zahlen

$$1, \quad \frac{-1 + i\sqrt{3}}{2} \quad \text{und} \quad \frac{-1 - i\sqrt{3}}{2}$$

alle dritte Potenz eins haben, ist mit jeder Kubikwurzel w einer Zahl y auch w mal einer dieser drei Zahlen Kubikwurzel; es gibt also (für $y \neq 0$) drei verschiedene Kubikwurzeln, und somit hat obige Formel für z gleich *neun* mögliche Interpretationen.

Daraus können wir die drei richtigen herausfiltern, wenn wir beachten, daß wir nicht nur das Produkt von u^3 und v^3 kennen, sondern auch das von u und v, nämlich $-p/3$. Damit ist der zweite Summand in der Formel für z eindeutig durch den ersten bestimmt, und es gibt nur die zu erwartenden drei Lösungen: Ist

$$u = \sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}}$$

irgendeiner der drei möglichen Werte der Wurzel, so ist

$$z = u - \frac{p}{3u}$$

eine Lösung der Gleichung $z^3 + pz + q = 0$ und

$$x = z - \frac{b}{3a}$$

eine Lösung der ursprünglichen Gleichung $ax^3 + bx^2 + cx + d = 0$.

Auch biquadratische Gleichungen lassen sich auflösen: Hier eliminiert man den kubischen Term von

$$ax^4 + bx^3 + cx^2 + dx + e = 0$$

durch die Substitution

$$z = x + \frac{b}{4a}\,;$$

dies führt auf eine Gleichung der Form $z^4 + pz^2 + qz + r = 0$. Für eine beliebige Zahl y folgt daraus für jede Nullstelle z dieser Gleichung die Beziehung

$$(z^2 + y)^2 = z^4 + 2yz^2 + y^2 = (2y - p)z^2 - qz + y^2 - r\,.$$

Falls rechts das Quadrat eines linearen Polynoms $sz + t$ steht, ist

$$(z^2 + y)^2 = (sz + t)^2 \Longrightarrow z = \pm\sqrt{-y \pm (sz + t)}\,,$$

wir können die Gleichung also auflösen.

Nun wird die rechte Seite $(2y - p)z^2 - qz + y^2 - r$ im allgemeinen kein Quadrat eines linearen Polynoms in z sein, wir können aber hoffen, daß es zumindest für gewisse spezielle Werte der bislang noch willkürlichen Konstante y eines ist.

Ein quadratisches Polynom $\alpha z^2 + \beta z + \gamma$ ist genau dann Quadrat eines linearen, wenn die beiden Nullstellen der quadratischen Gleichung $\alpha z^2 + \beta z + \gamma = 0$ übereinstimmen. Nach der obigen Lösungsformel für quadratische Gleichungen ist dies genau dann der Fall, wenn dort der Ausdruck unter der Wurzel verschwindet, d.h. wenn $\beta^2 - 4\alpha\gamma = 0$ ist. In unserem Fall muß also

$$q^2 - 4(2y - p)(y^2 - r) = -8y^3 + 4py^2 + 8ry + q^2 - 4pr$$

verschwinden. Dies ist eine kubische Gleichung für y; indem wir diese Gleichung lösen und eine der Lösungen für y einsetzen, erhalten wir die vier Lösungen der biquadratischen Gleichung.

Die erste Lösung einer kubischen Gleichung geht wohl aus SCIPIONE DEL FERRO (1465–1526) zurück, der von 1496 bis zu seinem Tod an der Universität Bologna lehrte. 1515 fand er eine Methode, um die Nullstellen von $x^3 + px = q$ für *positive* Werte von p und q zu bestimmen (Negative Zahlen waren damals in Europa noch nicht im Gebrauch). Er veröffentlichte diese jedoch nie, so daß NICCOLO FONTANA (1499–1557, oberes Bild), genannt TARTAGLIA (der Stotterer), dieselbe Methode 1535 noch einmal entdeckte und gleichzeitig auch noch eine Modifikation, um einen leicht verschiedenen Typ kubischer Gleichungen zu lösen. TARTAGLIA war mathematischer Autodidakt, war aber schnell als Fachmann anerkannt und konnte seinen Lebensunterhalt als Mathematiklehrer in Verona und Venedig verdienen.

Die Lösung allgemeiner kubischer Gleichungen geht auf den Mathematiker, Arzt und Naturforscher GIROLAMO CARDANO (1501–1576, unteres Bild) zurück, dem TARTAGLIA nach langem Drängen und unter dem Siegel der Verschwiegenheit seine Methode mitgeteilt hatte. LODOVICO FERRARI (1522–1565) kam 14-jährig als Diener zu CARDANO; als dieser merkte, daß FERRARI schreiben konnte, machte er ihn zu seinem Sekretär. 1540 fand er die Lösungsmethode für biquadratische Gleichungen; 1545 veröffentlichte CARDANO in seinem Buch *Ars magna* die Lösungsmethoden für kubische und biquadratische Gleichungen.

Nach der erfolgreichen Auflösung der kubischen und biquadratischen Gleichungen in der ersten Hälfte des sechzehnten Jahrhunderts beschäftigten sich natürlich viele Mathematiker mit dem nächsten Fall, der Gleichung fünften Grades. Hier gab es jedoch über 250 Jahre lang keinerlei Fortschritt, bis zu Beginn des neunzehnten Jahrhunderts ABEL glaubte, eine Lösung gefunden zu haben. Er entdeckte dann aber recht schnell seinen Fehler und bewies stattdessen 1824, daß es *unmöglich* ist, die Lösungen einer allgemeinen Gleichung fünften (oder höheren) Grades durch Grundrechenarten und Wurzeln auszudrücken.

Die Grundidee seines Beweises liegt in der Betrachtung von Symmetrien innerhalb der Lösungsmenge, ähnlich wie wir in einem späteren Abschnitt einige Differentialgleichungen durch Symmetriebetrachtungen lösen werden. Unmöglichkeitsbeweise sind allerdings deutlich aufwendiger als Lösungsversuche mit Hilfe von Symmetriebetrachtungen; daher kann über Einzelheiten des ABELschen Beweises hier nichts weiter gesagt werden. Interessenten finden ihn in fast jedem Algebralehrbuch im Kapitel über GALOIS-Theorie.

Der norwegische Mathematiker NILS HENRIK ABEL (1802–1829) ist trotz seinen frühen Todes (an Tuberkulose) Initiator vieler Entwicklungen der Mathematik des neunzehnten Jahrhunderts; Begriffe wie abelsche Gruppen, abelsche Integrale, abelsche Funktionen, abelsche Varietäten, die auch in der heutigen Mathematik noch allgegenwärtig sind, verdeutlichen seinen Einfluß. Zu seinem 200. Geburtstag stiftete die norwegische Regierung einen ABEL-Preises für Mathematik mit gleicher Ausstattung und Vergabebedingungen wie die Nobelpreise; erster Preisträger war 2003 JEAN-PIERRE SERRE (* 1926) vom Collège de France für seine Arbeiten über algebraische Geometrie, Topologie und Zahlentheorie.

Der ABELsche Satz besagt selbstverständlich nicht, daß Gleichungen höheren als vierten Grades *unlösbar* seien; er sagt nur, daß es *im allgemeinen* nicht möglich ist, die Lösungen durch Wurzelausdrücke in den Koeffizienten darzustellen: Für eine allgemeine Lösungsformel muß man also außer Wurzeln und Grundrechenarten noch weitere Funktionen zulassen. Beispielsweise fanden sowohl HERMITE als auch KRONECKER 1858 Lösungsformeln für Gleichungen fünften Grades mit sogenannten elliptischen Modulfunktionen; 1870 löste JORDAN damit Gleichungen beliebigen Grades.

Für die Berechnung von Eigenvektoren sind schon die Lösungen einer kubischen Gleichung nach CARDANOs Formel im allgemeinen zu kompliziert, als daß man ohne Computer damit rechnen könnte; dasselbe gilt erst recht für höhere Grade. Insbesondere sind die Formeln in vielen Fällen numerisch instabil, da annähernd gleich große Zahlen voneinander subtrahiert werden. Die Numerik geht daher aus gutem Grund anders vor, wenn sie Nullstellen von Polynomen berechnet.

d) Vielfachheiten von Eigenwerten

Ist x eine Nullstelle eines Polynoms $f(X)$, so kann $f(X)$ bekanntlich durch $(X - x)$ geteilt werden, und x heißt *r-fache Nullstelle* von $f(X)$, wenn $f(X)$ durch $(X - x)^r$ teilbar ist, nicht aber durch $(X - x)^{r+1}$.

Definition: Wir sagen, der Eigenwert λ von φ *bzw.* A habe die *algebraische Vielfachheit* r, wenn λ eine r-fache Nullstelle des charakteristischen Polynoms ist.

Im obigen Beispiel hatte also der Eigenwert Null die algebraische Vielfachheit zwei, die anderen beiden hatten algebraische Vielfachheit eins. Die Dimension des jeweiligen Eigenraums, die geometrische Vielfachheit also, war genauso groß, jedoch muß dies im allgemeinen nicht der Fall sein: Für die Matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

etwa hat das charakteristische Polynom

$$\det(A - \lambda E) = \begin{vmatrix} 1-\lambda & 1 \\ 0 & 1-\lambda \end{vmatrix} = (1-\lambda)^2$$

die doppelte Nullstelle eins, $\lambda = 1$ ist also ein Eigenwert mit algebraischer Vielfachheit zwei. Der zugehörige Eigenraum ist die Lösungsmenge des linearen Gleichungssystems

$$\begin{aligned} 0x_1 + 1x_2 &= 0 \\ 0x_1 + 0x_2 &= 0\,, \end{aligned}$$

also gerade die Menge aller Vektoren der Form $\binom{x}{0}$ und somit eindimensional. Die geometrische Vielfachheit des Eigenwerts eins ist daher nur eins.

