Mathematik 3

Skript zur Vorlesung an der Hochschule Heilbronn
(Stand: 20. August 2023)

Prof. Dr. V. Stahl

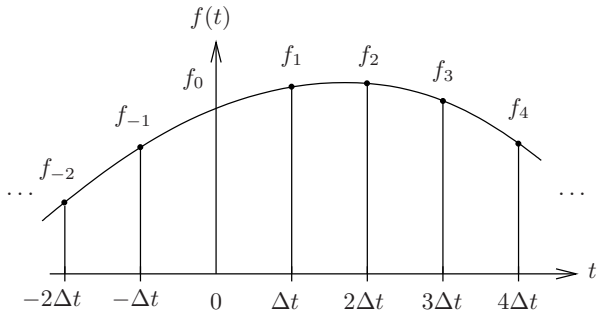
Inhaltsverzeichnis

1 z-Transformation

1.1 Abtastung.

Um zeitkontinuierliche Signale $f(t)$ digital verarbeiten zu können, muss man sie zunächst zu diskreten Zeitpunkten $t = k\Delta t$ abtasten, wobei Δt das Abtastintervall ist. Man erhält somit eine Folge von Abtastwerten

$$f_k \quad = \quad f(k\Delta t), \qquad k \in \mathbb{Z}.$$



Um die Sache zu vereinfachen, arbeiten wir immer mit Abtastintervall $\Delta t = 1$, d.h.

$$f_k \quad = \quad f(k), \quad k \in \mathbb{Z}.$$

Dies ist keine wirkliche Einschränkung: Möchte man ein anderes Abtastintervall Δt haben, kann man $f(t)$ zunächst um Faktor Δt in Zeitrichtung stauchen, d.h.

$$\hat{f}(t) \quad = \quad f(t\,\Delta t)$$

und diese Funktion mit Abtastintervall 1 abtasten:

$$\begin{aligned} \hat{f}_k \quad &= \quad \hat{f}(k) \\ &= \quad f(k\Delta t). \end{aligned}$$

Im Folgenden gehen wir wie auch bei der Laplace Transformation meistens implizit davon aus, dass $f(t) = 0$ für $t < 0$ und somit auch $f_k = 0$ für $k < 0$. Die Folge der Abtastwerte ist dann

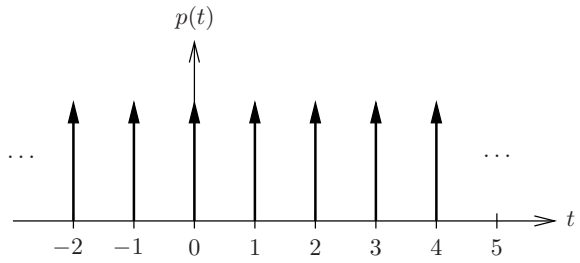
$$\langle f_0, f_1, f_2, \ldots \rangle.$$

1.2 Von der Laplace- zur z-Transformation

Impulszug. Die Funktion

$$p(t) \quad = \quad \sum_{k=-\infty}^{\infty} \delta(t-k)$$

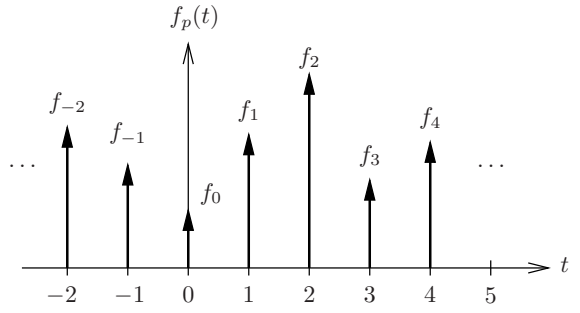
besteht aus Dirac Impulsen zu allen ganzzahligen Zeitpunkten $t \in \mathbb{Z}$. Sie wird daher Impulszug oder Kammfunktion genannt.



Multipliziert man eine Funktion $f(t)$ mit diesem Impulszug, erhält man

$$\begin{aligned} f_p(t) &= f(t)p(t) \\ &= f(t) \sum_{k=-\infty}^{\infty} \delta(t-k) \\ &= \sum_{k=-\infty}^{\infty} f(t)\delta(t-k) \\ &= \sum_{k=-\infty}^{\infty} f_k\delta(t-k). \end{aligned}$$

Im letzten Schritt wurde die Ausblendeigenschaft des Dirac Impulses verwendet. Es entsteht eine Funktion $f_p(t)$, die aus einer Folge von Impulsen besteht, deren Höhe gleich den Abtastwerten f_k von $f(t)$ ist.



Man kann somit $f_p(t)$ als Folge von Abtastwerten (multipliziert mit unendlich) interpretieren, andererseits ist $f_p(t)$ aber immer noch eine zeitkontinuierliche Funktion, die wir im nächsten Schritt Laplace transformieren:

$$
\begin{aligned}
f_p(t) \quad \circ\!\!-\!\!\bullet \quad & \int_0^\infty \sum_{k=-\infty}^{\infty} f_k \delta(t-k) e^{-st} dt \\
= \quad & \sum_{k=-\infty}^{\infty} \int_0^\infty f_k \delta(t-k) e^{-st} dt \\
= \quad & \sum_{k=-\infty}^{\infty} f_k \int_0^\infty \delta(t-k) e^{-st} dt \\
= \quad & \sum_{k=-\infty}^{\infty} f_k e^{-sk}
\end{aligned}
$$

Führt man nun die Abkürzung

$$z \;=\; e^s$$

ein, erhält man

$$e^{-sk} \;=\; (e^s)^{-k} \;=\; z^{-k}.$$

und damit

$$f_p(t) \quad \circ\!\!-\!\!\bullet \quad \sum_{k=-\infty}^{\infty} f_k z^{-k}.$$

Dieser Term hängt nur von den Abtastwerten f_k ab und wird als z-Transformation der Folge f_k bezeichnet.

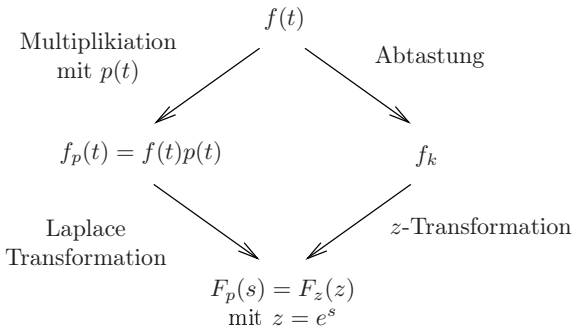
Definition 1.1 (z-Transformation)
Die z-Transformierte einer Folge $f_k, k \in \mathbb{Z}$ ist definiert durch

$$F(z) \quad = \quad \sum_{k=-\infty}^{\infty} f_k z^{-k}$$

Wie bei der Laplace Transformation hat man häufig den Spezialfall $f(t) = 0$ für $t < 0$ und folglich $f_k = 0$ für $k < 0$. Dann ist

$$F(z) \quad = \quad \sum_{k=0}^{\infty} f_k z^{-k}.$$

Der enge Zusammenhang zwischen Laplace- und z-Transformation wird durch folgendes Bild zusammengefasst:



Ist $F_p(s)$ die Laplace Transformierte von $f(t)p(t)$ und $F_z(z)$ die z-Transformierte von f_k, dann gilt

$$F_p(s) \quad = \quad F_z(e^s).$$

Wie bei der Fourier- und Laplace Transformation benutzt man auch bei der z-Transformation die Symbole

$$f_k \quad \circ\!\!-\!\!\bullet \quad F(z)$$
$$F(z) \quad \bullet\!\!-\!\!\circ \quad f_k.$$

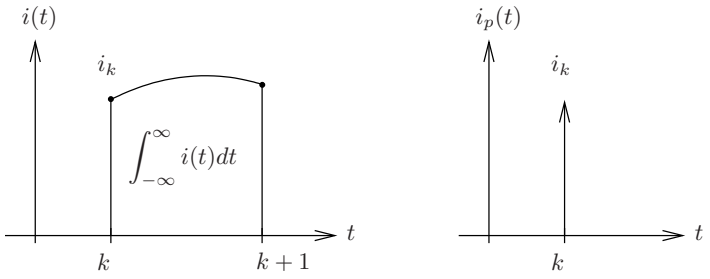
Beispiel. In der Praxis haben wir es i.a. mit zeitkontinuierlichen Signalen $f(t)$ zu tun. Um diese im Rechner verarbeiten zu können, muss die Abtastung durchgeführt werden. Wir haben diese Abtastung dadurch "realisiert", dass wir $f(t)$ mit einem Impulszug multipliziert haben, was natürlich nur theoretisch möglich ist. Es stellt sich daher die Frage, welcher physikalische Zusammenhang zwischen $f(t)$ und $f_p(t)$ besteht.

Sei $i(t)$ die Stromstärke zum Zeitpunkt t in einem elektrischen Leiter. Wir betrachten zunächst nur das Zeitintervall $[k, k+1)$ für ein $k \in \mathbb{Z}$ und nehmen an, dass die Stromstärke außerhalb dieses Intervalls Null ist. Da das Intervall kurz ist, wird angenommen, dass die die Stromstärke in diesem Intervall näherungsweise konstant i_k ist, d.h.

$$i(t) \quad = \quad \begin{cases} i_k & \text{für } t \in [k, k+1) \\ 0 & \text{sonst.} \end{cases}$$

Die transportierte Ladung ist damit

$$\int_{-\infty}^{\infty} i(t)dt \quad = \quad \int_{k}^{k+1} i(t)dt \;\approx\; \int_{k}^{k+1} i_k dt \;=\; i_k.$$



Nehmen wir nun an, dass die Stromstärke so modifiziert wird, dass alle Elektronen gleichzeitig bereits zum Zeit*punkt* k übertragen werden, im restlichen Intervall jedoch kein Strom mehr fließt. Die Stromstärke ist dann

$$i_p(t) \quad = \quad i(t)\delta(t-k).$$

Die transportierte Ladung mit dieser modifizierten Stromstärke ist aufgrund der Ausblendeigenschaft gleich wie zuvor:

$$\begin{aligned} \int_{-\infty}^{\infty} i_p(t)dt &= \int_{-\infty}^{\infty} i(t)\delta(t-k)dt \\ &= \int_{-\infty}^{\infty} i(k)\delta(t-k)dt \\ &= i_k \int_{-\infty}^{\infty} \delta(t-k)dt \\ &= i_k. \end{aligned}$$

Führt man diese Verschiebung nun in jedem Interval $[k, k+1)$, $k \in \mathbb{Z}$ durch,

erhält man die Stromstärke

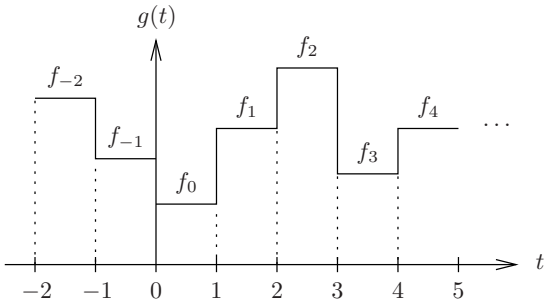
$$\begin{aligned} i_p(t) &= \sum_{k=-\infty}^{\infty} i(t)\delta(t-k) \\ &= i(t)\sum_{k=-\infty}^{\infty} \delta(t-k) \\ &= i(t)p(t). \end{aligned}$$

Die gepulste Stromstärke $i_p(t)$ transportiert in jedem Intervall $[k, k+1)$ näherungsweise die gleiche Ladung wie $i(t)$ und kann in diesem Sinne als gute Approximation an $i(t)$ verstanden werden. Lediglich der Zeitpunkt, zu dem die Eletronen geflossen sind, wurde auf den nächstgelegenen ganzzahligen Zeitpunkt k nach vorne verschoben. Die Funktion $i_p(t)$ ist zwar immer noch eine zeitkontinuierliche Funktion, andererseits aber nur von Null verschieden für ganzzahlige Zeitpunkte k.