Das Beispiel der Abbildung

$$\varphi\colon \mathbb{R}^2 \to \mathbb{R}^2; \quad \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x\cos\vartheta - y\sin\vartheta \\ y\cos\vartheta + x\sin\vartheta \end{pmatrix}$$

mit Abbildungsmatrix

$$A = \begin{pmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

zeigt, daß es überhaupt keine Eigenwerte geben muß, denn hier ist das charakteristische Polynom gleich

$$\begin{vmatrix} \cos\vartheta - \lambda & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta - \lambda \end{vmatrix} = (\cos\vartheta - \lambda)^2 + \sin^2\vartheta\,.$$

Abgesehen vom Fall $\sin\vartheta = 0$, wenn A gleich der positiven oder negativen Einheitsmatrix ist, hat dieses Polynom keine reelle Nullstelle, da es nur positive Werte annimmt. Es hat aber natürlich die beiden komplexen Nullstellen

$$\lambda_{1/2} = \cos\vartheta \pm i\sin\vartheta = e^{\pm i\vartheta}\,;$$

fassen wir φ als Abbildung von $\mathbb{C}^2$ nach $\mathbb{C}^2$ auf, gibt es also zwei Eigenwerte. Beide haben die algebraische und geometrische Vielfachheit eins; zugehörige Eigenvektoren sind etwa $\binom{1}{i}$ und $\binom{1}{-i}$. Wählen wir diese beiden Vektoren als Basis, so wird die Abbildungsmatrix von φ bezüglich dieser neuen Basis zur Diagonalmatrix

$$\begin{pmatrix} e^{i\vartheta} & 0 \\ 0 & e^{-i\vartheta} \end{pmatrix}\,.$$

Allgemein gilt für die algebraischen und geometrischen Vielfachheiten von Eigenvektoren

Satz: *a)* Die geometrische Vielfachheit eines Eigenwerts ist stets kleiner oder gleich der algebraischen Vielfachheit.
b) Die Summe der algebraischen Vielfachheiten der verschiedenen Eigenwerte einer linearen Abbildung ist kleiner oder gleich der Dimension des Vektorraums.

Beweis: a) Der Eigenwert λ der $n \times n$-Matrix A habe die geometrische Vielfachheit r, d.h. der zugehörige Eigenraum habe die Dimension r. Wir wählen eine Basis $\vec{b}_1, \ldots, \vec{b}_r$ dieses Eigenraums und ergänzen sie zu einer Basis des gesamten Vektorraums; bezüglich dieser Basis sei C die Abbildungsmatrix der linearen Abbildung

$$\varphi\colon \begin{cases} k^n \to k^n \\ \vec{v} \mapsto A\vec{v} \end{cases}\,.$$

Da $\vec{b}_1, \ldots, \vec{b}_r$ Eigenvektoren zum Eigenwert λ sind, ist $\varphi(\vec{b}_i) = \lambda\vec{b}_i$. In den ersten r Spalten von C steht also jeweils in der Diagonalen das Element λ und ansonsten überall die Null. A hat somit die Form

$$A = \begin{pmatrix} \lambda & 0 & \ldots & 0 & * & \ldots & * \\ 0 & \lambda & \ldots & 0 & * & \ldots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda & * & \ldots & * \\ 0 & 0 & \ldots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & M & \\ 0 & 0 & \ldots & 0 & & & \end{pmatrix},$$

wobei uns weder die mit $*$ bezeichneten Körperelemente noch die $(n-r) \times (n-r)$-Matrix M weiter zu interessieren brauchen.

Für $C - xE$ gilt dasselbe, nur daß jetzt $\lambda - x$ in der Diagonalen steht, d.h. diese Matrix hat die Form

$$\begin{pmatrix} \lambda - x & 0 & \ldots & 0 & * & \ldots & * \\ 0 & \lambda - x & \ldots & 0 & * & \ldots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda - x & * & \ldots & * \\ 0 & 0 & \ldots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & M - xE_{n-r} & \\ 0 & 0 & \ldots & 0 & & & \end{pmatrix},$$

wobei E_{n-r} die $(n-r) \times (n-r)$-Einheitsmatrix bezeichnet.

Zur Berechnung ihrer Determinanten verwenden wir den LAPLACEschen Entwicklungssatz: Da in der ersten Zeile (oder Spalte) nur an der ersten Stelle ein von Null verschiedener Eintrag steht, ist diese Determinante gleich $(\lambda - x)$ mal der Determinante jener Matrix, die durch Streichen der ersten Zeile und Spalte entsteht. Falls $r > 1$ ist, hat diese neue Matrix dieselbe Form, wir können den LAPLACEschen Entwicklungssatz also noch einmal anwenden *usw.;* wir erhalten schließlich

$$\det(C - xE) = (\lambda - x)^r \det(M - xE_{n-r}).$$

Somit ist $\det(C - xE)$ durch $(x - \lambda)^r$ teilbar.

Was uns wirklich interessiert, ist aber nicht $\det(C - xE)$, sondern $\det(A - xE)$. Ist B die Matrix des Basiswechsels von der Standardbasis des k^n auf die Basis $\{\vec{b}_1, \ldots, \vec{b}_n\}$, jene Matrix also, deren Spaltenvektoren die $\vec{b}_i$ sind, so ist $C = B^{-1}AB$ und

$$\begin{aligned} \det(C - xE) &= \det(B^{-1}AB - xE) = \det(B^{-1}AB - xB^{-1}EB) \\ &= \det\big(B(A - xE)B^{-1}\big) = \det B \det(A - xE)(\det B)^{-1} \\ &= \det(A - xE). \end{aligned}$$

A und C haben also dasselbe charakteristische Polynom, und somit ist auch das charakteristische Polynom von A durch $(x - \lambda)^r$ teilbar. Die algebraische Vielfachheit von λ ist daher mindestens r.

Unabhängig von diesem Ergebnis wollen wir noch festhalten, daß nach der gerade durchgeführten Rechnung für eine beliebige Matrix A und eine invertierbare Matrix B die beiden Matrizen A und BAB^{-1} dasselbe charakteristische Polynom haben; insbesondere haben also die Abbildungsmatrizen einer linearen Abbildung zu verschiedenen Basen dasselbe charakteristische Polynom.

b) Sind $\lambda_1, \ldots, \lambda_\ell$ die verschiedenen Eigenwerte von φ und sind $r_1, \ldots, r_\ell$ ihre algebraischen Vielfachheiten, so ist das charakteristische Polynom $\det(A - xE)$ teilbar durch

$$(x - \lambda_1)^{r_1} \cdots (x - \lambda_\ell)^{r_\ell}.$$

Dies ist ein Polynom vom Grad $r_1 + \cdots + r_\ell$, wohingegen das charakteristische Polynom Grad n hat; daher ist

$$r_1 + \cdots + r_\ell \leq n,$$

denn der Grad eines Teilers kann nicht größer sein als der des Polynoms selbst.

■

Zum Abschluß dieses Abschnitts sei noch ein Kriterium angegeben, wann es für eine lineare Abbildung φ eine Basis aus Eigenwerten gibt, wann also die Abbildungsmatrix bezüglich einer geeigneten Basis Diagonalgestalt hat:

Satz: Zur linearen Abbildung $\varphi: V \to V$ eines n-dimensionalen Vektorraums gibt es genau dann eine Basis aus Eigenvektoren von φ, wenn

1.) das charakteristische Polynom von φ als Produkt von Linearfaktoren geschrieben werden kann

2.) die geometrische Vielfachheit eines jeden Eigenwerts gleich der algebraischen ist.

Beweis: Zunächst sei $\varphi: V \to V$ eine lineare Abbildung derart, daß V eine Basis $\{\vec{b}_1, \ldots, \vec{b}_n\}$ aus Eigenvektoren von φ habe. Wir müssen zeigen, daß *1.)* und *2.)* erfüllt sind.

Da die Basisvektoren $\vec{b}_i$ Eigenvektoren sind, gibt es zu jedem $\vec{b}_i$ ein Körperelement λ_i, so daß $\varphi(\vec{b}_i) = \lambda \vec{b}_i$ ist; bezüglich dieser Basis hat die Abbildungsmatrix A von φ daher Diagonalgestalt, und das charakteristische Polynom

$$\det(A - \lambda E) = \begin{vmatrix} \lambda_1 - \lambda & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n - \lambda \end{vmatrix} = \prod_{i=1}^{n} (\lambda_i - \lambda)$$

zerfällt in der Tat in Linearfaktoren. Die algebraische Vielfachheit des Eigenwerts λ_i ist gleich der Anzahl jeder Indizes $j \in \{1, \ldots, n\}$, für die $\lambda_j = \lambda_i$ ist; dies ist auch die geometrische Vielfachheit, denn der Eigenraum wird aufgespannt von den Vektoren $\vec{b}_j$ zu diesen j. Also sind *1)* und *2)* erfüllt.

Umgekehrt erfülle die Abbildung φ die Bedingungen *1)* und *2);* wir müssen zeigen, daß es eine Basis aus Eigenvektoren von φ gibt.

Wegen *1)* läßt sich das charakteristische Polynom in der Form

$$(\lambda_1 - \lambda)^{r_1} \cdot \cdots \cdot (\lambda_s - \lambda)^{r_s}$$

schreiben, wobei wir annehmen können, daß die λ_i paarweise verschieden sind. Dann ist r_i die algebraische Vielfachheit von λ_i. Da das charakteristische Polynom den Grad n hat, folgt, daß

$$r_1 + \cdots + r_s = n$$

ist. Außerdem gibt es wegen *2.)* zu jedem λ_i einen r_i-dimensionalen Eigenraum, also r_i linear unabhängige Eigenvektoren. Da Eigenvektoren zu verschiedenen Eigenwerten nach dem Lemma vom Anfang dieses Abschnitts stets linear unabhängig sind, ist auch das System all dieser Eigenvektoren linear unabhängig und somit eine Basis, denn es besteht aus $n = \dim V$ Vektoren. Damit ist eine Basis aus Eigenvektoren von φ gefunden.

■

e) Eigenwerte symmetrischer und Hermitescher Matrizen

Wie wir im letzten Paragraphen gesehen haben, kann die geometrische Vielfachheit eines Eigenwerts kleiner sein als die algebraische, und im Falle einer reellen Matrix müssen nicht auch die Eigenwerte reell sein. In diesem Abschnitt wollen wir sehen, daß solche Dinge bei symmetrischen (und auch den noch zu definierenden HERMITEschen) Matrizen nicht möglich sind.

Symmetrische und HERMITEsche Matrizen hängen eng mit (HERMITEschen) Skalarprodukten zusammen: Für zwei Vektoren

$$\vec{v} = \sum_{i=1}^{n} v_i \vec{b}_i \quad \text{und} \quad \vec{w} = \sum_{i=1}^{n} w_i \vec{b}_i$$

aus einem endlichdimensionalen EUKLIDischen Vektorraum V mit Basis $\{\vec{b}_1, \ldots, \vec{b}_n\}$ ist wegen der Linearität des Skalarprodukts in beiden Argumenten

$$\vec{v} \cdot \vec{w} = \left(\sum_{i=1}^{n} v_i \vec{b}_i \right) \cdot \left(\sum_{j=1}^{n} w_j \vec{b}_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} v_i w_j \, \vec{b}_i \cdot \vec{b}_j \, .$$

Setzen wir

$$c_{ij} \underset{\text{def}}{=} \vec{b}_i \cdot \vec{b}_j \, ,$$

so ist wegen der Symmetrie des Skalarprodukts $c_{ij} = c_{ji}$, wir haben also eine symmetrische $n \times n$-Matrix C.