Treppenfunktion. Wenn Ihnen die Herleitung der z-Transformation mit Hilfe des Impulszugs zu theoretisch ist, kann man die Sache auch durch Abtastung von $f(t)$ mit einen Sample and Hold Verstärker erklären. Dies ist die Art der Abtastung, wie sie in der Praxis durchgeführt wird. Hierbei entsteht eine Treppenfunktion $g(t)$ mit

$$g(t) \quad = \quad f(\lfloor t \rfloor),$$

wobei $\lfloor t \rfloor$ die nächstgelegene, ganze Zahl $\leq t$ ist.



Die Funktion $g(t)$ ist eine gute Approximation an $f(t)$, da die Fläche unter $f(t)$ näherungsweise gleich ist wie die Fläche unter $g(t)$.

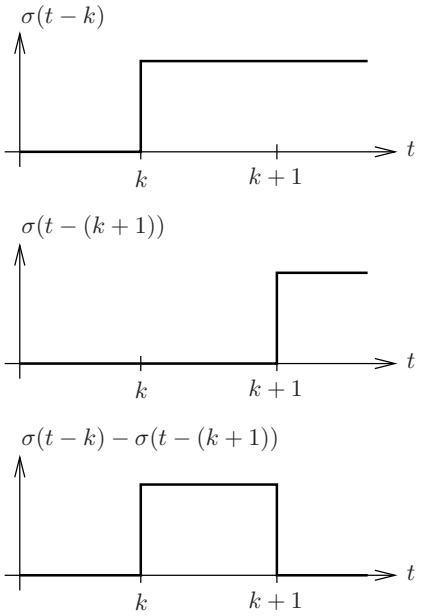
Die Laplace Transformierte von $g(t)$ ist

$$\begin{aligned}
g(t) \quad \circ\!\!-\!\!\bullet \quad & \int_{-\infty}^{\infty} g(t)e^{-st}dt \\
= \quad & \sum_{k=-\infty}^{\infty} \int_{t=k}^{k+1} g(t)e^{-st}dt \\
= \quad & \sum_{k=-\infty}^{\infty} \int_{t=k}^{k+1} f_k e^{-st}dt \\
= \quad & \sum_{k=-\infty}^{\infty} f_k \int_{t=k}^{k+1} e^{-st}dt \\
= \quad & \sum_{k=-\infty}^{\infty} f_k \left[-\frac{1}{s} e^{-st} \right]_k^{k+1} \\
= \quad & \sum_{k=-\infty}^{\infty} f_k \left(-\frac{e^{-s(k+1)} - e^{-sk}}{s} \right) \\
= \quad & \sum_{k=-\infty}^{\infty} f_k \left(\frac{1 - e^{-s}}{s} \right) e^{-sk} \\
= \quad & \frac{1 - e^{-s}}{s} \sum_{k=-\infty}^{\infty} f_k e^{-sk}.
\end{aligned}$$

Man hätte $g(t)$ auch kompakt mit Hilfe von Sprungfunktionen beschreiben können:

$$g(t) \quad = \quad \sum_{k=0}^{\infty} f_k\big(\sigma(t-k) - \sigma(t-(k+1))\big).$$

Die Differenz der beiden Sprungfunktionen ist Eins ist wenn $t \in [k, k+1)$ und Null sonst.



Mit Hilfe der Korrespondenz

$$\sigma(t) \quad \circ\!\!-\!\!\bullet \quad \frac{1}{s}$$

und dem Verschiebungssatz der Laplace Transformation erhält man

$$\sigma(t-k) \quad \circ\!\!-\!\!\bullet \quad \frac{e^{-sk}}{s}$$
$$\sigma(t-(k+1)) \quad \circ\!\!-\!\!\bullet \quad \frac{e^{-s(k+1)}}{s}.$$

Mit der Linearität der Laplace Transformation kommen wir somit zum gleichen Ergebnis wie oben:

$$g(t) \quad \circ\!\!-\!\!\bullet \quad \sum_{k=-\infty}^{\infty} f_k \left(\frac{e^{-sk}}{s} - \frac{e^{-s(k+1)}}{s} \right)$$
$$= \quad \frac{1-e^{-s}}{s} \sum_{k=-\infty}^{\infty} f_k e^{-sk}.$$

Versuchen wir nun zu verstehen, was die beiden Faktoren

$$\frac{1-e^{-s}}{s} \quad \text{und} \quad \sum_{k=-\infty}^{\infty} f_k e^{-sk}$$

der Laplace Transformierten von $g(t)$ bedeuten.

- Der erste Faktor ist unabhängig von den Abtastwerten f_k und entspricht einem Rechteckimpuls im Zeitbereich.

$$\frac{1-e^{-s}}{s} \quad \bullet\!\!-\!\!\circ \quad r(t) = \begin{cases} 1 & \text{falls } 0 \le t \le 1 \\ 0 & \text{sonst} \end{cases}$$

- Der zweite Faktor entspricht im Zeitbereich einer Summe von verschobenen Dirac Impulsen an der Stelle $t = k$ mit Intensität f_k.

$$\begin{aligned} e^{-sk} \quad &\bullet\!\!-\!\!\circ \quad \delta(t-k) \\ f_k e^{-sk} \quad &\bullet\!\!-\!\!\circ \quad f_k \delta(t-k) \\ \sum_{k=-\infty}^{\infty} f_k e^{-sk} \quad &\bullet\!\!-\!\!\circ \quad \sum_{k=-\infty}^{\infty} f_k \delta(t-k) \;=\; f_p(t). \end{aligned}$$

  Wir erhalten also eine Folge von Dirac Impulsen, deren Höhe genau die Abtastwerte f_k sind, d.h. die Funktion $f_p(t)$ bei der Herleitung mit dem Impulszug, siehe Seite 4.

Da die Multiplikation im Bildbereich einer Faltung im Zeitbereich entspricht, hätte man die Treppenfunktion $g(t)$ auch unter Verwendung des Faltungssatzes durch

$$g(t) \quad = \quad r(t) * f_p(t)$$

konstruieren können.

Mit der Abkürzung

$$z \quad = \quad e^{s}$$

gilt

$$\sum_{k=-\infty}^{\infty} f_k e^{-sk} \quad = \quad \sum_{k=-\infty}^{\infty} f_k z^{-k}$$

Man sieht, dass auch auf diesem Weg wieder die z-Transformierte von f_k entsteht.

1.3 Konvergenz

Abschließend stellt sich die Frage, für welche werte von $z \in \mathbb{C}$ die z-Transformierte von f_k existiert. Hier wird nur der Spezialfall $f_k = 0$ für $k < 0$ bzw. $f(t) = 0$ für $t < 0$ betrachtet.

Von der Laplace Transformierten

$$\sum_{k=0}^{\infty} f_k e^{-sk} \quad \bullet\!\!-\!\!\circ \quad \sum_{k=0}^{\infty} f_k \delta(t-k)$$

ist bekannt, dass sie entweder für kein s existiert, für alle s oder für genau die s mit $\mathrm{re}(s) > a$ für ein bestimmtes $a \in \mathbb{R}$. Mit

$$z = e^s = e^{\mathrm{re}(s)} e^{j\mathrm{im}(s)}$$

gilt

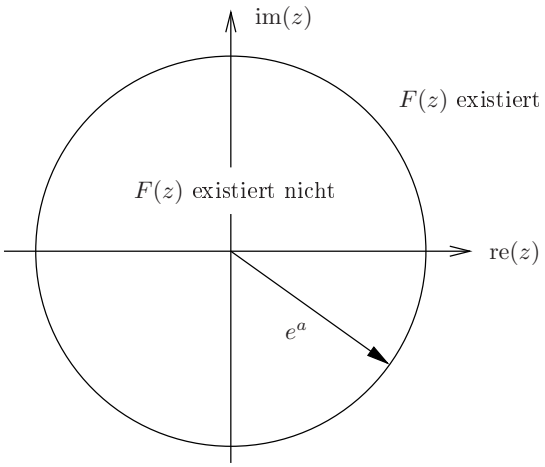
$$|z| = |e^{\mathrm{re}(s)} e^{j\mathrm{im}(s)}| = |e^{\mathrm{re}(s)}| \underbrace{|e^{j\mathrm{im}(s)}|}_{=1} = e^{\mathrm{re}(s)}.$$

Da die e-Funktion streng monoton steigend ist, gilt

$$\begin{array}{lll} \mathrm{re}(s) > a & \text{genau dann wenn} & e^{\mathrm{re}(s)} > e^a \\ & \text{genau dann wenn} & |z| > e^a \end{array}$$

d.h. wenn z in der komplexen Ebene außerhalb des Kreises mit Radius e^a liegt.

Zusammengefasst bedeutet dies, dass $F(z)$ entweder für kein z existiert oder für alle z (außer $z = 0$, da e^s nie Null werden kann) oder dass es einen Kreis in der komplexen Ebene gibt, so dass $F(z)$ für genau die $z \in \mathbb{C}$ existiert, die außerhalb dieses Kreises liegen.



1.4 Beispiele

In diesem Abschnitt werden die z-Transformierten einiger Beispiele berechnet. Beachten Sie, dass es bei der z-Transformation um eine Folge f_k geht und nicht etwa wie bei der Laplace Transformation um eine Funktion $f \in \mathbb{R} \to \mathbb{R}$.

Beispiel 1.2 Der diskrete Dirac Impuls δ_k ist definiert durch

$$\delta_k = \begin{cases} 1 & \text{falls } k = 0 \\ 0 & \text{sonst.} \end{cases}$$

Die z-Transformierte berechnet man wie folgt:

$$\begin{aligned} \delta_k \quad \circ\!\!-\!\!\bullet \quad & \sum_{k=-\infty}^{\infty} \delta_k z^{-k} \\ = \quad & z^0 \\ = \quad & 1 \end{aligned}$$

Auch im diskreten Fall gilt die Ausblendeigenschaft des Dirac Impulses. Ist $f \in \mathbb{Z} \to \mathbb{R}$ eine beliebige Folge, dann gilt für alle $k, \hat{k}$

$$f_k \delta_{k-\hat{k}} \quad = \quad f_{\hat{k}} \delta_{k-\hat{k}}.$$

Der Beweis geschieht durch Fallunterscheidung.

- Für $k \neq \hat{k}$ sind beide Seiten aufgrund des Faktors $\delta_{k-\hat{k}}$ gleich Null.
- Für $k = \hat{k}$ steht auf beiden Seiten $f_{\hat{k}} \delta_0$.

Beispiel 1.3 Die diskrete Einheitssprungfunktion σ_k ist definiert durch

$$\sigma_k = \begin{cases} 1 & \text{für } k \geq 0 \\ 0 & \text{sonst.} \end{cases}$$

Die z-Transformierte berechnet man wie folgt:

$$\begin{aligned} \sigma_k \quad \circ\!\!-\!\!\bullet \quad & \sum_{k=-\infty}^{\infty} \sigma_k z^{-k} \\ = \quad & \sum_{k=0}^{\infty} z^{-k}. \end{aligned}$$

Für die Berechnung einer unendlichen Summe dieser Bauart gibt es einen Trick. Zunächst ist klar, dass die Summe nur endlich sein kann, wenn die Summanden für große k gegen Null gehen. Dies ist genau dann der Fall wenn $|z| > 1$.