Die Matrix C legt das Skalarprodukt eindeutig fest, denn für zwei beliebige Vektoren $\vec{v}, \vec{w}$ wie oben ist

$$\vec{v} \cdot \vec{w} = \sum_{i=1}^{n} \sum_{j=1}^{n} v_i w_j \cdot c_{ij} \, .$$

Diese Formel definiert umgekehrt auch für jede symmetrische Matrix $C \in \mathbb{R}^{n\times n}$ eine bilineare Abbildung $V \times V \to \mathbb{R}$, allerdings muß diese nicht positiv definit und damit kein Skalarprodukt sein.

Ist V ein HERMITEscher Vektorraum, wieder mit Basis $\{\vec{b}_1, \ldots, \vec{b}_n\}$, so ist jetzt für zwei Vektoren

$$\vec{v} = \sum_{i=1}^{n} v_i \vec{b}_i \quad \text{und} \quad \vec{w} = \sum_{i=1}^{n} w_i \vec{b}_i \quad \text{mit} \quad v_i, w_i \in \mathbb{C}$$

$$\vec{v} \cdot \vec{w} = \left(\sum_{i=1}^{n} v_i \vec{b}_i\right) \cdot \left(\sum_{j=1}^{n} w_j \vec{b}_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i \overline{w}_j \, \vec{b}_i \cdot \vec{b}_j \, .$$

Setzen wir auch hier wieder

$$c_{ij} \underset{\text{def}}{=} \vec{b}_i \cdot \vec{b}_j \, ,$$

so ist nun $c_{ij} = \overline{c_{ji}}$. Matrizen mit dieser Eigenschaft wollen wir als HERMITEsch bezeichnen.

Um dies etwas kompakter ausdrücken zu können, definieren wir

Definition: *a)* Für eine Matrix $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ bezeichnen wir die Matrix $\overline{A} = \left(\overline{a_{ij}}\right)$ als die zu A *konjugiert komplexe Matrix.*
b) $A \in \mathbb{C}^{n\times n}$ heißt HERMITEsch, falls ${}^tA = \overline{A}$ ist.
c) Zu einem Vektor $\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ heißt $\overline{\vec{v}} = \begin{pmatrix} \overline{v_1} \\ \vdots \\ \overline{v_n} \end{pmatrix}$ der *konjugiert komplexe* Vektor.

(Letztere Schreibweise sieht zwar grausam aus, läßt sich aber nicht vermeiden, wenn man Vektoren mit Pfeilen kennzeichnet. Alternativen wie der Fettdruck von Vektoren funktionieren weder an der Tafel noch in einer Mitschrift, und für Frakturbuchstaben wie $\mathfrak{u}, \mathfrak{v}, \mathfrak{w}$ können sich leider nur wenige Studenten begeistern.)

Schließlich wollen wir Vektoren hier mit $1 \times n$-Matrizen identifizieren; insbesondere rechnen wir mit dem „transponierten Vektor“

$${}^t\vec{v} = (v_1, \ldots, v_n) \, .$$

Mit dieser Bezeichnung kann das Standardskalarprodukt zweier Vektoren $\vec{v}, \vec{w} \in \mathbb{R}^n$ als Matrixprodukt ${}^t\vec{v}\vec{w}$ geschrieben werden; das Standard-HERMITEsche Produkt in $\mathbb{C}^n$ ist entsprechend ${}^t\vec{v}\,\overline{\vec{w}}$.

Da die komplexe Konjugation auf $\mathbb{R}$ keine Wirkung hat, ist eine HERMITEsche Matrix mit reellen Einträgen einfach eine symmetrische Matrix; wir können uns im folgenden bei den Beweisen daher auf HERMITEsche Matrizen beschränken und erhalten trotzdem Ergebnisse, die auch für reelle symmetrische Matrizen gelten.

Das Hauptziel dieses Abschnitts ist

Satz: A sei eine symmetrische reelle oder HERMITEsche (komplexe) Matrix.
a) Dann sind alle Eigenwerte von A reell.
b) Eigenvektoren zu verschiedenen Eigenwerten sind orthogonal bezüglich des Standard- *bzw.* HERMITEschen Skalarprodukts.
c) Für jeden Eigenwert von A ist die geometrische Vielfachheit gleich der algebraischen Vielfachheit.
d) $\mathbb{R}^n$ *bzw.* $\mathbb{C}^n$ hat eine Orthonormalbasis aus Eigenvektoren von A.

Beweis: a) Ist $\lambda \in \mathbb{C}$ ein Eigenwert von A, so gibt es nach Definition einen Vektor $\vec{v} \neq \vec{0}$, so daß $A\vec{v} = \lambda\vec{v}$ ist. Da die komplexe Konjugation mit sämtlichen Grundrechenarten vertauschbar ist, folgt, daß

$$\overline{A}\,\overline{\vec{v}} = \overline{\lambda}\,\overline{\vec{v}}\,, \quad \text{d.h.} \quad {}^t\vec{v}\overline{A}\,\overline{\vec{v}} = {}^t\vec{v}\overline{\lambda}\,\overline{\vec{v}} = \overline{\lambda}\,{}^t\vec{v}\,\overline{\vec{v}}\,.$$

Bislang gilt alles noch für beliebige $n \times n$-Matrizen; um die Symmetrie *bzw.* HERMITE-Eigenschaft von A ins Spiel zu bringen, betrachten wir den Vektor ${}^t(A\vec{v}) = {}^t\vec{v}\,{}^tA$. Da nach Voraussetzung ${}^tA = \overline{A}$ ist, können wir die rechte Seite der Gleichung auch als ${}^t\vec{v}\,\overline{A}$ schreiben, und die linke Seite als ${}^t(\lambda v) = \lambda\,{}^t\vec{v}$, da $\vec{v}$ Eigenvektor von A ist. Somit können wir die Zahl ${}^t\vec{v}\overline{A}\,\overline{\vec{v}}$ auch schreiben als

$${}^t\vec{v}\overline{A}\,\overline{\vec{v}} = \left({}^t\vec{v}\,\overline{A}\right)\overline{\vec{v}} = \lambda\,{}^t\vec{v}\,\overline{\vec{v}}\,.$$

Somit haben wir die beiden Darstellungen

$${}^t\vec{v}\overline{A}\,\overline{\vec{v}} = \lambda\,{}^t\vec{v}\,\overline{\vec{v}} \quad \text{und} \quad {}^t\vec{v}\overline{A}\,\overline{\vec{v}} = \overline{\lambda}\,{}^t\vec{v}\,\overline{\vec{v}}\,,$$

die nur dann beide richtig sein können, wenn $\lambda = \overline{\lambda}$ und somit reell ist; denn ${}^t\vec{v}\,\overline{\vec{v}}$ kann wegen der Definitheit HERMITEscher Skalarprodukte für einen Vektor $\vec{v} \neq \vec{0}$ nicht verschwinden.

b) $\vec{v}$ sei Eigenvektor zum Eigenwert λ, und $\vec{w}$ sei Eigenvektor zum davon verschiedenen Eigenwert μ, d.h.

$$A\vec{v} = \lambda\vec{v} \quad \text{und} \quad A\vec{w} = \mu\vec{w} \quad \text{und} \quad \lambda \neq \mu\,.$$

Dann ist

$$\lambda\,{}^t\vec{v}\overline{\vec{w}} = {}^t(\lambda\vec{v})\overline{\vec{w}} = {}^t(A\vec{v})\overline{\vec{w}} = {}^t\vec{v}\,{}^tA\overline{\vec{w}} = {}^t\vec{v}\overline{A}\,\overline{\vec{w}} = {}^t\vec{v}\overline{A\vec{w}} = {}^t\vec{v}\overline{\mu\vec{w}} = \overline{\mu}\,{}^t\vec{v}\overline{\vec{w}}\,.$$

Wie wir schon wissen, sind alle Eigenwerte reell, d.h. $\overline{\mu} = \mu \neq \lambda$. Die obige Gleichungskette kann daher nur richtig sein, wenn ${}^t\vec{v}\,\overline{\vec{w}}$ verschwindet, d.h. wenn $\vec{v}$ und $\vec{w}$ orthogonal sind.

Beim Beweis von *c)* gehen wir im wesentlichen genauso vor wie im vorigen Abschnitt, als wir zeigten, daß die geometrische Vielfachheit eines Eigenwerts stets kleiner oder gleich der algebraischen ist; die zusätzliche Annahme über die Matrix A wird zeigen, daß hier die beiden Vielfachheiten sogar gleich sind.

λ sei also ein Eigenwert von A mit geometrischer Vielfachheit r, d.h. der zugehörige Eigenraum habe die Dimension r. Wir wählen eine Basis $\{\vec{b}_1, \ldots, \vec{b}_r\}$ davon und ergänzen sie zu einer Basis $\mathcal{B} = \{\vec{b}_1, \ldots, \vec{b}_n\}$ des gesamten Vektorraums $V = \mathbb{R}^n$ oder $\mathbb{C}^n$. Indem wir nötigenfalls das GRAM-SCHMIDTsche Orthogonalisierungsverfahren anwenden und anschließend die Längen aller Vektoren auf eins normieren, können wir annehmen, daß es sich dabei um eine Orthonormalbasis handelt.

Nun betrachten wir die lineare Abbildung

$$\varphi\colon V \to V; \quad \vec{v} \mapsto A\vec{v}\,.$$

Bezüglich der Standardbasis hat sie A als Abbildungsmatrix; für uns interessanter ist aber die Abbildungsmatrix C bezüglich der neuen Basis $\mathcal{B}$. Dazu sei B die Matrix mit Spaltenvektoren $\vec{b}_i$; da der Eintrag an der Stelle (i,j) eines Matrixprodukts das (Standard-)Skalarprodukt des i-ten Zeilenvektors des ersten Faktors mit dem j-ten Spaltenvektor des zweiten Faktors ist, steht an der Stelle (i,j) der Matrix ${}^tB\overline{B}$

das (Standard) HERMITEsche Produkt der Vektoren $\vec{b}_i$ und $\vec{b}_j$. Da $\mathcal{B}$ als Orthonormalbasis gewählt wurde, ist daher

$${}^tB\overline{B} = E \quad \text{und damit} \quad {}^tB = \overline{B}^{\,-1} = \overline{B^{-1}}\,.$$

Aus dieser Formel folgt, daß mit A auch C eine HERMITEsche Matrix ist, denn

$${}^tC = {}^t(B^{-1}AB) = {}^tB\,{}^tA\,{}^tB^{-1} = \overline{B^{-1}}\,\overline{A}\,\overline{B} = \overline{B^{-1}AB} = \overline{C}\,.$$

Die ersten r Basisvektoren $\vec{b}_i$ sind Eigenvektoren von A zum Eigenwert λ; für $i \leq r$ ist daher $\varphi(\vec{b}_i) = \lambda\vec{b}_i$, d.h. in der i-ten Spalte von C steht an der i-ten Stelle die reelle Zahl λ und ansonsten überall die Null, genau wie auch im vorigen Abschnitt. Im Gegensatz zu dort haben wir nun aber eine HERMITEsche Matrix; da in der i-ten Spalte abgesehen von λ auf der Hauptdiagonalen nur Nullen stehen, muß daher dasselbe auch für die i-te Zeile gelten; die Matrix C hat also die Form

$$C = \begin{pmatrix} \lambda & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \lambda & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & M & \\ 0 & 0 & \ldots & 0 & & & \end{pmatrix},$$

wobei M eine $(n-r) \times (n-r)$-Matrix ist, die uns nicht weiter zu interessieren braucht. Damit hat $C - xE$ die Form

$$\begin{pmatrix} \lambda - x & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \lambda - x & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda - x & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & M - xE_{n-r} & \\ 0 & 0 & \ldots & 0 & & & \end{pmatrix},$$

wobei E_{n-r} die $(n-r) \times (n-r)$-Einheitsmatrix bezeichnet.