Sei nun

$$S = \sum_{k=0}^{\infty} z^{-k}.$$

Dann ist

$$\begin{aligned} z^{-1}S &= z^{-1} \sum_{k=0}^{\infty} z^{-k} \\ &= \sum_{k=0}^{\infty} z^{-(k+1)} \\ &= \sum_{k=1}^{\infty} z^{-k}. \end{aligned}$$

Weiterhin gilt

$$\begin{aligned} S - z^{-1}S &= \sum_{k=0}^{\infty} z^{-k} - \sum_{k=1}^{\infty} z^{-k} \\ &= z^0 \\ &= 1. \end{aligned}$$

Damit erhält man die Gleichung

$$\begin{aligned} S - z^{-1}S &= 1 \\ S(1 - z^{-1}) &= 1 \\ S &= \frac{1}{1 - z^{-1}}. \end{aligned}$$

Multipliziert man Zähler und Nenner mit z erhält man

$$S = \frac{z}{z-1}$$

und damit

$$\sigma_k \quad \circ\!\!-\!\!\bullet \quad \frac{z}{z-1} \qquad \text{falls } |z| > 1.$$

Beispiel 1.4 Sei

$$f_k \quad = \quad \sigma_k a^k.$$

Die z-Transformierte berechnet man wie folgt:

$$\begin{aligned} f_k \quad \circ\!\!-\!\!\bullet \quad & \sum_{k=-\infty}^{\infty} f_k z^{-k} \\ = \quad & \sum_{k=0}^{\infty} a^k z^{-k} \\ = \quad & \sum_{k=0}^{\infty} \left(\frac{a}{z}\right)^k. \end{aligned}$$

Nun wendet man den gleichen Trick wie im vorigen Beispiel an. Die Summe ist endlich genau dann wenn $|z| > |a|$. Sei

$$\begin{aligned} S &= \sum_{k=0}^{\infty} \left(\frac{a}{z}\right)^k \\ \frac{a}{z}S &= \frac{a}{z}\sum_{k=0}^{\infty} \left(\frac{a}{z}\right)^k \\ &= \sum_{k=0}^{\infty} \left(\frac{a}{z}\right)^{k+1} \\ &= \sum_{k=1}^{\infty} \left(\frac{a}{z}\right)^k \\ S - \frac{a}{z}S &= \sum_{k=0}^{\infty} \left(\frac{a}{z}\right)^k - \sum_{k=1}^{\infty} \left(\frac{a}{z}\right)^k \\ &= \left(\frac{a}{z}\right)^0 \\ &= 1 \\ S\left(1 - \frac{a}{z}\right) &= 1 \\ S &= \frac{1}{1 - \frac{a}{z}} \\ &= \frac{z}{z-a}. \end{aligned}$$

Damit gilt

$$f_k \quad \circ\!\!-\!\!\bullet \quad \frac{z}{z-a} \qquad \text{falls } |z| > |a| \ .$$

Beispiel 1.5 Sei

$$f_k \;=\; \langle 4,5,7,\ 4,5,7,\ 4,5,7,\ldots\rangle,$$

d.h. eine unendliche Wiederholung von $4,5,7$, die bei $k=0$ beginnt. Die z-Transformierte berechnet man wie folgt:

$$\begin{aligned} f_k \;\circ\!\!-\!\!\bullet\; & 4z^{-0}+5z^{-1}+7z^{-2}+ \\ & 4z^{-3}+5z^{-4}+7z^{-5}+\ldots \\ = \;& \left(4z^{-0}+5z^{-1}+7z^{-2}\right)\left(1+z^{-3}+z^{-6}+\ldots\right). \end{aligned}$$

Der zweite Faktor kann wieder in geschlossener Form dargestellt werden. Sei also

$$\begin{aligned} S &= 1+z^{-3}+z^{-6}+\ldots \\ z^{-3}S &= \quad z^{-3}+z^{-6}+\ldots \\ S-z^{-3}S &= 1 \\ S\left(1-z^{-3}\right) &= 1 \\ S &= \frac{1}{1-z^{-3}} \\ S &= \frac{z^3}{z^3-1} \end{aligned}$$

Damit gilt

$$\begin{aligned} f_k \;\circ\!\!-\!\!\bullet\; & \left(4z^{-0}+5z^{-1}+7z^{-2}\right)\frac{z^3}{z^3-1} \\ = \;& \frac{4z^3+5z^2+7z}{z^3-1}. \end{aligned}$$

1.5 Eigenschaften der z-Transformation

Theorem 1.6 (Linearität)

$$\begin{aligned} f_k + g_k &\circ\!\!-\!\!\bullet\ F(z) + G(z) \\ uf_k &\circ\!\!-\!\!\bullet\ uF(z). \end{aligned}$$

Beweis.

$$\begin{aligned} f_k + g_k &\circ\!\!-\!\!\bullet \sum_{k=-\infty}^{\infty} (f_k + g_k) z^{-k} \\ &= \sum_{k=-\infty}^{\infty} f_k z^{-k} + \sum_{k=-\infty}^{\infty} g_k z^{-k} \\ &= F(z) + G(z) \\ uf_k &\circ\!\!-\!\!\bullet \sum_{k=-\infty}^{\infty} u f_k z^{-k} \\ &= u \sum_{k=-\infty}^{\infty} f_k z^{-k} \\ &= uF(z). \end{aligned}$$

Theorem 1.7 (Dämpfung)

$$a^k f_k \ \circ\!\!-\!\!\bullet\ F\left(\frac{z}{a}\right)$$

Beweis.

$$\begin{aligned} a^k f_k &\circ\!\!-\!\!\bullet \sum_{k=-\infty}^{\infty} a^k f_k z^{-k} \\ &= \sum_{k=-\infty}^{\infty} f_k \left(\frac{z}{a}\right)^{-k} \\ &= F\left(\frac{z}{a}\right) \end{aligned}$$

Für die Zeitverschiebung gibt es wie bei der Laplace Transformation mehrere Varianten.

Theorem 1.8 (Zeitverschiebung)
Sei

$$f_k \quad \circ\!\!-\!\!\bullet \quad F(z).$$

Dann gilt für jedes $m \in \mathbb{Z}$

$$f_{k-m} \quad \circ\!\!-\!\!\bullet \quad z^{-m}F(z).$$

Beweis.

$$\begin{aligned} f_{k-m} \quad \circ\!\!-\!\!\bullet \quad & \sum_{k=-\infty}^{\infty} f_{k-m} z^{-k} \\ = \quad & \sum_{k=-\infty}^{\infty} f_k z^{-(k+m)} \\ = \quad & \sum_{k=-\infty}^{\infty} f_k z^{-k} z^{-m} \\ = \quad & z^{-m} \sum_{k=-\infty}^{\infty} f_k z^{-k} \\ = \quad & z^{-m} F(z). \end{aligned}$$

Sehr häufig hat die Folge im Zeitbereich die Form $\sigma_k f_k$, da man Folgen untersuchen möchte, die Null sind für negative k. In diesem Fall gilt entsprechend folgendes Theorem.

Theorem 1.9 (Zeitverschiebung)
Sei

$$\sigma_k f_k \quad \circ\!\!-\!\!\bullet \quad F(z).$$

Dann gilt für jedes $m \in \mathbb{Z}$

$$\sigma_{k-m} f_{k-m} \quad \circ\!\!-\!\!\bullet \quad z^{-m}F(z).$$

Der Faktor σ_{k-m} bewirkt, dass das Signal im Zeitbereich mit $k-m$ Nullen beginnt. Es ist bei der Anwendung des Verschiebungssatzes somit wichtig, den Faktor σ_k explizit mit zu verschieben!

Beispiel 1.10 Es wurde bereits gezeigt, dass

$$\sigma_k a^k \quad \circ\!\!-\!\!\bullet \quad \frac{z}{z-a}.$$

Mit der Zeitverschiebung im Fall $m = 1$ erhält man somit

$$\sigma_{k-1} a^{k-1} \quad \circ\!\!-\!\!\bullet \quad \frac{1}{z-a}.$$

Andererseits ist

$$\begin{aligned} \sigma_k a^{k-1} &= a^{-1}\sigma_k a^k \\ &\circ\!\!-\!\!\bullet \quad \frac{a^{-1}z}{z-1}. \end{aligned}$$

Möchte man den Faktor σ_k nicht mitverschieben, wird der Verschiebungssatz etwas komplizierter. Insbesondere entstehen unterschiedliche Formeln je nachdem ob man nach links oder nach rechts verschiebt.

Theorem 1.11 (Zeitverschiebung)
Sei

$$\sigma_k f_k \quad \circ\!\!-\!\!\bullet \quad F(z).$$

Dann gilt für jedes $m \in \mathbb{N}$

$$\begin{aligned} \sigma_k f_{k+m} &\quad \circ\!\!-\!\!\bullet \quad z^m \left(F(z) - \sum_{k=0}^{m-1} f_k z^{-k} \right) \\ \sigma_k f_{k-m} &\quad \circ\!\!-\!\!\bullet \quad z^{-m} \left(F(z) + \sum_{k=-m}^{-1} f_k z^{-k} \right). \end{aligned}$$

Beispiel 1.12 Auf o.g. Beispiel angewandt, erhält man mit $m = 1$

$$\begin{aligned} \sigma_k a^k &\quad \circ\!\!-\!\!\bullet \quad \frac{z}{z-a} \\ \sigma_k a^{k-1} &\quad \circ\!\!-\!\!\bullet \quad z^{-1} \left(\frac{z}{z-a} + \sum_{k=-1}^{-1} a^k z^{-k} \right) \\ &= \quad z^{-1} \left(\frac{z}{z-a} + a^{-1} z^1 \right) \\ &= \quad \frac{1}{z-a} + a^{-1} \\ &= \quad \frac{1 + a^{-1}(z-a)}{z-a} \\ &= \quad \frac{1 + a^{-1}z - 1}{z-1} \\ &= \quad \frac{a^{-1}z}{z-a}. \end{aligned}$$

Beweis.

$$
\begin{aligned}
\sigma_k f_{k+m} \quad \circ\!\!-\!\!\bullet \quad & \sum_{k=-\infty}^{\infty} \sigma_k f_{k+m} z^{-k} \\
= \quad & \sum_{k=0}^{\infty} f_{k+m} z^{-k} \\
= \quad & \sum_{k=m}^{\infty} f_k z^{-(k-m)} \\
= \quad & z^m \sum_{k=m}^{\infty} f_k z^{-k} \\
= \quad & z^m \left(\sum_{k=0}^{\infty} f_k z^{-k} - \sum_{k=0}^{m-1} f_k z^{-k} \right) \\
= \quad & z^m \left(\sum_{k=-\infty}^{\infty} \sigma_k f_k z^{-k} - \sum_{k=0}^{m-1} f_k z^{-k} \right) \\
= \quad & z^m \left(F(z) - \sum_{k=0}^{m-1} f_k z^{-k} \right) \\
\sigma_k f_{k-m} \quad \circ\!\!-\!\!\bullet \quad & \sum_{k=-\infty}^{\infty} \sigma_k f_{k-m} z^{-k} \\
= \quad & \sum_{k=0}^{\infty} f_{k-m} z^{-k} \\
= \quad & \sum_{k=-m}^{\infty} f_k z^{-(k+m)} \\
= \quad & z^{-m} \sum_{k=-m}^{\infty} f_k z^{-k} \\
= \quad & z^{-m} \left(\sum_{k=0}^{\infty} f_k z^{-k} + \sum_{k=-m}^{-1} f_k z^{-k} \right) \\
= \quad & z^{-m} \left(\sum_{k=-\infty}^{\infty} \sigma_k f_k z^{-k} + \sum_{k=-m}^{-1} f_k z^{-k} \right) \\
= \quad & z^{-m} \left(F(z) + \sum_{k=-m}^{-1} f_k z^{-k} \right).
\end{aligned}
$$

Theorem 1.13 (Ableitung im Bildbereich)

$$kf_k \quad \circ\!\!-\!\!\bullet \quad -zF'(z)$$

Beweis. Aus

$$F(z) = \sum_{k=-\infty}^{\infty} f_k z^{-k}$$

folgt

$$\begin{aligned} F'(z) &= \sum_{k=-\infty}^{\infty} f_k(-k)z^{-k-1} \\ &= -z^{-1} \sum_{k=-\infty}^{\infty} kf_k z^{-k} \end{aligned}$$

Multiplikation mit $-z$ auf beiden Seiten ergibt

$$\begin{aligned} -zF'(z) &= \sum_{k=-\infty}^{\infty} kf_k z^{-k} \\ &\circ\!\!-\!\!\bullet \quad kf_k. \end{aligned}$$

Beispiel 1.14 In Beispiel 1.3 wurde gezeigt, dass

$$\sigma_k \quad \circ\!\!-\!\!\bullet \quad \frac{z}{z-1}.$$

Mit der Ableitung im Bildbereich gilt damit

$$\begin{aligned} k\sigma_k \quad \circ\!\!-\!\!\bullet \quad & -z\left(\frac{z}{z-1}\right)' \\ = \quad & \frac{z}{(z-1)^2}. \end{aligned}$$

Auf gleiche Weise kann man auch die z-Transformierte von k^n für beliebige $n \in \mathbb{N}$ berechnen.