Wie wir uns schon im vorigen Abschnitt überlegten beim Beweis, daß die geometrische Vielfachheit eines Eigenwerts immer kleiner oder gleich der algebraischen ist, haben A und C dasselbe charakterische Polynom; da wir die Matrix C besser kennen, rechnen wir mit ihr.

Wie in Abschnitt *d)* folgt auf Grund der obigen Form der Matrix $C - xE$ aus dem LAPLACEschen Entwicklungssatz, daß

$$\det(A - xE) = \det(C - xE) = (\lambda - x)^r \det(M - xE_{n-r})$$

ist, wobei E_{n-r} die $(n-r) \times (n-r)$-Einheitsmatrix bezeichnet. Wir müssen zeigen, daß die algebraische Vielfachheit von λ *genau* gleich r ist, daß also λ keine Nullstelle von $\det(M - xE_{n-r})$ sein kann.

Wäre λ Nullstelle von $\det(M - xE_{n-r})$, so hätte M den Eigenwert λ, es gäbe also einen $(n-r)$-dimensionalen Eigenvektor $\vec{w}$ von M. Wegen der speziellen Form der Matrix C ist für jeden Eigenvektor

$$\vec{w} = \begin{pmatrix} w_{r-1} \\ \vdots \\ w_n \end{pmatrix} \quad \text{von } M \text{ der Vektor} \quad \vec{v} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ w_{r-1} \\ \vdots \\ w_n \end{pmatrix}$$

ein Eigenvektor von C und damit von A – Eigenvektoren hängen schließlich nur von der linearen Abbildung ab, nicht von einer speziellen Abbildungsmatrix. Dies widerspricht aber der Voraussetzung, daß der Eigenraum zum Eigenwert λ von $\vec{b}_1, \ldots, \vec{b}_r$ erzeugt wird, denn $\vec{v}$ ist linear unabhängig von diesen $\vec{b}_i$.

Also hat λ die algebraische Vielfachheit r, und *c)* ist gezeigt.

d) ist nun eine einfache Folgerung aus den übrigen Aussagen und dem sogenannten *Fundamentalsatz der Algebra,* wonach jedes reelle oder komplexe Polynom über den komplexen Zahlen in Linearfaktoren zerfällt:

Wir wissen, daß die Summe der algebraischen Vielfachheiten aller Eigenwerte gleich der Dimension n des Vektorraums ist und daß alle Eigenwerte reell sind; da die algebraischen gleich den geometrischen

Vielfachheiten sind, gibt es also n Eigenvektoren, die eine Basis von V bilden.

Für jeden einzelnen Eigenraum können wir die Eigenvektoren nach GRAM-SCHMIDT so wählen, daß sie eine Orthonormalbasis bilden; da Eigenvektoren zu verschiedenen Eigenwerten stets orthogonal sind, ist die Vereinigungsmenge dieser Basen Orthonormalbasis von V. ■

f) Hauptvektoren und und die Jordan-Zerlegung

Falls die lineare Abbildung $\varphi: V \to V$ Eigenwerte hat, deren geometrische Vielfachheit kleiner als die algebraische ist, haben wir keine Chance auf eine Basis, bezüglich derer die Abbildungsmatrix von φ Diagonalgestalt hat: Die Elemente einer solchen Basis wären allesamt Eigenvektoren, und bei zu kleiner geometrischer Vielfachheit gibt es nicht genügend linear unabhängige Eigenvektoren. Außerdem gibt es offensichtlich keine Chance auf eine Diagonalgestalt, wenn das charakteristische Polynom von φ nicht in Linearfaktoren zerfällt, denn dann ist schon die Summe der *algebraischen* Vielfachheiten der Eigenwerte kleiner als die Dimension von V.

Das zweite dieser Probleme konnten wir zumindest beim Beispiel der Matrix

$$\begin{pmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{pmatrix}$$

dadurch lösen, daß wir zu einem größeren Körper übergegangen sind, nämlich von den reellen zu den komplexen Zahlen.

Tatsächlich läßt es sich *immer* dadurch lösen, daß man zu einem größeren Körper übergeht: Nach dem *Fundamentalsatz der Algebra,* zerfällt jedes Polynom mit komplexen (also insbesondere auch mit reellen) Koeffizienten über den komplexen Zahlen in Linearfaktoren. Für andere Körper als die reellen oder komplexen Zahlen zeigt die Algebra, daß es zu jedem Polynom über einem Körper stets einen Erweiterungskörper gibt, der als Vektorraum über dem Ausgangskörper endliche Dimension hat, so daß das gegebene Polynom dort in Linearfaktoren zerfällt. Mit Methoden, die im allgemeinen nicht konstruktiv sind, folgt sogar,

daß es stets einen (im allgemeinen unendlichdimensionalen) Erweiterungskörper gibt, über dem *jedes* Polynom in Linearfaktoren zerfällt, den sogenannte *algebraischen Abschluß* des Ausgangskörpers. Einzelheiten findet man in jedem Lehrbuch der Algebra.

Somit können wir das Problem, daß das charakteristische Polynom eventuell nicht genügend viele Nullstellen hat, im wesentlichen ignorieren. Ernster ist das Problem mit Eigenwerten, deren geometrische Vielfachheit kleiner ist als die algebraische. Damit wollen wir uns in diesem Abschnitt beschäftigen.

Die Lösung wird darin bestehen, daß wir solchen Eigenwerten Räume zuordnen, die größer sind als die Eigenräume, aber immer noch eine gut an die Abbildung angepaßte Basis haben. Insbesondere sollen sie, genau wie die Eigenräume, *invariant* sein unter der betrachteten Abbildung:

Definition: $\varphi\colon V \to V$ sei eine lineare Abbildung. Ein Untervektorraum $U \leq V$ heißt invariant unter φ oder kurz φ-invariant, wenn $\varphi(U) \leq U$ ist.

Die φ-Invarianz der Eigenräume im Sinne dieser Definition ist klar, denn auf einem Eigenraum ist φ einfach die Multiplikation mit dem zugehörigen Eigenwert.

Für das folgende wollen wir der Einfachheit halber annehmen, daß V endliche Dimension habe. Dann ist erst recht jeder φ-invariante Unterraum U endlichdimensional, wir können also eine endliche Basis $\{\vec{b}_1, \dots, \vec{b}_r\}$ von U finden und diese ergänzen zu einer Basis $\{\vec{b}_1, \dots, \vec{b}_n\}$ von V. Da $\varphi(U) \leq U$ ist, liegen die Bilder der ersten r Basisvektoren wieder in U, d.h. die Abbildungsmatrix bezüglich dieser Basis hat die Form

$$\begin{pmatrix} \boxed{A} & \boxed{C} \\ \mathbf{0} & \boxed{B} \end{pmatrix}$$

mit einer $r \times r$-Matrix A, der Abbildungsmatrix von $\varphi_{|U}\colon U \to U$, einer $(n-r)\times(n-r)$-Matrix B und einer $(n-r)\times r$-Matrix C. Die fette Null soll hier, wie auch in den noch folgenden Matrizen, stets eine Nullmatrix der jeweils korrekten Größe bezeichnen.

Noch besser wird die Situation, wenn U ein φ-invariantes Komplement hat, wenn es also einen weiteren φ-invarianten Untervektorraum W gibt, so daß $V = U + W$ ist und $U \cap W = \{\vec{0}\}$. (Wir sagen dann, $V = U \oplus W$ sei die *direkte Summe* von U und W.) In diesem Fall können wir für $\vec{b}_{r+1}$ bis $\vec{b}_n$ die Vektoren einer Basis von W nehmen, und da nun auch W auf sich selbst abgebildet wird, haben wir eine Abbildungsmatrix der Form

$$\begin{pmatrix} \boxed{A} & \mathbf{0} \\ \mathbf{0} & \boxed{B} \end{pmatrix}.$$

Allgemein sagen wir für s Untervektorräume $U_1, \dots, U_s$ von V, daß V die direkte Summe

$$V = U_1 \oplus \cdots \oplus U_s = \bigoplus_{i=1}^{s} U_i$$

sei, wenn

$$V = U_1 + \cdots + U_s = \sum_{i=1}^{s} U_i \quad \text{und} \quad U_i \cap \sum_{j \neq i} U_j = \{\vec{0}\}$$

ist. Falls hierbei die U_i allesamt φ-invariant sind, können wir ihre Basen aneinandersetzen und erhalten eine Basis, bezüglich derer die Abbildungsmatrix die Gestalt

$$\begin{pmatrix} \boxed{A_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boxed{A_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boxed{A_s} \end{pmatrix}$$

hat, wobei die A_i die Abbildungsmatrizen der Einschränkungen $\varphi_{|U_i}$ zu Abbildungen von U_i nach U_i sind.

Kandidaten für Untervektorräume U_i liefern die Haupträume:

Definition: *a)* Ein Vektor $\vec{v} \in V$ heißt *Hauptvektor* von φ zum Eigenwert λ, wenn es ein $\ell \in \mathbb{N}_0$ gibt, so daß $(\varphi - \lambda\,\mathrm{id})^{\ell}(\vec{v}) = \vec{0}$ ist. Falls $(\varphi - \lambda\,\mathrm{id})^{\ell-1} \neq \vec{0}$ ist, bezeichnen wir ℓ als die *Stufe* des Hauptvektors.

b) Die Menge aller Hauptvektoren von φ zum Eigenwert λ heißt *Hauptraum* zu λ und wird mit H_λ bezeichnet.

Insbesondere sind die Hauptvektoren der Stufe eins genau die Eigenvektoren zum Eigenwert λ: Der Nullvektor ist nämlich kein Hauptvektor erster Stufe, da er bereits von $(\varphi - \lambda\,\mathrm{id})^0 = \mathrm{id}$ auf $\vec{0}$ abgebildet wird.

Es ist klar, daß die Hauptvektoren einen Untervektorraum bilden, denn mit $(\varphi - \lambda\,\mathrm{id})$ sind auch dessen Schachtelungen

$$(\varphi - \lambda\,\mathrm{id})^\ell = (\varphi - \lambda\,\mathrm{id}) \circ \cdots \circ (\varphi - \lambda\,\mathrm{id})$$

lineare Abbildungen, und die Hauptvektoren der Stufe höchstens ℓ sind gerade die Elemente des Kerns dieser Abbildung. Da wir von einem endlichdimensionalen Vektorraum V ausgehen, kann die Folge dieser Kerne nicht unbeschränkt wachsen, es gibt also ein maximales ℓ, das als Stufe eines Hauptvektors auftreten kann. Mit diesem ℓ ist der Hauptraum H_λ gerade der Kern von $(\varphi - \lambda\,\mathrm{id})^\ell$.

Der Nutzen der Haupträume ergibt sich aus folgendem

Lemma: H_λ ist ein φ-invarianter Unterraum von V. Bezeichnet ℓ die größte Stufe eines Hauptvektors aus H_λ, so ist $\mathrm{Bild}(\varphi - \lambda\,\mathrm{id})^\ell$ ein φ-invariantes Komplement.

Beweis: Beginnen wir mit der Invarianz von H_λ unter φ.

Ist $\vec{v}$ ein Hauptvektor der Stufe j, so ist

$$\begin{aligned}(\varphi - \lambda\,\mathrm{id})^j(\vec{v}) &= (\varphi - \lambda\,\mathrm{id})^{j-1}\big((\varphi - \lambda\,\mathrm{id})(\vec{v})\big)\\ &= (\varphi - \lambda\,\mathrm{id})^{j-1}\big(\varphi(\vec{v}) - \lambda\vec{v}\big) = \vec{0}\,,\end{aligned}$$

$\varphi(\vec{v}) - \lambda\vec{v}$ ist also ein Hauptvektor der Stufe höchstens $j-1$ und somit insbesondere ein Element von H_λ. Da mit $\vec{v}$ auch $\lambda\vec{v}$ in H_λ liegt, ist damit auch $\varphi(\vec{v}) = \big(\varphi(\vec{v}) - \lambda\vec{v}\big) + \lambda\vec{v} \in H_\lambda$ ein Hauptvektor.

Die Invarianz von $\mathrm{Bild}(\varphi - \lambda\,\mathrm{id})^\ell$ folgt genauso: Für $\vec{v} = (\varphi - \lambda\,\mathrm{id})^\ell(\vec{w})$ ist

$$\varphi(\vec{v}) - \lambda\vec{v} = (\varphi - \lambda\,\mathrm{id})(\vec{v}) = (\varphi - \lambda\,\mathrm{id})^{\ell+1}(\vec{w}) = (\varphi - \lambda\,\mathrm{id})^\ell(\varphi(\vec{w}) - \lambda\vec{w})$$

wieder ein Element des Bilds und damit auch $\varphi(\vec{v})$ selbst.

Als nächstes müssen wir zeigen, daß der Durchschnitt der beiden Räume nur aus dem Nullvektor besteht. Dazu sei $\vec{v}$ ein Vektor aus diesem Durchschnitt. Dann liegt $\vec{v}$ sowohl im Kern als auch im Bild der linearen Abbildung $(\varphi - \lambda\,\mathrm{id})^\ell$, es gibt also einen Vektor $\vec{w} \in V$ derart, daß $\vec{v} = (\varphi - \lambda\,\mathrm{id})^\ell(\vec{w})$ ist, und $(\varphi - \lambda\,\mathrm{id})^\ell(\vec{v}) = (\varphi - \lambda\,\mathrm{id})^{2\ell}(\vec{w}) = \vec{0}$. Damit liegt $\vec{w}$ aber im Hauptraum zu λ, d.h. $\vec{v} = (\varphi - \lambda\,\mathrm{id})^\ell(\vec{w}) = \vec{0}$.

Nach der Dimensionsformel ist

$$\dim \mathrm{Bild}(\varphi - \lambda\,\mathrm{id})^\ell = \dim V - \dim \mathrm{Kern}(\varphi - \lambda\,\mathrm{id})^\ell\,,$$

also ist

$$\dim \mathrm{Kern}(\varphi - \lambda\,\mathrm{id})^\ell + \dim \mathrm{Bild}(\varphi - \lambda\,\mathrm{id})^\ell = \dim V\,,$$

die beiden Untervektorräume erzeugen somit ganz V. ■

Die Zerlegung von V nach diesem Lemma heißt FITTING-Zerlegung.

Der deutsche Mathematiker HANS FITTING (1906–1938) beschäftigte sich vor allem mit der Untersuchung von Operatoren (und Operatorenringen). Trotz seines frühen Todes konnte er damit wesentliche Beiträge zur Algebra leisten, vor allem auch zur Erforschung der Struktur von Gruppen.

Das Schöne an der FITTING-Zerlegung ist, daß sie rekursiv fortgesetzt werden kann: Da $\mathrm{Bild}(\varphi - \lambda\,\mathrm{id})^\ell$ auch φ-invariant ist, können wir für die Einschränkung von φ auf diesen Unterraum einen Hauptraum zu einem anderen Eigenwert abspalten *usw.* Bevor wir uns das genauer überlegen, wollen wir uns aber zunächst eine gute Basis für den Hauptraum H_λ verschaffen.

Lemma: Der Hauptraum H_λ hat eine Basis, bezüglich derer die Abbildungsmatrix von φ eine obere Dreiecksmatrix ist. Alle Hauptdiagonaleinträge dieser Matrix sind gleich λ.

Beweis: Wir beginnen mit einer Basis $\{\vec{b}_1, \ldots, \vec{b}_{r_1}\}$ des Eigenraums zu λ und ergänzen diese zu einer Basis $\{\vec{b}_1, \ldots, \vec{b}_{r_2}\}$ des Raums aller Hauptvektoren der Stufe höchstens zwei und so weiter, bis eine Basis $\{\vec{b}_1, \ldots, \vec{b}_r\}$ des gesamten Hauptraums erreicht ist.

Der Vektor $\vec{b}_i$ sei Hauptvektor der Stufe ℓ; dann ist

$$\begin{aligned}(\varphi - \lambda\,\mathrm{id})^{\ell}(\vec{b}_i) &= (\varphi - \lambda\,\mathrm{id})^{\ell-1}\big((\varphi - \lambda\,\mathrm{id})(\vec{b}_i)\big)\\ &= (\varphi - \lambda\,\mathrm{id})^{\ell-1}\big(\varphi(\vec{b}_i) - \lambda\vec{b}_i\big) = \vec{0}\,,\end{aligned}$$

$\varphi(\vec{b}_i) - \lambda\vec{b}_i$ ist also ein Hauptvektor der Stufe höchstens $\ell - 1$. Nach Konstruktion der Basis ist $\varphi(\vec{b}_i) - \lambda\vec{b}_i$ daher eine Linearkombination von Basisvektoren $\vec{b}_j$ mit Indizes echt kleiner i, d.h.

$$\varphi(\vec{b}_i) = \lambda\vec{b}_i + \sum_{j=1}^{i-1} \alpha_{ij}\vec{b}_j\,.$$

Bezüglich der Basis $\{\vec{b}_1, \ldots, \vec{b}_r\}$ hat die Abbildungsmatrix daher in der Tat die gewünschte Form. ∎

Abspaltung immer weiterer Haupträume auch vom invarianten Komplement führt schließlich zum

Satz: Zu einer linearen Abbildung $\varphi\colon V \to V$ eines endlichdimensionalen k-Vektorraums V gibt es genau dann eine Basis von V, bezüglich derer die Abbildungsmatrix von φ eine Dreiecksmatrix ist, wenn das charakteristische Polynom von φ über k als Produkt von Linearfaktoren geschrieben werden kann. Alsdann kann die Basis so gewählt werden, daß die Abbildungsmatrix die Form

$$\begin{pmatrix} \boxed{A_1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boxed{A_2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boxed{A_s} \end{pmatrix}$$

hat mit Dreiecksmatrizen A_i, die auf der Hauptdiagonalen den i-ten Eigenwert λ_i stehen haben.

Beweis: Falls es zu φ eine Basis gibt, bezüglich derer die Abbildungsmatrix A von φ eine Dreiecksmatrix ist, ist bezüglich dieser Basis auch

$A - \lambda E$ eine Dreiecksmatrix. Da die Determinante einer Dreiecksmatrix gerade das Produkt der Diagonaleinträge ist, bekommen wir als charakteristisches Polynom $\det(A - \lambda E)$ ein Produkt von Linearfaktoren.

Falls umgekehrt das charakteristische Polynom in Linearfaktoren zerfällt, gibt es auf jeden Fall Eigenwerte; λ_1 sei einer davon, und H_1 sei der zugehörige Hauptraum. Dazu gibt es nach dem gerade bewiesenen Lemma eine Basis, bezüglich derer die Abbildungsmatrix von $\varphi_{|H_{\lambda_1}}$ eine obere Dreiecksmatrix ist, deren sämtliche Hauptdiagonaleinträge gleich λ_1 sind.

Nach dem Lemma von der FITTING-Zerlegung gibt es zu H_{λ_1} ein φ-invariantes Komplement V_1, so daß $V = H_{\lambda_1} \oplus V_1$ ist. Wir ergänzen die Basis von H_{λ_1} durch eine Basis von V_1 zu einer Basis von V; bezüglich dieser Basis hat φ dann eine Abbildungsmatrix der Form

$$A_1 = \begin{pmatrix} \boxed{D_1} & \mathbf{0} \\ \mathbf{0} & \boxed{B_1} \end{pmatrix}$$

mit einer oberen Dreiecksmatrix

$$D_1 = \begin{pmatrix} \lambda_1 & * & \ldots & * \\ 0 & \lambda_1 & \ldots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_1 \end{pmatrix}.$$

Entwicklung des charakteristischen Polynoms $\det(A - \lambda E)$ nach der ersten Spalte, gefolgt von der Entwicklung des Rests nach seiner ersten Spalte und so weiter, bis die ersten $r_1 = \dim H_{\lambda_1}$ Spalten aufgebraucht sind, zeigt, daß

$$\det(A_1 - \lambda E) = (\lambda_1 - \lambda)^{r_1} \cdot \det(B_1 - \lambda E)$$

ist, das charakteristische Polynom von $\varphi_{|V_1}$ ist also ein Teiler des charakteristischen Polynoms von φ und zerfällt somit auch in Linearfaktoren.