1.6 Diskrete Faltung

Definition 1.15 (Diskrete Faltung)
Die Faltung $f * g$ zweier Folgen $f, g \in \mathbb{Z} \to \mathbb{R}$ ist definiert durch

$$(f * g)_k = \sum_{\ell=-\infty}^{\infty} f_\ell g_{k-\ell}$$

für alle $k \in \mathbb{Z}$.

Notation 1.16
Die um $\hat{k}$ Takte verzögerte Folge f wird mit $f_{\cdot-\hat{k}}$ bezeichnet, d.h.

$$\left(f_{\cdot-\hat{k}}\right)_k = f_{k-\hat{k}}$$

für alle $k \in \mathbb{Z}$.

Mit dieser Notation lässt sich die Faltung zweier Folgen f, g kompakter formulieren:

$$\begin{aligned}
(f * g)_k &= \sum_{\ell=-\infty}^{\infty} f_\ell g_{k-\ell} \\
&= \sum_{\ell=-\infty}^{\infty} f_\ell (g_{\cdot-\ell})_k \\
&= \sum_{\ell=-\infty}^{\infty} (f_\ell g_{\cdot-\ell})_k \\
&= \left(\sum_{\ell=-\infty}^{\infty} f_\ell g_{\cdot-\ell}\right)_k \qquad \text{für alle } k \in \mathbb{Z} \text{ bzw.} \\
f * g &= \sum_{\ell=-\infty}^{\infty} f_\ell g_{\cdot-\ell}.
\end{aligned}$$

Beachten Sie, dass nun auf beiden Seiten der Gleichung Folgen stehen, nicht Zahlen. Der ℓ-te Summand

$$f_\ell g_{\cdot-\ell}$$

ist die mit Faktor f_ℓ gewichtete, um ℓ Takte verschobene Folge g. Die Faltung $f * g$ beschreibt somit eine gewichtete Summe von verschobenen Kopien von g. In der Akustik werden solche Kopien als Echo bezeichnet. Durchläuft ein Schallsignal g einen Raum, kommt beim Empfänger tatsächlich nicht nur g an sondern aufgrund der Reflektionen an den Wänden auch dessen Echos.

Besonders relevant sind Folgen mit

$$f_k = g_k = 0 \text{ für } k < 0.$$

In diesem Fall ist

$$\begin{aligned}(f * g)_k &= \sum_{\ell=-\infty}^{\infty} f_\ell g_{k-\ell} \\ &= \sum_{\ell=0}^{\infty} f_\ell g_{k-\ell} \\ &= \sum_{\ell=0}^{k} f_\ell g_{k-\ell}.\end{aligned}$$

Insbesondere ist dann auch $(f * g)_k = 0$ für $k < 0$.

Theorem 1.17 (Faltungssatz)
Sei

$$\begin{aligned}f_k &\circ\!\!-\!\!\bullet\ F(z) \text{ und} \\ g_k &\circ\!\!-\!\!\bullet\ G(z).\end{aligned}$$

Dann gilt

$$(f * g)_k \circ\!\!-\!\!\bullet\ F(z)G(z).$$

Beweis. Die nachfolgenden Summen laufen immer von $-\infty$ bis ∞. Um die Notation zu vereinfachen werden die Summationsgrenzen weggelassen.

$$\begin{aligned}(f * g)_k &\circ\!\!-\!\!\bullet \sum_\ell f_\ell g_{k-\ell} \\ &\circ\!\!-\!\!\bullet \sum_k \sum_\ell f_\ell g_{k-\ell} z^{-k} \\ &= \sum_\ell \sum_k f_\ell g_{k-\ell} z^{-k} \\ &= \sum_\ell f_\ell \underbrace{\left(\sum_k g_{k-\ell} z^{-k} \right)}_{\circ\!\!-\!\!\bullet z^{-\ell} G(z)} \\ &= \left(\sum_\ell f_\ell z^{-\ell} \right) G(z) \\ &= F(z)G(z)\end{aligned}$$

Theorem 1.18 (Eigenschaften der Faltung)
Seien $f, g, h \in \mathbb{Z} \to \mathbb{R}$ und $u \in \mathbb{R}$.

- *Linearität*

$$\begin{aligned}(f+g)*h &= f*h+g*h\\ (uf)*g &= u(f*g)\end{aligned}$$

- *Kommutativgesetz*

$$f*g = g*f$$

- *Assoziativgesetz*

$$f*(g*h) = (f*g)*h$$

- *Dirac Impuls ist neutrales Element der Faltung*

$$f*\delta = f$$

- *Zeitinvarianz.*

$$f_{\cdot-\hat{k}}*g = (f*g)_{\cdot-\hat{k}}.$$

Beweis. Die Eigenschaften lassen sich natürlich alle im Zeitbereich beweisen. Viel einfacher geht's jedoch über die z-Transformation und den Faltungssatz.

- Linearität

$$\begin{aligned}(f+g)*h \quad &\circ\!\!-\!\!\bullet \quad (F(z)+G(z))H(z)\\ &= \quad F(z)H(z)+G(z)H(z)\\ &\bullet\!\!-\!\!\circ \quad f*h+g*h\\ (uf)*g \quad &\circ\!\!-\!\!\bullet \quad uF(z)G(z)\\ &\bullet\!\!-\!\!\circ \quad u(f*g)\end{aligned}$$

- Kommutativgesetz

$$\begin{aligned}f*g \quad &\circ\!\!-\!\!\bullet \quad F(z)G(z)\\ &= \quad G(z)F(z)\\ &\bullet\!\!-\!\!\circ \quad g*f\end{aligned}$$

- Assoziativgesetz

$$\begin{aligned}f*(g*h) \quad &\circ\!\!-\!\!\bullet \quad F(z)\big(G(z)H(z)\big)\\ &= \quad \big(F(z)G(z)\big)H(z)\\ &\bullet\!\!-\!\!\circ \quad (f*g)*h\end{aligned}$$

- Dirac Impuls ist neutrales Element der Faltung

$$\begin{aligned} f * \delta \quad &\circ\!\!-\!\!\bullet \quad F(z) \cdot 1 \\ &\bullet\!\!-\!\!\circ \quad f \end{aligned}$$

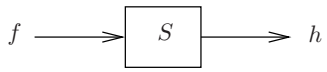
- Zeitinvarianz

$$\begin{aligned} f_{.-\hat{k}} * g \quad &\circ\!\!-\!\!\bullet \quad (z^{-\hat{k}} F(z)) G(z) \\ &= \quad z^{-\hat{k}} (F(z) G(z)) \\ &\bullet\!\!-\!\!\circ \quad (f * g)_{.-\hat{k}} \end{aligned}$$

1.7 Lineare zeitinvariante Systeme

Unter einem System versteht man in diesem Kontext eine Funktion S, die eine Folge f in eine Folge $h = S(f)$ transformiert, d.h. eine Funktion von Folgen:

$$S \in (\mathbb{Z} \to \mathbb{R}) \to (\mathbb{Z} \to \mathbb{R}).$$



Die Notation $h_k = S(f_k)$ ist irreführend, da der Abtastwert h_k nicht aus dem Abtastwert f_k berechnet wird sondern i.a. von allen Abtastwerten von f abhängt. Korrekt muss man schreiben

$$\begin{array}{rcll} h & = & S(f) & \text{oder} \\ h_k & = & [S(f)]_k & \text{für alle } k \in \mathbb{Z} \end{array}$$

um auszudrücken, dass S auf die gesamte Folge f angewandt wird und von der resultierenden Folge $S(f)$ der k-te Abtastwert genommen wird.

Beispiel 1.19 Nachfolgend ein paar Beispiele für Systeme.

- Verzögerung um einen Takt.

$$[S(f)]_k = f_{k-1}$$

- Verstärker mit Verstärkungsfaktor c.

$$[S(f)]_k = cf_k$$

- Differenzierer.

$$[S(f)]_k = f_k - f_{k-1}.$$

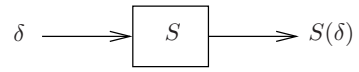
- Summierer.

$$[S(f)]_k = \sum_{\ell=-\infty}^{k} f_\ell$$

- Faltung mit einer Folge g.

$$\begin{array}{rcl} [S(f)]_k & = & (f * g)_k \\ & = & \displaystyle\sum_{\ell=-\infty}^{\infty} f_\ell g_{k-\ell} \end{array}$$

Definition 1.20 (Impulsantwort)
Die Impulsantwort eines Systems S ist die Folge $S(\delta)$.



Beispiel 1.21

- Verzögerungsglied

$$\begin{aligned} [S(f)]_k &= f_{k-1} \\ [S(\delta)]_k &= \delta_{k-1} \\ &= \begin{cases} 1 & \text{falls } k = 1 \\ 0 & \text{sonst.} \end{cases} \end{aligned}$$

- Verstärker

$$\begin{aligned} [S(f)]_k &= cf_k \\ [S(\delta)]_k &= (c\delta)_k \\ &= \begin{cases} c & \text{falls } k = 0 \\ 0 & \text{sonst.} \end{cases} \end{aligned}$$

- Differenzierer

$$\begin{aligned} [S(f)]_k &= f_k - f_{k-1} \\ [S(\delta)]_k &= \delta_k - \delta_{k-1} \\ &= \begin{cases} 1 & \text{falls } k = 0 \\ -1 & \text{falls } k = 1 \\ 0 & \text{sonst.} \end{cases} \end{aligned}$$

- Summierer

$$\begin{aligned} [S(f)]_k &= \sum_{\ell=-\infty}^{k} f_\ell \\ [S(\delta)]_k &= \sum_{\ell=-\infty}^{k} \delta_\ell \\ &= \begin{cases} 1 & \text{falls } k \geq 0 \\ 0 & \text{sonst.} \end{cases} \\ &= \sigma_k \\ S(\delta) &= \sigma \end{aligned}$$

- Faltung mit g

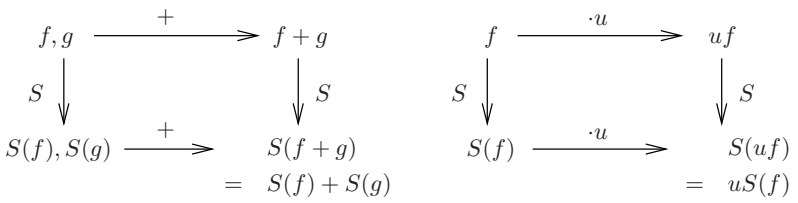
$$\begin{aligned} [S(f)]_k &= (f * g)_k \\ S(\delta) &= \delta * g \\ &= g \end{aligned}$$

Definition 1.22 (Lineares System)
Ein System S heißt linear wenn

$$\begin{aligned} S(f+g) &= S(f)+S(g) \\ S(uf) &= uS(f) \end{aligned}$$

für alle Folgen $f, g \in \mathbb{Z} \to \mathbb{R}$ und $u \in \mathbb{R}$.