Insbesondere hat es (mindestens) eine Nullstelle λ_2; wir können deren Hauptraum H_{λ_2} in V_1 betrachten und damit V_1 genau wie oben weiter zerlegen in H_{λ_2} und dessen invariantes Komplement V_2. Nimmt man

nun als Basisvektoren von V zunächst die Basisvektoren von H_{λ_1} wie oben, dann entsprechende Basisvektoren für H_{λ_2} und schließlich noch solche für V_2, hat die Abbildungsmatrix A_2 bezüglich dieser neuen Basis die Form

$$A_2 = \begin{pmatrix} \boxed{D_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boxed{D_2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boxed{B_2} \end{pmatrix}$$

mit einer neuen Dreiecksmatrix

$$D_2 = \begin{pmatrix} \lambda_2 & * & \dots & * \\ 0 & \lambda_2 & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_2 \end{pmatrix} .$$

Auf diese Weise lassen sich sukzessive immer weitere Haupträume abspalten, bis schließlich eine Basis erreicht ist, bezüglich derer die Abbildungsmatrix von φ die Form

$$A = \begin{pmatrix} \boxed{D_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boxed{D_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boxed{D_s} \end{pmatrix}$$

hat mit oberen Dreiecksmatrizen

$$D_i = \begin{pmatrix} \lambda_i & * & \dots & * \\ 0 & \lambda_i & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_i \end{pmatrix}$$

zu den Eigenwerten von φ. Ist D_i eine $r_i \times r_i$-Matrix, so ist das charakteristische Polynom von φ

$$\det(A - \lambda E) = (\lambda_1 - \lambda)^{r_1}(\lambda_2 - \lambda)^{r_2} \cdots (\lambda_1 - \lambda)^{r_s} ,$$

die r_i sind also gerade die algebraischen Vielfachheiten der λ_i. ■

Für spätere Anwendungen wollen wir das gerade bewiesene Ergebnis noch etwas umformulieren:

Satz: Falls das charakteristische Polynom von $\varphi: V \to V$ in Linearfaktoren zerfällt, gibt es eine Basis von V, bezüglich derer die Abbildungsmatrix A von φ als $A = D + N$ geschrieben werden kann, wobei D eine Diagonalmatrix ist und N eine obere Dreiecksmatrix mit Nullen in der Hauptdiagonalen. Außerdem ist $DN = ND$.

Beweis: Wir nehmen natürlich die Basis aus dem gerade beendeten Beweis; die Diagonalmatrix D soll genau aus den Diagonalelementen der Abbildungsmatrix A bestehen, also die Eigenwerte entsprechend ihrer algebraischen Vielfachheiten als Diagonalelemente enthalten, und $N = A - D$. Für jede einzelne Dreiecksmatrix D_i aus dem obigen Beweis kommutiert der Diagonalanteil mit dem Rest, da der Diagonalanteil gerade das λ_i-fache der Einheitsmatrix ist. Damit ist auch $DN = ND$, denn bei beiden Multiplikationen treffen, abgesehen von den Nullen, immer nur Einträge aus einem D_i aufeinander. ■

Diese Zerlegung aus diesem Satz bezeichnet man nach dem französischen Mathematiker CAMILLE JORDAN als JORDAN-Zerlegung.

MARIE ENNEMOND CAMILLE JORDAN (1838–1922) arbeitete bei der Herleitung dieser und weiterer Zerlegungen nicht mit komplexen Matrizen, sondern mit Matrizen über endlichen Körpern, motiviert durch Fragen aus der Gruppentheorie und Lösbarkeitsfragen für nichtlineare Gleichungen. Weitere Arbeiten beschäftigen sich mit der Anwendung gruppentheoretischer Methoden auf die Geometrie sowie mit der Topologie, wo er z.B. bewies, daß jede doppeltpunktfreie geschlossene Kurve die Ebene in zwei Gebiete zerlegt. Außerdem entwickelte er neue Methoden zum Nachweis der Konvergenz von FOURIER-Reihen.

Ziel unserer Betrachtungen in diesem Paragraphen war die Berechnung von Potenzen und Exponentialfunktionen einer Matrix. Mit der JORDAN-Zerlegung ist dies im wesentlichen erreicht: Da D und N miteinander

kommutieren, gilt für Potenzen der Summe $D + N$ der „übliche“ binomische Lehrsatz, d.h.

$$(D+N)^m = \sum_{\ell=0}^{m} \binom{m}{\ell} D^{m-\ell} N^\ell ,$$

und

$$e^{D+N} = e^D \cdot e^N .$$

Die Potenzen von D sind sehr einfach zu berechnen: D^j ist wieder eine Diagonalmatrix, ihre Diagonalelemente sind die j-ten Potenzen der Diagonalelemente von D; genauso ist e^D einfach die Diagonalmatrix mit den Exponentialfunktionen der Einträge von D als Einträgen.

N ist eine obere Dreiecksmatrix mit Nullen in der Hauptdiagonalen, wir wissen also bereits, daß es einen Exponenten gibt, ab dem alle Potenzen gleich der Nullmatrix sind, so daß die Exponentialreihe zu einer endlichen Summe wird und auch in der binomischen Formel selbst für große m nur relativ wenige Summanden auftreten.

Mit der JORDAN-Zerlegung können wir diese Aussage nun noch etwas präzisieren: Die lineare Abbildung ψ zu N bildet den i-ten Basisvektor $\vec{b}_i$ ab in den von Basisvektoren $\vec{b}_j$ mit $j \le i-1$ erzeugten Unterraum. Für diese Basisvektoren gilt eine analoge Aussage, $\psi^{(2)}(\vec{b}_i)$ liegt daher im Unterraum, den die $\vec{b}_j$ mit $j \le i-2$ aufspannen. Induktiv folgt, daß $\psi^{(\ell)}(\vec{b}_i)$ im von den $\vec{b}_j$ mit $j \le i-\ell$ aufgespannten Untervektorraum liegt; falls $i-\ell$ negativ wird, ist das natürlich der Nullraum.

Die Abbildungsmatrix N^ℓ von $\psi^{(\ell)}$ ist daher ebenfalls eine obere Dreiecksmatrix mit Nullen in der Hauptdiagonalen; zusätzlich stehen auch noch in den $\ell-1$ schrägen Reihen oberhalb und parallel zur Hauptdiagonale lauter Nullen, und spätestens wenn ℓ größer oder gleich der größten Stufe eines Hauptvektors wird, ist N^ℓ gleich der Nullmatrix.

Als Beispiel betrachten wir die Matrix

$$A = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} ,$$

die offensichtlich von der oben betrachteten Form ist; hier ist

$$D = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{und} \quad N = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 3 \\ 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} .$$

Eine kurze Rechnung zeigt, daß

$$N^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{und} \quad N^3 = 0$$

ist, also ist beispielsweise

$$A^{10} = D^{10} + 10\,D^9 N + 45\,D^8 N^2 = \begin{pmatrix} 1024 & 5120 & 0 & 0 & 0 \\ 0 & 1024 & 0 & 0 & 0 \\ 0 & 0 & 1 & 20 & 390 \\ 0 & 0 & 0 & 1 & 40 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} .$$

Mehr als Potenzen interessiert uns die Exponentialfunktion einer Matrix; auch diese läßt sich über die JORDAN-Zerlegung berechnen: Da D und N kommutieren, ist $e^A = e^{D+N} = e^D \cdot e^N$ mit

$$e^D = \begin{pmatrix} e^2 & 0 & 0 & 0 & 0 \\ 0 & e^2 & 0 & 0 & 0 \\ 0 & 0 & e & 0 & 0 \\ 0 & 0 & 0 & e & 0 \\ 0 & 0 & 0 & 0 & e \end{pmatrix}$$

und

$$e^N = E + N + \frac{1}{2}N^2 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 7 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} ,$$

also ist

$$e^A = e^D \cdot e^N = \begin{pmatrix} e^2 & e^2 & 0 & 0 & 0 \\ 0 & e^2 & 0 & 0 & 0 \\ 0 & 0 & e & 2e & 7e \\ 0 & 0 & 0 & e & 4e \\ 0 & 0 & 0 & 0 & e \end{pmatrix} .$$

Entsprechend läßt sich auch e^{At} berechnen:

$$e^{Nt} = E + Nt + \frac{1}{2}N^2t^2 = \begin{pmatrix} 1 & t & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2t & 3t+4t^2 \\ 0 & 0 & 0 & 1 & 4t \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} ,$$

und da auch Dt und Nt kommutieren, ist

$$e^{At} = e^{Dt} \cdot e^{Nt} = \begin{pmatrix} e^{2t} & te^{2t} & 0 & 0 & 0 \\ 0 & e^{2t} & 0 & 0 & 0 \\ 0 & 0 & e^t & 2te^t & 3te^t + 4t^2e^t \\ 0 & 0 & 0 & e^t & 4te^t \\ 0 & 0 & 0 & 0 & e^t \end{pmatrix} .$$

Zur Vorsicht sei noch einmal ausdrücklich darauf hingewiesen, daß es für diese Rechnungen sehr wesentlich war, daß D und N miteinander kommutieren; es reicht nicht, wenn wir die Matrix A nur auf *irgendeine* Dreiecksgestalt bringen und dann als Summe einer Diagonalmatrix und einer oberen Dreiecksmatrix mit Nullen in der Hauptdiagonalen schreiben. Für

$$A = \begin{pmatrix} 1 & 3 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} 0 & 3 \\ 0 & 0 \end{pmatrix} \underset{\text{def}}{=} D + N$$

beispielsweise ist

$$\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 0 & 3 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 3 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 6 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 3 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} ,$$

und in der Tat ist

$$D^2 + 2DN + N^2 = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} + \begin{pmatrix} 0 & 6 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 6 \\ 0 & 4 \end{pmatrix}$$

verschieden von

$$A^2 = \begin{pmatrix} 1 & 9 \\ 0 & 4 \end{pmatrix} ,$$

und genauso ist

$$e^A = \begin{pmatrix} e & 3e^2 - 3e \\ 0 & e^2 \end{pmatrix} \neq e^D \cdot e^N = \begin{pmatrix} e & 3e \\ 0 & e^2 \end{pmatrix} .$$

Der Grund für die Verschiedenheit der Ergebnisse beim Quadrat liegt natürlich darin, daß wir im allgemeinen nur sagen können, daß

$$(D+N)^2 = D(D+N) + N(D+N) = D^2 + DN + ND + N^2$$

ist, aber wir können $DN + ND$ nicht zusammenfassen zu $2DN$. Mit wachsendem Exponenten verschlimmert sich die Situation drastisch; schon

$$(D+N)^3 = D^3 + D^2N + DND + ND^2 + DN^2 + NDN + N^2D + N^3$$

hat acht Summanden; die m-te Potenz hat 2^m, und von denen überleben viele auch dann, wenn N^r schon für relativ kleine r verschwindet. Für die Matrixexponentialfunktion, in die alle Potenzen eingehen, ist also ziemlich klar, daß es für nichtkommutierende Matrizen D und N keinen vernünftigen Zusammenhang zwischen e^{D+N} und $e^D \cdot e^N$ geben kann.

g) Ein Beispiel

Wir haben Eigenvektoren und Hauptvektoren in erster Linie eingeführt, um Differentialgleichungen zu lösen; daher soll das etwas ausführlichere Beispiel in diesem Abschnitt ebenfalls mit einer Differentialgleichung beginnen: Gesucht sind die Lösungen des Differentialgleichungssystems

$$\begin{aligned} \dot{x}(t) &= 2x(t) - y(t) - z(t) \\ \dot{y}(t) &= x(t) + 5y(t) + 2z(t) \\ \dot{z}(t) &= -x(t) - 2y(t) + z(t) . \end{aligned}$$

Hier ist

$$A = \begin{pmatrix} 2 & -1 & -1 \\ 1 & 5 & 2 \\ -1 & -2 & 1 \end{pmatrix} ,$$

und das charakteristische Polynom von A ist

$$\det(A - \lambda E) = -\lambda^3 + 8\lambda^2 - 21\lambda + 18 = -(\lambda - 2)(\lambda - 3)^2 .$$

Wir haben also den Eigenwert zwei mit algebraischer und somit auch geometrischer Vielfachheit eins und den Eigenwert drei mit algebraischer Vielfachheit zwei. In der Matrix

$$A - 2E = \begin{pmatrix} 0 & -1 & -1 \\ 1 & 3 & 2 \\ -1 & -2 & -1 \end{pmatrix}$$

ist die mittlere Spalte gleich der Summe der beiden äußeren,

$$\vec{v}_1 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

erzeugt also den Eigenraum. In

$$A - 3E = \begin{pmatrix} -1 & -1 & -1 \\ 1 & 2 & 2 \\ -1 & -2 & -2 \end{pmatrix}$$

stimmen die zweite und die dritte Spalte miteinander überein,

$$\vec{v}_2 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

ist also ein Eigenvektor und erzeugt auch den Eigenraum, denn da die erste Spalte kein Vielfaches der zweiten ist, hat die Matrix den Rang zwei. Die geometrische Vielfachheit des Eigenwerts drei ist also nur eins: Um zu einer Dreiecksmatrix zu kommen, müssen wir einen Hauptvektor zweiter Stufe berechnen. Aus

$$(A - 3E)^2 = \begin{pmatrix} 1 & 1 & 1 \\ -1 & -1 & -1 \\ 1 & 1 & 1 \end{pmatrix}$$

sieht man, daß sich

$$\vec{v}_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$$

als von $\vec{v}_2$ linear unabhängiger Kandidat anbietet. Da

$$A\vec{v}_3 = \begin{pmatrix} 3 \\ -4 \\ 1 \end{pmatrix} = 3\vec{v}_3 - \vec{v}_2$$

ist, hat A bezüglich der Basis $\vec{v}_1, \vec{v}_2, \vec{v}_3$ die Form

$$M = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & -1 \\ 0 & 0 & 3 \end{pmatrix} ,$$

der erste „Kasten“ ist also einfach eine 1×1-Matrix und der zweite ist

$$\begin{pmatrix} 3 & -1 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} .$$

Da das Quadrat des zweiten Summanden verschwindet, ist

$$e^{\begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} t} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} t = \begin{pmatrix} 1 & -t \\ 0 & 1 \end{pmatrix}$$

und

$$e^{\begin{pmatrix} 3 & -1 \\ 0 & 3 \end{pmatrix} t} = e^{3t} \begin{pmatrix} 1 & -t \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} e^{3t} & -te^{3t} \\ 0 & e^{3t} \end{pmatrix} .$$

Damit ist

$$e^{Mt} = \begin{pmatrix} e^{2t} & 0 & 0 \\ 0 & e^{3t} & -te^{3t} \\ 0 & 0 & e^{3t} \end{pmatrix} .$$

Um daraus e^{At} zu berechnen, müssen wir die Standardbasis des $\mathbb{R}^3$ durch die Hauptvektoren ausdrücken; man überzeugt sich leicht, daß

$$\vec{e}_1 = \vec{v}_1 + \vec{v}_2, \quad \vec{e}_2 = \vec{v}_1 + \vec{v}_2 - \vec{v}_3 \quad \text{und} \quad \vec{e}_3 = \vec{v}_1 - \vec{v}_3$$

ist. Bezüglich der Basis $\vec{v}_1, \vec{v}_2, \vec{v}_3$ ist also

$$e^{At}\vec{e}_1 = \begin{pmatrix} e^{2t} & 0 & 0 \\ 0 & e^{3t} & -te^{3t} \\ 0 & 0 & e^{3t} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} e^{2t} \\ e^{3t} \\ 0 \end{pmatrix} ,$$

was bezüglich der Standardbasis der Vektor

$$e^{2t}\vec{v}_1 + e^{3t}\vec{v}_2 = \begin{pmatrix} e^{2t} \\ -e^{2t} \\ e^{2t} \end{pmatrix} + \begin{pmatrix} 0 \\ e^{3t} \\ -e^{3t} \end{pmatrix} = \begin{pmatrix} e^{2t} \\ e^{3t} - e^{2t} \\ e^{2t} - e^{3t} \end{pmatrix} .$$

Also ist, bezüglich der Standardbasis ausgedrückt,

$$e^{At}\begin{pmatrix}1\\0\\0\end{pmatrix} = \begin{pmatrix}e^{2t}\\ e^{3t}-e^{2t}\\ e^{2t}-e^{3t}\end{pmatrix}.$$

Genauso überlegt man sich, daß

$$e^{At}\begin{pmatrix}0\\1\\0\end{pmatrix} = e^{2t}\vec{v}_1 + \left(e^{3t}+te^{3t}\right)\vec{v}_2 - e^{3t}\vec{v}_3 = \begin{pmatrix}e^{2t}-e^{3t}\\ 2e^{3t}-e^{2t}+te^{3t}\\ e^{2t}-e^{3t}-te^{3t}\end{pmatrix}$$

und

$$e^{At}\begin{pmatrix}0\\0\\1\end{pmatrix} = \begin{pmatrix}e^{2t}-e^{3t}\\ e^{3t}-e^{2t}+te^{3t}\\ e^{2t}-te^{3t}\end{pmatrix}$$

ist. Da in den Spalten einer Matrix die Bilder der Basisvektoren stehen, ist somit

$$e^{At} = \begin{pmatrix}e^{2t} & e^{2t}-e^{3t} & e^{2t}-e^{3t}\\ e^{3t}-e^{2t} & 2e^{3t}-e^{2t}+te^{3t} & e^{3t}-e^{2t}+te^{3t}\\ e^{2t}-e^{3t} & e^{2t}-e^{3t}-te^{3t} & e^{2t}-te^{3t}\end{pmatrix}.$$

Die Lösung, die den Anfangsbedingungen

$$x(0)=x_0,\quad y(0)=y_0 \quad\text{und}\quad z(0)=z_0$$

genügt, ist also

$$\begin{pmatrix}x(t)\\y(t)\\z(t)\end{pmatrix} = \begin{pmatrix}e^{2t} & e^{2t}-e^{3t} & e^{2t}-e^{3t}\\ e^{3t}-e^{2t} & 2e^{3t}-e^{2t}+te^{3t} & e^{3t}-e^{2t}+te^{3t}\\ e^{2t}-e^{3t} & e^{2t}-e^{3t}-te^{3t} & e^{2t}-te^{3t}\end{pmatrix}\begin{pmatrix}x_0\\y_0\\z_0\end{pmatrix}$$
$$= \begin{pmatrix}(x_0+y_0+z_0)e^{2t}-(y_0+z_0)e^{3t}\\ \left(x_0+(t+2)y_0+(t+1)z_0\right)e^{3t}-(x_0+y_0+z_0)e^{2t}\\ (x_0+y_0+z_0)e^{2t}-\left(x_0+(t+1)y_0+tz_0\right)e^{3t}\end{pmatrix}.$$

h) Ergänzung: Die Jordan-Normalform

Die im vorigen Abschnitt konstruierte Normalform für Abbildungsmatrizen wird für alle Zwecke dieser Vorlesung ausreichen. Trotzdem ist sie nicht ganz befriedigend, da die Dreiecksmatrizen immer noch sehr willkürlich und damit komplizierter als notwendig sind. Für Interessenten sei in diesem Abschnitt gezeigt, wie sich die bislang erreichte

Dreiecksgestalt noch weiter vereinfachen läßt, indem man die bislang noch ziemlich willkürlichen Basen der Haupträume etwas geschickter wählt. Für das folgende werden wir die Ergebnisse dieses Abschnitts nicht benötigen; er kann also gefahrlos überlesen werden.

Die Potenzen einer oberen Dreiecksmatrix mit Nullen in der Hauptdiagonalen verschwinden, wie wir gesehen haben, ab einem meist überschaubar kleinen Exponenten, aber die Potenzen bis dahin muß man doch mühsam von Hand ausrechnen. Eine Ausnahme, bei der alles klar ist, bilden Matrizen der Form

$$N = \begin{pmatrix}0&1&0&\dots&0&0\\0&0&1&\dots&0&0\\\vdots&\vdots&&\ddots&\vdots&\vdots\\0&0&0&\dots&0&1\\0&0&0&\dots&0&0\end{pmatrix},$$

bei denen direkt oberhalb der Hauptdiagonale lauter Einsen stehen, während alle anderen Einträge verschwinden, d.h.

$$N=(n_{ij})\quad\text{mit}\quad n_{ij}=\begin{cases}1 & \text{falls } j-i=1\\ 0 & \text{sonst}\end{cases}.$$

Bei der sukzessiven Potenzierung von N verschiebt sich einfach die Reihe von Einsen jeweils um eins weiter nach außen, d.h.

$$N^\ell=(n_{ij}^{(\ell)})\quad\text{mit}\quad n_{ij}^{(\ell)}=\begin{cases}1 & \text{falls } j-i=\ell\\ 0 & \text{sonst}\end{cases},$$

denn die zu N gehörige lineare Abbildung ψ bildet einfach den i-ten Basisvektor auf den $(i-1)$-ten ab oder auf den Nullvektor, falls es keinen $(i-1)$-ten Basisvektor mehr gibt, und entsprechend ist $\psi^{(\ell)}(\vec{b}_i)=\vec{b}_{i-\ell}$ beziehungsweise $\vec{0}$.

In diesem speziellen Fall sind die Potenzen von N also ohne jeden Aufwand zu berechnen, und tatsächlich genügen solche Matrizen N schon vollständig für eine Normalform der Abbildungsmatrix, die sogenannte JORDAN-Normalform:

Satz: Falls das charakteristische Polynom von $\varphi\colon V\to V$ als Produkt von Linearfaktoren geschrieben werden kann, gibt es eine Basis von V,

bezüglich derer die Abbildungsmatrix von φ die Form

$$A = \begin{pmatrix} \boxed{J_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boxed{J_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boxed{J_s} \end{pmatrix}$$

hat mit oberen Dreiecksmatrizen

$$J_i = \begin{pmatrix} \lambda_i & 1 & 0 & \dots & 0 \\ 0 & \lambda_i & 1 & \dots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & \lambda_i \end{pmatrix}$$

zu den Eigenwerten von φ. Die Anzahl der Kästchen J_i zu einem festen Eigenwert ist die geometrische Vielfachheit dieses Eigenwerts, die Summe ihrer Zeilenzahlen die algebraische.