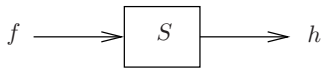
Die beiden Linearitätsbedingungen lassen sich durch das folgende kommutative Diagram darstellen.



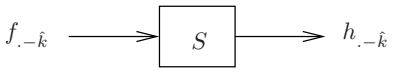
Alle o.g. Beispiele von Systemen sind linear. Nichtlinear sind z.B.

$$\begin{aligned} [S(f)]_k &= f_k^2 \\ [S(f)]_k &= f_k + 1 \\ [S(f)]_k &= 1 \end{aligned}$$

Ein System heißt zeitinvariant, wenn die Verzögerung des Inputsignals f um $\hat{k}$ Takte bewirkt, dass auch das Outputsignal $h = S(f)$ um $\hat{k}$ Takte verzögert wird, d.h. wenn



dann

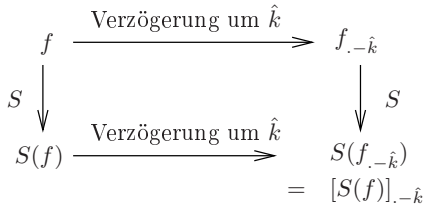


Definition 1.23 (Zeitinvariantes System)
Ein System S heißt zeitinvariant wenn

$$S(f_{.-\hat{k}}) \quad = \quad S(f)_{.-\hat{k}}$$

für alle Folgen f.

Genau wie die Linearität lässt sich auch die Zeitinvarianz durch ein kommutatives Diagramm darstellen:



Linearität und Zeitinvarianz wird oft mit LTI (linear time invariant) abgekürzt.

Alle o.g. Systeme sind zeitinvariant. Nicht zeitinvariant ist z.B.

$$[S(f)]_k \quad = \quad kf_k$$

Es gilt

$$\begin{aligned} \left[S(f_{.-\hat{k}})\right]_k &= k(f_{.-\hat{k}})_k \\ &= kf_{k-\hat{k}} \\ \left[S(f)_{.-\hat{k}}\right]_k &= [S(f)]_{k-\hat{k}} \\ &= (k-\hat{k})f_{k-\hat{k}}. \end{aligned}$$

Der eigentliche Grund weshalb wir uns mit der Faltung beschäftigen besteht darin, dass jedes LTI System S durch eine Faltung mit der Impulsantwort von S berechnet werden kann.

Theorem 1.24
Sei S ein LTI System mit Impulsantwort $S(\delta) = g$. Dann gilt

$$S(f) \quad = \quad f * g$$

für jede Folge f.

Beweis. Ausgehend von

$$f * g \quad = \quad \sum_{\ell} f_\ell g_{\cdot-\ell}$$

erhält man mit $g = \delta$

$$\begin{aligned} f &= f * \delta \\ &= \sum_{\ell} f_\ell \delta_{\cdot-\ell}. \end{aligned}$$

Folglich ist

$$\begin{aligned} S(f) &= S\left(\sum_{\ell} f_\ell \delta_{\cdot-\ell}\right) \\ &= \sum_{\ell} S\left(f_\ell \delta_{\cdot-\ell}\right) && \text{erste Linearitätsbedingung von } S \\ &= \sum_{\ell} f_\ell S(\delta_{\cdot-\ell}) && \text{zweite Linearitätsbedingung von } S \\ &= \sum_{\ell} f_\ell \left[S(\delta)\right]_{\cdot-\ell} && \text{Zeitinvarianz von } S \\ &= \sum_{\ell} f_\ell g_{\cdot-\ell} \\ &= f * g. \end{aligned}$$

In Umkehrung zum vorigen Theorem ist auch jedes System, das durch eine Faltung mit einer Folge g berechnet werden kann, linear und zeitinvariant. Damit ist die Klasse der LTI Systeme identisch mit der Klasse der Systeme, die durch eine Faltung berechnet werden können.

Theorem 1.25
Sei $g \in \mathbb{Z} \to \mathbb{R}$ eine Folge und S definiert durch

$$S(f) = f * g$$

für jede Folge $f \in \mathbb{Z} \to \mathbb{R}$. Dann ist S linear und zeitinvariant.

Beweis. Der Beweis folgt direkt aus den Eigenschaften der Faltung, siehe Theorem 1.18.

$$\begin{aligned} S(f^{(1)} + f^{(2)}) &= (f^{(1)} + f^{(2)}) * g \\ &= f^{(1)} * g + f^{(2)} * g \\ &= S(f^{(1)}) + S(f^{(2)}) \\ S(uf) &= (uf) * g \\ &= u(f * g) \\ &= uS(f) \\ \left[S(f_{.-\hat{k}})\right] &= f_{.-\hat{k}} * g \\ &= (f * g)_{.-\hat{k}} \\ &= S(f)_{.-\hat{k}}. \end{aligned}$$

Theorem 1.24 und 1.25 stellen den eins zu eins Zusammenhang zwischen LTI Systemen und der Faltung her.

Theorem 1.26
Ein System S ist linear und zeitinvariant genau dann wenn

$$S(f) = f * S(\delta)$$

für jede Folge f.

Definition 1.27 Kausales System
Ein lineares zeitinvariantes System S heißt kausal wenn

$$S(\delta)_k = 0 \text{ für alle } k < 0.$$

Kausalität ist eine Eigenschaft, die man oft stillschweigend voraussetzt. Anschaulich bedeutet Kausalität, dass der Output zum Zeitpunkt k nur von den Inputwerten zu Zeitpunkten $\leq k$ abhängt:

$$\begin{aligned} S(f)_k &= (f * S(\delta))_k \\ &= \sum_{\ell=-\infty}^{\infty} f_\ell S(\delta)_{k-\ell} \\ &= \sum_{\ell=-\infty}^{k} f_\ell S(\delta)_{k-\ell} \quad \text{da } S(\delta)_{k-\ell} = 0 \text{ für } \ell > k. \end{aligned}$$

Ein nicht kausales System würde bereits Outputwerte erzeugen noch bevor man den zugehörigen Input anlegt, was technisch natürlich nicht realisierbar ist.

Nichtkausale System sind trotzdem aus mehreren Gründen interessant. Einerseits muss k keine Zeitvariable sein sondern kann auch einen Ort beschreiben wie z.B. in der Bildverarbeitung. Andererseits treten nichtkausale Systeme auch dann auf, wenn k ein Zeitvariable ist wie z.B. beim idealen Tiefpassfilter. Für die Praxis stellt sich dann die Aufgabe, so ein System mit möglichst kleinem Fehler durch ein kausales System zu approximieren.

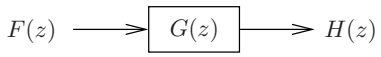
1.8 Übertragungsfunktion

Sei S ein LTI System mit Impulsantwort $S(\delta) = g$. Das System bildet somit eine Folge f auf eine Folge

$$h \quad = \quad f * g$$

ab. Mit der z-Transformation und dem Faltungssatz erhält man

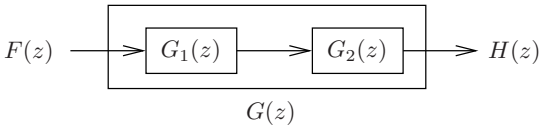
$$H(z) \quad = \quad F(z)G(z).$$



Im Bildbereich bewirkt ein LTI System somit eine simple Multiplikation mit $G(z)$. Die Funktion $G(z)$, d.h. die z-Transformierte der Impulsantwort g wird daher Übertragungsfunktion des Systems genannt.

In der Praxis spielt die Zusammenschaltung von LTI Systemen eine wichtige Rolle. Mit Hilfe der z-Transformation lässt sich die Übertragungsfunktion eines komplexen Systems einfach berechnen. Für die Hintereinanderschaltung zweier Systeme mit Übertragungsfunktionen $G_1(z)$ und $G_2(z)$ gilt

$$\begin{aligned} H(z) \quad &= \quad \big(F(z)G_1(z)\big)G_2(z) \\ &= \quad (G_1(z)G_2(z))F(z). \end{aligned}$$



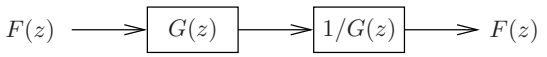
Die Übertragungsfunktion des Gesamtsystems ist somit

$$G(z) \quad = \quad G_1(z)G_2(z).$$

Die Impulsantwort g des Gesamtsystems erhält man daher aus den Impulsantworten $g^{(1)}$ und $g^{(2)}$ der hintereinandergeschalteten Systeme durch Faltung

$$g \quad = \quad g^{(1)} * g^{(2)}.$$

Unmittelbare Anwendung finden diese Überlegungen wenn man ein Signal f durch einen gestörten Kanal mit Übertragungsfunktion $G(z)$ überträgt und beim Empfänger das Originalsignal f rekonstruieren möchte. Dies erreicht man durch ein nachgeschaltetes System mit Übertragungsfunktion $1/G(z)$.



Leider besitzt das System $1/G(z)$ in der Regel keine inverse z-Transformierte und kann daher im Zeitbereich nicht realisiert werden. So ist z.B. leicht verständlich, dass ein Verzögerungsglied nicht ausgeglichen werden kann, da das zugehörige inverse System "in die Zukunft schauen" müsste. Mit geeigneten Approximationen und Vereinfachungen lassen sich jedoch häufig zumindest näherungsweise inverse Systeme zur Kanalkompensation realisieren.

1.9 Frequenzantwort

Eine wichtige Eigenschaft von LTI Systemen ist ihr Antwortverhalten auf Schwingungen. Ist das Eingangssignal f_k eine Schwingung mit Frequenz $\hat{\omega}$, dann ist auch das Ausgangssignal eine Schwingung mit *gleicher* Frequenz $\hat{\omega}$. Das System bewirkt also allenfalls eine Verstärkung und Phasenverschiebung.

Sei

$$f_k \quad = \quad e^{j\hat{\omega}k}$$

eine komplexe Schwingung und g_k die Impulsantwort des Systems. Dann berechnet sich die Systemantwort h_k wie folgt.

$$\begin{aligned}
h_k &= (f * g)_k \\
&= \sum_{\ell=-\infty}^{\infty} g_\ell f_{k-\ell} \\
&= \sum_{\ell=-\infty}^{\infty} g_\ell e^{j\hat{\omega}(k-\ell)} \\
&= \sum_{\ell=-\infty}^{\infty} g_\ell e^{j\hat{\omega}k} e^{-j\hat{\omega}\ell} \\
&= e^{j\hat{\omega}k} \sum_{\ell=-\infty}^{\infty} g_\ell e^{-j\hat{\omega}\ell} \\
&= f_k \sum_{\ell=-\infty}^{\infty} g_\ell \left(e^{j\hat{\omega}}\right)^{-\ell} \\
&= f_k \sum_{\ell=-\infty}^{\infty} g_\ell z^{-\ell} \qquad \text{für } z = e^{j\hat{\omega}} \\
&= f_k G(z) \\
&= f_k G(e^{j\hat{\omega}}).
\end{aligned}$$

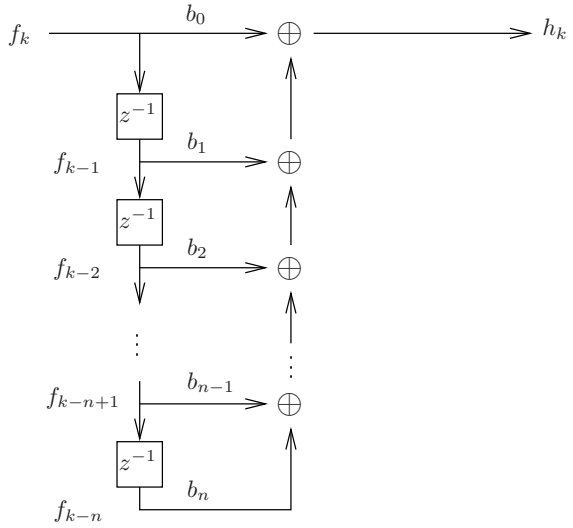
Die Schwingung f_k wird somit nur mit dem konstanten Faktor $G(e^{j\hat{\omega}})$ multipliziert, wobei G die Übertragungsfunktion des Systems ist. Da dieser Faktor von der Frequenz $\hat{\omega}$ abhängt, bewirkt ein LTI System eine frequenzabhängige Verstärkung bzw. Phasenverschiebung. Man nennt $G(e^{j\hat{\omega}})$ daher Frequenzantwort des LTI Systems. Der Faktor spielt eine entscheidende Rolle beim Entwurf von digitalen Filtern wie z.B. einem Tiefpass, die bestimmte Frequenzbereiche unterdrücken oder verstärken sollen.

1.10 Digitale Filter

Die Faltung mit einer endlichen Folge

$$b = \langle b_0, b_1, \ldots, b_n \rangle$$

lässt sich technisch durch eine Zusammenschaltung von Verzögerungsgliedern z^{-1}, Multiplizierern und Addierern realisieren. Man nennt eine solche Schaltung auch digitalen Filter. Zum Zeitpunkt k liegt am Filtereingang der Abtastwert f_k, am Ausgang erhält man h_k. Aufgrund der Verzögerungsglieder werden im Filter die n vergangenen Abtastwerte $f_{k-1}, \ldots, f_{k-n}$ gespeichert.



Wie aus dem Bild ersichtlich gilt

$$\begin{aligned} h_k &= b_0 f_k + b_1 f_{k-1} + \ldots + b_n f_{k-n} \\ &= (b * f)_k. \end{aligned}$$

Im Bildbereich gilt

$$H(z) = B(z)F(z)$$

so dass die Übertragunsfunktion des Filters

$$B(z) = b_0 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_n z^{-n}$$

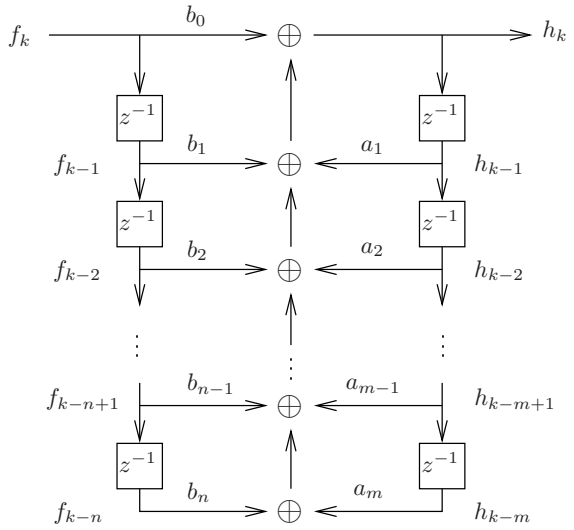
ist.

Die Impulsantwort des Filters ist

$$b * \delta = b$$

und somit eine endliche Folge. Man nennt solche Filter daher auch FIR Filter (finite impulse response).

Durch eine einfache Erweiterung lassen sich mit endlichem Hardwareaufwand auch Filter mit unendlicher Impulsantwort realisieren. Hierzu wird der Filterausgang durch eine Kette von Verzögerungsgliedern geschickt, mit Gewichten $a_1, a_2, \ldots, a_m$ multipliziert und zurückgeführt. Man spricht daher auch von rekursiven oder IIR Filtern (infinite impulse response).



Wie aus dem Bild ersichtlich ist der Filterausgang zum Zeitpunkt k

$$h_k = b_0 f_k + b_1 f_{k-1} + \ldots + b_n f_{k-n} + a_1 h_{k-1} + a_2 h_{k-2} + \ldots + a_m h_{k-m}.$$

Zur Berechnung der Übertragungsfunktion dieses Filters bringt man zunächst alle h-Terme auf eine Seite:

$$h_k - a_1 h_{k-1} - a_2 h_{k-2} - \ldots - a_m h_{k-m} = b_0 f_k + b_1 f_{k-1} + \ldots + b_n f_{k-n}.$$

Die Terme auf beiden Seiten kann man durch Faltungen ausdrücken:

$$(\langle 1, -a_1, -a_2, \ldots, -a_m \rangle * h)_k = (\langle b_0, b_1, \ldots, b_n \rangle * f)_k$$

Mit

$$a = \langle 1, -a_1, -a_2, \ldots, -a_m \rangle$$

gilt damit

$$a * h = b * f$$

bzw.

$$\begin{aligned} A(z)H(z) &= B(z)F(z) \\ H(z) &= \frac{B(z)}{A(z)}F(z). \end{aligned}$$

Die Übertragungsfunktion des rekursiven Filters ist somit

$$G(z) \quad = \quad \frac{B(z)}{A(z)}.$$

Die Impulsantwort g ist per Definition die inverse z-Transformierte der Übertragungsfunktion $G(z)$. Da

$$G(z) \quad = \quad \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_n z^{-n}}{1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_m z^{-m}}$$

lässt sich g_k mit Polynomdivision und Partialbruchzerlegung berechnen, siehe Kapitel 1.11. Als weitere Alternative kann man auch einen rekursiven Filter aus den Koeffizienten des Zähler- und Nennerpolynoms von $G(z)$ konstruieren und die Impulsantwort im Zeitbereich bestimmen. Aus

$$h_k \quad = \quad b_0 f_k + b_1 f_{k-1} + \ldots + b_n f_{k-n} + a_1 h_{k-1} + a_2 h_{k-2} + \ldots + a_m h_{k-m}$$

erhält man mit $f = \delta$ die Impulsantwort

$$g_k \quad = \quad b_0 \delta_k + b_1 \delta_{k-1} + \ldots + b_n \delta_{k-n} + a_1 g_{k-1} + a_2 g_{k-2} + \ldots + a_m g_{k-m}.$$

Diese lässt sich nun sukzessive für $k = 0, 1, 2, \ldots$ bestimmen.

$$\begin{aligned}
g_0 &= b_0 \\
g_1 &= b_1 + a_1 g_0 \\
&= b_1 + a_1 b_0 \\
g_2 &= b_2 + a_1 g_1 + a_2 g_0 \\
&= b_2 + a_1 (b_1 + a_1 b_0) + a_2 b_0 \\
g_3 &= b_3 + a_1 g_2 + a_2 g_1 + a_3 g_0 \\
&= b_3 + a_1 (b_2 + a_1 (b_1 + a_1 b_0) + a_2 b_0) + a_2 (b_1 + a_1 b_0) + a_3 b_0 \\
&\vdots
\end{aligned}$$

1.11 Inverse z-Transformation

Wie in Kapitel 1.10 gezeigt, hat die z-Transformierte eines rekursiven Filters die Form

$$F(z) = \frac{p_0 + p_1 z^{-1} + p_2 z^{-2} + \ldots + p_n z^{-n}}{q_0 + q_1 z^{-1} + q_2 z^{-2} + \ldots + q_m z^{-m}}, \quad q_0 = 1.$$

Ohne Beschränkung der Allgemeinheit nehmen wir an, dass $n < m$, da man sonst mit Polynomdivision einen ganzrationalen Teil in z^{-1} abspalten kann. Ein Polynom in z^{-1} in den Zeitbereich zu transformieren, ist trivial, da man nur die Koeffizienten ablesen muss.

Bei der Rücktransformation in den Zeitbereich ist eine Folge f_k gesucht, mit

$$f_k \quad \circ\!\!-\!\!\bullet \quad F(z).$$

Hierfür gibt es zwei Methoden:

Partialbruchzerlegung. Indem man Zähler und Nenner mit z^m erweitert, kann man $F(z)$ als rationale Funktion in z darstellen.

$$F(z) = \frac{p_0 z^m + p_1 z^{m-1} + p_2 z^{m-2} + \ldots + p_n z^{m-n}}{q_0 z^m + q_1 z^{m-1} + q_2 z^{m-2} + \ldots + q_m}, \quad q_0 = 1.$$

Nach Abspalten des ganzrationalen Teils p_0/q_0 mit Polynomdivision kann man den gebrochenen Rest durch Partialbruchzerlegung als gewichtete Summe von einfachen Termen der Form

$$\frac{1}{(z-a)^n}$$

darstellen, die mit Hilfe der Korrespondenz

$$\frac{1}{(z-a)^n} \quad \bullet\!\!-\!\!\circ \quad a^{k-n}\binom{k-1}{n-1}$$

in den Zeitbereich zurücktransformiert werden können. Die hierbei auftretenden Bionmialkoeffizienten sind definiert durch

$$\binom{a}{b} \quad = \quad \begin{cases} \frac{a!}{b!(a-b)!} & \text{falls } 0 \le b \le a \\ 0 & \text{sonst} \end{cases}$$

wobei a, b ganze Zahlen sind.

Die Herleitung dieser Korrespondenz geschieht über Induktion.

- Für $n = 1$ wurde in Beispiel 1.10 gezeigt, dass

$$\frac{1}{z-a} \quad \bullet\!\!-\!\!\circ \quad \sigma_{k-1} a^{k-1} = a^{k-1}\binom{k-1}{0}$$

- Wir nehmen nun an, dass für ein festes n gezeigt wurde, dass

$$\frac{1}{(z-a)^n} \quad \bullet\!\!-\!\!\circ \quad a^{k-n}\binom{k-1}{n-1}.$$

- Zu zeigen ist, dass dies auch für $n+1$ gilt, d.h.

$$\frac{1}{(z-a)^{n+1}} \quad \bullet\!\!-\!\!\circ \quad a^{k-n-1}\binom{k-1}{n}.$$

Umformen im Zeitbereich und Anwenden der Ableitung im Bildbereich ergibt

$$\begin{aligned}
& a^{k-n-1}\binom{k-1}{n} \\
= \quad & \frac{1}{a}a^{k-n}\binom{k-1}{n-1}\frac{k-n}{n} \\
= \quad & \frac{1}{na}\left(k\underbrace{a^{k-n}\binom{k-1}{n-1}}_{\text{Ind.Hyp.}} - n\underbrace{a^{k-n}\binom{k-1}{n-1}}_{\text{Ind.Hyp.}}\right) \\
\circ\!\!-\!\!\bullet \quad & \frac{1}{na}\left(-z\left(\frac{1}{(z-a)^n}\right)' - n\frac{1}{(z-a)^n}\right) \\
= \quad & \frac{1}{na}\left(-z\left(\frac{-n}{(z-a)^{n+1}}\right) - \frac{n(z-a)}{(z-a)^{n+1}}\right) \\
= \quad & \frac{1}{na}\left(\frac{nz-nz+na}{(z-a)^{n+1}}\right) \\
= \quad & \frac{1}{(z-a)^{n+1}}.
\end{aligned}$$

Polynomdivision. Die zweite Möglichkeit, $F(z)$ in den Zeitbereich zurückzutransformieren geschieht über Polynomdivision. Im Gegensatz zur Methode mit der Partialbruchzerlegung erhält man hierbei keinen Term für f_k sondern sukzessive Werte für $f_0, f_1, \ldots$.