Beweis: Wir gehen aus von der Zerlegung von V in die Haupträume zu den Eigenwerten von φ und betrachten einen festen Hauptraum H_λ. Die Einschränkung von φ auf diesen Untervektorraum läßt sich zerlegen in eine Summe

$$\varphi_{|H_\lambda} = \lambda\,\mathrm{id} + \psi\,;$$

dabei ist die Abbildungsmatrix von $\lambda\,\mathrm{id}$ bezüglich jeder beliebigen Basis gleich dem λ-fachen der Einheitsmatrix, und zumindest bezüglich der im vorigen Abschnitt konstruierten Basis $\{\vec{b}_1, \dots, \vec{b}_r\}$ ist die Abbildungsmatrix von ψ eine obere Dreiecksmatrix N mit Nullen in der Hauptdiagonalen.

ψ bildet den Basisvektor $\vec{b}_i$ daher ab in das Erzeugnis der Basisvektoren $\vec{b}_1$ bis $\vec{b}_{i-1}$; insbesondere geht $\vec{b}_1$ auf den Nullvektor. Wiederholte Anwendung von ψ zeigt, daß für jeden Basisvektor $\vec{b}_i$ gilt: $\psi^{(i-1)}(\vec{b}_i) = \vec{0}$, wobei der Exponent von ψ für die wiederholte Anwendung der Abbildung stehen soll. Insbesondere ist also $\psi^{(r)}(\vec{v}) = \vec{0}$ für alle $\vec{v} \in H_\lambda$.

Es könnte sein, daß es schon eine kleinere Zahl s gibt, so daß $\psi^{(s)}$ die Nullabbildung ist; die kleinste solche Zahl bezeichnen wir als den *Nilpotenzgrad* von ψ.

Hat der Nilpotenzgrad seinen größtmöglichen Wert r, so sind die Vektoren

$$\vec{b}_r,\ \psi(\vec{b}_r),\ \dots,\ \psi^{(s-1)}(\vec{b}_r)$$

allesamt ungleich dem Nullvektor. Sie sind auch linear unabhängig, denn ist

$$\alpha_0\vec{b}_r + \alpha_1\psi(\vec{b}_r) + \dots + \alpha_{r-1}\psi^{(r-1)}(\vec{b}_r) = \vec{0}\,,$$

so ist auch für jedes j

$$\begin{aligned} &\psi^{(j)}\big(\alpha_0\vec{b}_s + \alpha_1\psi(\vec{b}_s) + \dots + \alpha_{r-1}\psi^{(r-1)}(\vec{b}_r)\big) \\ &= \alpha_0\psi^{(j)}(\vec{b}_r) + \alpha_1\psi^{(j+1)}(\vec{b}_r) + \dots + \alpha_{r-1}\psi^{(j+r-1)}(\vec{b}_r)\big) = \vec{0}\,. \end{aligned}$$

Da $\psi^{(s)}$ für $s \geq r$ die Nullabbildung ist, treten hier nur die Summanden $\alpha_i\psi^{(i+j)}(\vec{b}_s)$ mit $i < r - j$ wirklich auf, für $j = r - 1$ also nur der Summand $\alpha_0\psi^{(r-1)}(\vec{b}_r)$. Da $\psi^{(r-1)}(\vec{b}_r)$ ungleich dem Nullvektor ist, muß also $\alpha_0 = 0$ sein. Anwendung von $\psi^{(r-2)}$ zeigt als nächstes, daß $\alpha_1 = 0$ ist, und genauso zeigt man sukzessive das Verschwinden aller α_i. Also können wir

$$\vec{c}_1 = \psi^{(r-1)}(\vec{b}_r), \quad \vec{c}_2 = \psi^{(r-2)}(\vec{b}_r), \quad \dots, \quad \vec{c}_r = \vec{b}_r$$

als Basisvektoren von H_λ wählen, und bezüglich dieser Basis ist

$$\psi(\vec{c}_i) = \begin{cases} c_{i-1} & \text{für } i > 1 \\ \vec{0} & \text{für } i = 1 \end{cases} \quad \text{und} \quad \varphi(\vec{c}_i) = \begin{cases} \lambda\vec{c}_i + \vec{c}_{i-1} & \text{für } i > 1 \\ \lambda\vec{c}_1 & \text{für } i = 1 \end{cases}.$$

Die Abbildungsmatrix von φ hat somit die einfache Gestalt

$$\begin{pmatrix} \lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & \lambda & 1 & \dots & 0 & 0 \\ 0 & 0 & \lambda & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda \end{pmatrix},$$

und das ist gerade eines der JORDAN-Kästchen aus der Formulierung des Satzes.

Falls der Nilpotenzgrad s von ψ kleiner als r ist, können wir nicht so argumentieren. Wir können aber immerhin einen Vektor $\vec{v} \in H_\lambda$ finden, so daß $\psi^{(s-1)}(\vec{v}) \neq \vec{0}$ ist, denn erst $\psi^{(s)}$ ist die Nullabbildung. Genau wie oben folgt, daß

$$\vec{c}_1 = \psi^{(s-1)}(\vec{v}), \quad \vec{c}_2 = \psi^{(s-2)}(\vec{v}), \quad \ldots, \quad \vec{c}_s = \vec{v}$$

linear unabhängig sind, allerdings spannen sie nur einen s-dimensionalen Teilraum U von H_λ auf. Dieser Teilraum ist ψ-invariant und damit auch φ-invariant, denn ψ bildet einfach die Basisvektoren aufeinander beziehungsweise auf den Nullvektor ab, und die Abbildungsmatrizen bezüglich dieser Basis sehen genauso aus wie oben; auch zu U gehört also ein JORDAN-Kästchen.

Um weitere Kästchen zu bekommen, brauchen wir ein invariantes Komplement von U in H_λ. Dazu wählen wir irgendeine lineare Abbildung $\omega\colon V \to k$, für die $\omega(\vec{v}) \neq 0$ ist und setzen

$$W = \left\{ \vec{w} \in H_\lambda \;\middle|\; \omega(\vec{w}) = \omega\big(\psi(\vec{w})\big) = \cdots = \omega\big(\psi^{(s-1)}(\vec{w})\big) = 0i \right\}.$$

Der Durchschnitt $U \cap W$ besteht nur aus dem Nullvektor, denn jeder Vektor aus U läßt sich als

$$\vec{w} = \alpha_1 \vec{c}_1 + \ldots + \alpha_s \vec{c}_s$$

schreiben, und wenn $\vec{w}$ auch in W liegt, ist

$$\omega\big(\psi^{(j)}(\vec{w})\big) = \alpha_1 \psi^{(j+s-1)}(\vec{v}) + \ldots + \alpha_{s-1}\psi^{j+1}(\vec{v}) + \alpha_s \psi^{(j)}(\vec{v}) = 0$$

für $j = 0, \ldots, s-1$. Da $\psi^{(\ell)}(\vec{v})$ für $\ell \geq s$ gleich dem Nullvektor ist, folgt für $j = s-1$, daß $\alpha_{s-1} = 0$ ist, und erniedrigt man j immer weiter, folgt nacheinander das Verschwinden aller Koeffizienten α_i. Somit ist $U \cap W$ in der Tat der Nullraum.

Die Dimension von W läßt sich zumindest nach unten leicht abschätzen: Bezüglich einer Basis von H_λ wird jede Gleichung $\omega\big(\psi^{(j)}(\vec{w})\big) = 0$ zu einer linearen Gleichung in den Koeffizienten von $\vec{w}$, der Untervektorraum W ist also die Lösungsmenge eines homogenen linearen Gleichungssystems aus s Gleichungen in $\dim H_\lambda$ Variablen. Daher ist

$$\dim W \geq \dim H_\lambda - s \quad \text{und} \quad \dim U \oplus W = \dim U + \dim W \geq \dim H_\lambda .$$

Da $U \oplus W$ Untervektorraum von H_λ ist, geht das nur, wenn das Gleichheitszeichen gilt, d.h. $H_\lambda = U \oplus W$.

Wir müssen uns noch überlegen, daß W unter ψ invariant ist. Dazu müssen wir zeigen, daß für alle $\vec{w} \in W$ gilt

$$\omega\big(\psi(\vec{w})\big) = \omega\Big(\psi\big(\big(\psi(\vec{w})\big)\Big) = \cdots = \omega\Big(\psi^{(s-1)}\big((\vec{w})\big)\Big) = 0\,,$$

d.h.

$$\omega\big(\psi(\vec{w})\big) = \omega\big(\psi^{(2)}(\vec{w})\big) = \cdots = \omega\big(\psi^{(s)}(\vec{w})\big) = 0$$

falls

$$\omega(\vec{w}) = \omega\big(\psi(\vec{w})\big) = \cdots = \omega\big(\psi^{(s-1)}(\vec{w})\big) = 0$$

ist. Die einzige neue Bedingung ist $\omega\big(\psi^{(s)}(\vec{w})\big) = 0$, und die ist trivialerweise erfüllt, da $\psi^{(s)}$ die Nullabbildung ist. Also ist W invariant unter ψ und somit ein invariantes Komplement von U.

Auch $\psi_{|W}$ ist eine nilpotente Abbildung von einem Nilpotenzgrad $s' \leq s$; wenn wir also einen Vektor $\vec{w} \in W$ hernehmen, für den $\psi^{(s')}(\vec{w}) \neq \vec{0}$ ist, können wir die gleiche Kontruktion wie oben mit $\vec{v}$ noch einmal durchführen und erhalten einen neuen invarianten Unterraum $U' \leq W$ mit einer Basis, bezüglich derer φ ein JORDAN-Kästchen als Abbildungsmatrix hat.

Falls $U' = W$ ist, sind wir damit fertig; andernfalls können wir wieder wie oben ein invariantes Komplement W' von U' in W finden und einen weiteren Teilraum abspalten, *usw.* Jeder solche Teilraum führt auf ein JORDAN-Kästchen, und das Verfahren bricht schließlich ab, da wir in einem endlichdimensionalen Vektorraum arbeiten.

In jedem der konstruierten Teilräume liegt genau ein eindimensionaler Teilraum aus Eigenvektoren (und dem Nullvektor), nämlich der vom ersten Basisvektor aufgespannte. Der Eigenraum zu λ wird also von diesen ersten Basisvektoren aufgespannt und seine Dimension, die geometrische Vielfachheit von λ, ist damit gleich der Anzahl der JORDAN-Kästchen zu λ. Die algebraische Vielfachheit ist wegen der speziellen Gestalt der Abbildungsmatrix natürlich die Anzahl der λ