Beispiel 1.28

$$\begin{array}{rl}
z^{-1} \;:\; 1-2z^{-1}+z^{-2} & = z^{-1}+2z^{-2}+3z^{-3}+\ldots \\
\underline{z^{-1}-2z^{-2}+z^{-3}} & \\
\quad 2z^{-2}-z^{-3} & \\
\quad \underline{2z^{-2}-4z^{-3}+2z^{-4}} & \\
\quad\quad 3z^{-3}-2z^{-4} & \\
\quad\quad \underline{3z^{-3}-6z^{-4}+3z^{-5}} & \\
\quad\quad\quad 4z^{-4}-3z^{-5} & \\
\quad\quad\quad\quad \ldots &
\end{array}$$

Damit ist

$$f_0 = 0, \quad f_1 = 1, \quad f_2 = 2, \quad f_3 = 3, \ldots$$

2 Lineare DGL Systeme mit konstanten Koeffizienten

2.1 Beispiel: Modellierung eines Räuber Beute Systems

In einem Ökosystem werden zwei Tierarten beobachtet. Die eine Tierart sind Räuber, deren Anzahl zum Zeitpunkt t wird mit $x(t)$ beschrieben wird. Die andere Tierart sind Beutetiere, deren Anzahl zum Zeitpunkt t mit $y(t)$ beschrieben wird.

Zunächst wird qualitativ überlegt, wie sich der Bestand der Tierarten in einem kurzen Zeitintervall $[t, t + \Delta t]$ ändert. Falls Δt klein ist, kommen in dem Intervall wenige Tiere dazu und es kann vernachlässigt werden, dass auch dies Tiere wieder Junge bekommen. Daher darf für kleine Δt näherungsweise angenommen werden, dass die Zunahme in dem Intervall proportional zu Dauer Δt des Intervalls ist.

Änderung der Räuber. Wenn es keine Beutetiere geben würde, dann wäre die Anzahl Räuber, die im Zeitintervall $[t, t+\Delta t]$ sterben, proportional zu ihrer Anzahl, d.h.

$$x(t + \Delta t) - x(t) = ax(t)\Delta t, \quad a < 0.$$

Andererseits ist die Anzahl neu geborenen Räuber proportional zur Anzahl bereits vorhandener Räuber, und zur Anzahl von Beutetieren. Damit gilt

$$x(t + \Delta t) - x(t) = \big(ax(t) + bx(t)y(t)\big)\Delta t, \quad a < 0, b > 0.$$

Änderung der Beutetiere. Ohne Räuber würden sich die Beutetiere proportional zu ihrer Anzahl vermehren, d.h.

$$y(t + \Delta t) - y(t) = cy(t)\Delta t, \quad c > 0.$$

Die von den Räubern gefressenen Beutetiere sind im eingeschwungenen Zustand proportional zur Anzahl der Räuber und proportional zur Anzahl der Beutetiere. Damit gilt

$$y(t + \Delta t) - y(t) = \big(cy(t) + dx(t)y(t)\big)\Delta t, \quad c > 0, d < 0.$$

Nach Division durch Δt erhält man

$$\begin{aligned} \frac{x(t + \Delta t) - x(t)}{\Delta t} &= ax(t) + bx(t)y(t) \\ \frac{y(t + \Delta t) - y(t)}{\Delta t} &= cy(t) + dx(t)y(t). \end{aligned}$$

Die Betrachtungen galten nur für ein sehr kurzes Zeitintervall der Länge Δt. Führt man den Grenzübergang $\Delta t \to 0$ durch, entstehen auf der linken Seite Ableitungen:

$$\begin{aligned} x'(t) &= ax(t) + bx(t)y(t) \\ y'(t) &= cy(t) + dx(t)y(t). \end{aligned}$$

Das so entstehende System von Differentialgleichungen ist nach seinen Entdeckern Volterra und Lotka benannt. Da Produkte $x(t)y(t)$ auftreten, handelt es sich um ein nichtlineares DGL System. Trotzdem kann man es z.B. mit der Methode des Eulerschen Polygonzugs zumindest näherungsweise lösen. Umformen ergibt für kleine Werte von Δt

$$\begin{aligned} x(t+\Delta t) &= x(t) + \big(ax(t) + bx(t)y(t)\big)\Delta t \\ y(t+\Delta t) &= y(t) + \big(cy(t) + dx(t)y(t)\big)\Delta t. \end{aligned}$$

Kennt man also $x(t)$ und $y(t)$ zum einem bestimmten Zeitpunkt t, dann kann man mit diesen Gleichungen $x(t+\Delta t)$ und $y(t+\Delta t)$ näherungsweise berechnen. Sind die Anfangswerte $x(0)$ und $y(0)$ bekannt, kann man somit $x(n\Delta t)$ und $y(n\Delta t)$ für $n = 1, 2, \ldots$ bestimmen, siehe Bild 2.1 oben. Man sieht, dass $x(t)$ und $y(t)$ periodisch sind mit der selben Periodendauer. Dabei hinkt die Anzahl der Räuber etwas hinter der Anzahl Beutetiere hinterher, was auch verständlich ist: Zuerst müssen mehr Beutetiere vorhanden sein, dann steigt die Anzahl der Räuber an.

Besonders interessant ist die Darstellung, wenn man $(x(t), y(t))$ als Punkt in einem Koordinatensystem einzeichnet und die Bewegung dieses Punktes über der Zeit beobachtet. Hierbei entsteht beim Volterra Lotka System immer eine geschlossene Kurve da $x(t)$ und $y(t)$ periodisch sind mit gemeinsamer Periodendauer, siehe Bild 2.1 unten.

Interessant ist auch zu untersuchen, ob es für das Ökosystem einen stationären Zustand gibt, d.h. eine Anzahl von Räuber- und Beutetieren so dass ihr Bestand sich zeitlich nicht ändert. Die Forderung nach einem zeitlich konstanten Bestand bedeutet

$$x'(t) = 0 \text{ und } y'(t) = 0 \text{ für alle } t.$$

Eingesetzt in die DGL erhält man

$$\begin{aligned} ax(t) + bx(t)y(t) &= 0 \\ cy(t) + dx(t)y(t) &= 0. \end{aligned}$$

Eine offensichtliche Lösung ist

$$x(t) = 0 \text{ und } y(t) = 0 \text{ für alle } t.$$

Klar, wenn es überhaupt keine Tiere gibt, bleibt ihre Anzahl konstant. Wird also zusätzlich $x(t) \neq 0$ und $y(t) \neq 0$ gefordert, kann man die erste Gleichung durch $x(t)$ dividieren und die zweite durch $y(t)$ und erhält

$$\begin{aligned} a + by(t) &= 0 \\ c + dx(t) &= 0. \end{aligned}$$

Die zeitkonstanten Lösungsfunktionen sind somit

$$\begin{aligned} x(t) &= -c/d \\ y(t) &= -a/b. \end{aligned}$$

Wählt man die Anfangspopulation $x(0) = -c/d$ und $y(0) = -a/c$, dann wird diese für alle t gleich bleiben.

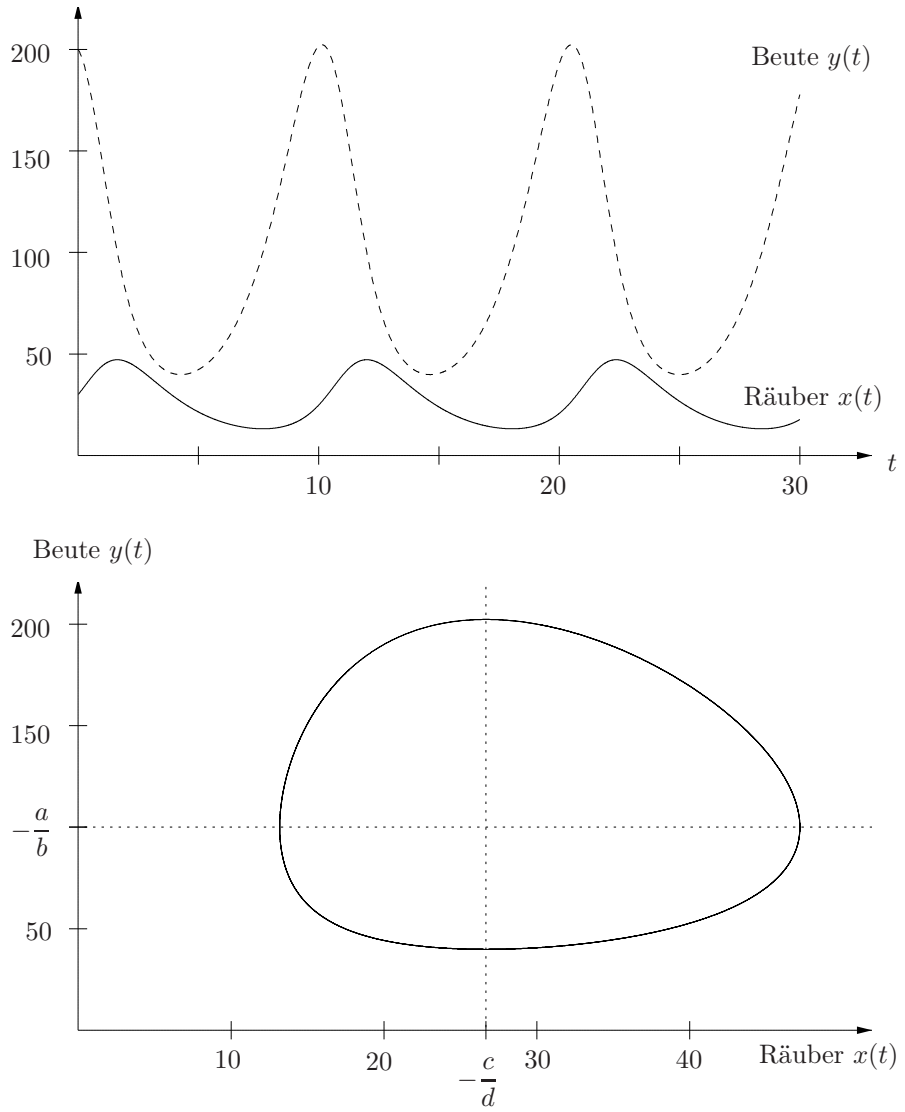


Abbildung 2.1: Differentialgleichungssystem nach Volterra und Lotka mit den Parametern $a = -0.5, b = 0.005, c = 0.8, d = -0.03$. Die Anfangswerte sind $x(0) = 30$ und $y(0) = 200$. Der Gleichgewichtszustand ist $x(t) = -c/d = 26.67$ und $y(t) = -a/b = 100$.

2.2 Determinanten

Die Lösung von linearen DGL Systemen lässt sich mit Hilfe der Eigenwerte und Eigenvektoren von Matrizen ausdrücken. Diese wiederum können mit Hilfe der Determinante einer Matrix berechnet werden.

Berechnung der Determinante. Die Determinante einer $n \times n$ Matrix A wird mit $\det(A)$ bezeichnet und ist eine Zahl, die wie folgt definiert ist. Zunächst zwei Hilfsgrößen:

$$s_{ij} \quad = \quad (-1)^{i+j}.$$

Man kann sich s_{ij} mit Hilfe eines Schachbretts vorstellen, wobei i, j die Koordinaten sind: Die weißen Felder entsprechen 1, die schwarzen -1.

Streicht man aus der $n \times n$ Matrix A die i-te Zeile und j-te Spalte, entsteht eine $(n-1) \times (n-1)$ Matrix. Diese wird mit $A^{(i,j)}$ bezeichnet. Und nun zur Determinante.

- Der Spezialfall $n = 1$ ist sehr einfach.

$$\det(a_{11}) = a_{11}$$

- Im allgemeinen Fall $n > 1$ hat man mehrere Möglichkeiten, die Determinante zu berechnen.
  - Man kann sich eine belieibge Zeile i aussuchen und die Determinante nach dieser Zeile entwickeln:

$$\det(A) = \sum_{j=1}^{n} s_{ij} a_{ij} \det(A^{(i,j)})$$

  - Genauso kann man sich eine beliebige Spalte j aussuchen und die Determinante nach dieser Spalte entwickeln:

$$\det(A) = \sum_{i=1}^{n} s_{ij} a_{ij} \det(A^{(i,j)})$$

  In allen Fällen kommt das gleiche Ergebnis heraus.

Die Berechnunt einer $n \times n$ Determinante erfordert somit die Berechnung von n Determinanten von $(n-1) \times (n-1)$ Matrizen. Die Sache ist also teuer, da der Aufwand mit $n!$ wächst. Sinnvollerweise sucht man sich zur Entwicklung eine Zeile oder Spalte mit vielen Nullen, da Summanden mit $a_{ij} = 0$ wegfallen und für diese $\det(A^{(i,j)})$ nicht berechnet werden muss.

Für den Spezialfall $n = 2$ gibt es eine Formel, die man sich leicht merken kann. Man berechnet das Produkt der Diagonalelemente minus das Produkt der Gegendiagonalelemente.

$$\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

Eigenschaften der Determinante.

- $\det(A) = 0$ genau dann wenn A singulär (d.h. nicht invertierbar) ist.
- $\det(AB) = \det(A)\det(B)$
- $\det(cA) = c^n \det(A)$
- $\det(A^{\mathrm{T}}) = \det(A)$
- $\det(A^{-1}) = \det(A)^{-1}$
- Cramersche Regel. Die Lösung des LGS $A\vec{x} = \vec{b}$ kann berechnet werden durch

$$x_i = \frac{\det(A_i^{\vec{b}})}{\det(A)}$$

  wobei $A_i^{\vec{b}}$ die Matrix ist, die man erhält wenn man die i-te Spalte von A durch $\vec{b}$ ersetzt.

Berechnung der Determinante mit dem Gauß Algorithmus Mit dem Gauß Algorithmus wird eine Matrix durch eine Folge von Umformungsschritten auf Dreiecksform gebracht. Die Determinante einer Dreiecksmatrix ist das Produkt ihrer Diagonalelemente und lässt sich somit schnell berechnen. Die Umformungsschritte beim Gauß Algorithmus sind wie folgt:

- Vertauschen von zwei Zeilen. Hierbei kehrt sich das Vorzeichen der Determinante um.
- Addition des Vielfachen einer Zeile zu einer anderen Zeile. Hierbei ändert sich die Determinante nicht.
- Multiplikation einer Zeile mit einem Faktor c. Hierbei wird die Determinante mit Faktor c multipliziert.

Man kann somit aus der leicht zu berechnenden Determinante der Dreiecksmatrix auf die Determinante der ursprünglichen Matrix zurückrechnen. Der Aufwand für den Gauß Algorithmus wächst nur mit n^3 und ist daher viel effizienter als die Entwicklung nach einer Zeile oder Spalte.

2.3 Eigenwerte und Eigenvektoren

Definition 2.1 (Eigenwert, Eigenvektor)
Ein Vektor $\vec{v} \in \mathbb{C}^n$, $\vec{v} \neq \vec{0}$ heißt Eigenvektor von $A \in \mathbb{R}^{n\times n}$ mit Eigenwert λ wenn

$$A\vec{v} \quad = \quad \lambda\vec{v}$$

Weiterhin heißt λ Eigenwert von A wenn A einen Eigenvektor mit Eigenwert λ hat.

Der Nullvektor $\vec{0}$ ist somit per Definition kein Eigenvektor von A, obwohl

$$A\vec{0} \quad = \quad \lambda\vec{0}$$

für jedes λ gilt.

Beispiel 2.2 Sei

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

Dann ist

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

Eigenvektor von A mit Eigenwert $\lambda_1 = 2$ da

$$A\vec{v}_1 \quad = \quad \begin{pmatrix} 1 & 1 \\ 3 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad = \quad \begin{pmatrix} 2 \\ 2 \end{pmatrix} \quad = \quad 2\vec{v}.$$

Weiterhin ist

$$\vec{v} = \begin{pmatrix} 1 \\ -3 \end{pmatrix}$$

Eigenvektor von A mit Eigenwert $\lambda_2 = -2$ da

$$A\vec{v}_2 \quad = \quad \begin{pmatrix} 1 & 1 \\ 3 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ -3 \end{pmatrix} \quad = \quad \begin{pmatrix} -2 \\ 6 \end{pmatrix} \quad = \quad -2\vec{v}.$$

Definition 2.3 (Eigenraum)
Sei λ Eigenwert von A. Der Eigenraum von A zum Eigenwert λ ist definiert durch

$$E_\lambda \quad = \quad \{\vec{v} \,|\, A\vec{v} = \lambda\vec{v}\}.$$

Der Eigenraum E_λ ist somit die Menge aller Eigenvektoren von A zum Eigenwert λ und dem Nullvektor.

Theorem 2.4
Für jeden Eigenwert λ von A ist der Eigenraum E_λ abgeschlossen unter Addition und skalarer Multiplikation und bildet daher einen Vektorraum.

Beweis.

- Abgeschlossenheit unter Addition. Seien

$$\vec{v}_1, \vec{v}_2 \in E_\lambda,$$

d.h.

$$A\vec{v}_1 = \lambda\vec{v}_1, \qquad A\vec{v}_2 = \lambda\vec{v}_2.$$

Dann ist

$$\begin{aligned} A(\vec{v}_1 + \vec{v}_2) &= A\vec{v}_1 + A\vec{v}_2 \\ &= \lambda\vec{v}_1 + \lambda\vec{v}_2 \\ &= \lambda(\vec{v}_1 + \vec{v}_2). \end{aligned}$$

Damit ist

$$\vec{v}_1 + \vec{v}_2 \in E_\lambda.$$

- Abgeschlossenheit unter skalarer Multiplikation. Sei

$$\vec{v} \in E_\lambda \text{ d.h. } A\vec{v} = \lambda\vec{v}$$

und $u \in \mathbb{C}$. Dann gilt

$$\begin{aligned} A(u\vec{v}) &= u(A\vec{v}) \\ &= u(\lambda\vec{v}) \\ &= \lambda(u\vec{v}). \end{aligned}$$

Damit ist

$$u\vec{v} \in E_\lambda.$$

Der Eigenwert $\lambda = 0$ kann durchaus auftreten, nämlich immer dann wenn A singulär ist.

Beispiel 2.5 Sei

$$A = \begin{pmatrix} 1 & 3 \\ 2 & 6 \end{pmatrix}.$$

Die zweite Spalte ist das dreifache der ersten Spalte, d.h. die Spalten sind linear abhängig und somit A singulär. Entsprechend ist

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

bzw.

$$A\vec{v} = \vec{0} = 0\vec{v} \text{ für } \vec{v} = \begin{pmatrix} 3 \\ -1 \end{pmatrix}.$$

Damit ist $\vec{v}$ Eigenvektor von A zum Eigenwert $\lambda = 0$.

Berechnung der Eigenwerte und Eigenvektoren. Für eine gegebene Matrix $A \in \mathbb{R}^{n\times n}$ sollen die Eigenvektoren und ihre Eigenwerte berechnet werden. Gesucht sind also $\vec{v} \neq \vec{0}$ und λ so dass

$$A\vec{v} = \lambda\vec{v}.$$

Umformen ergibt

$$\begin{aligned} A\vec{v} - \lambda\vec{v} &= \vec{0} \\ A\vec{v} - \lambda E\vec{v} &= \vec{0} \\ (A - \lambda E)\vec{v} &= \vec{0}. \end{aligned}$$

Um $\vec{v}$ ausklammern zu können musste die Einheitsmatrix E im zweiten Schritt eingefügt werden. Es entsteht somit ein homogenes lineares Gleichungssystem mit Koeffizientenmatrix

$$A - \lambda E.$$

Gesucht ist eine nichttriviale Lösung $\vec{v} \neq \vec{0}$. Eine solche Lösung kann nur existieren wenn $A - \lambda E$ singulär ist, d.h.

$$\det(A - \lambda E) = 0.$$

Für $A \in \mathbb{R}^{n\times n}$ ist $\det(A - \lambda E)$ ein Polynom vom Grad n von λ und heißt charakteristisches Polynom von A. Die Nullstellen λ_i dieses Polynoms sind die Eigenwerte von A. Aus dem Fundamentalsatz der Algebra folgt somit, dass A höchstens n unterschiedliche Eigenwerte haben kann. Die Vielfachheit einer Nullstelle λ_i des Polynoms heißt *algebraische Vielfachheit* des Eigenwerts λ_i.

Hat man die Eigenwerte λ_i berechnet, erhält man die Eigenvektoren $\vec{v}_i$ zu einem Eigenwert λ_i durch Lösen des homogenen singulären LGS

$$(A - \lambda_i E)\vec{v}_i = \vec{0}.$$

Die Lösungsmenge dieses LGS bildet einen Vektorraum. In der Regel ist dies eine Ursprungsgerade, d.h. ein eindimensionaler Vektorraum. Die Dimension kann aber durchaus größer sein und heißt *geometrische Vielfachheit* des Eigenwerts λ_i.

Die algebraische Vielfachheit ist immer größer oder gleich der geometrischen Vielfachheit. Ist λ_i also einfache Nullstelle des charakteristischen Polynoms (d.h. der Normalfall), dann ist auch der zugehörige Eigenraum eindimensional und wird von einem einzigen Basisvektor aufgespannt.

Ohne Beweis sei noch ein interessanter Zusammenhang zwische den Eigenwerten und der Determinante einer Matrix erwähnt: Hat A die Eigenwerte λ_i mit algebraischer Vielfachheit r_i dann gilt

$$\det(A) = \prod_i \lambda_i^{r_i}$$

Beispiel 2.6 Sei

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

Damit ist

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

Das charakteristische Polynom ist

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

Die Nullstellen des charakteristischen Polynoms liefern die Eigenwerte

$$\lambda_1 = 2 \text{ und } \lambda_2 = -2.$$

- Eigenvektoren zum Eigenwert $\lambda_1 = 2$.

  $$A - \lambda_1 E = \begin{pmatrix} -1 & 1 \\ 3 & -3 \end{pmatrix}.$$

  Die Lösungsmenge von

  $$(A - \lambda_1 E)\vec{v} = \vec{0}$$

  ist der Eigenraum

  $$E_2 = \left\{ a \begin{pmatrix} 1 \\ 1 \end{pmatrix} \mid a \in \mathbb{C} \right\}.$$

  Eine Basis dieses Eigenraums ist der Vektor

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

- Eigenvektoren zum Eigenwert $\lambda_2 = -2$.

  $$A - \lambda_2 E = \begin{pmatrix} 3 & 1 \\ 3 & 1 \end{pmatrix}.$$

  Die Lösungsmenge von

  $$(A - \lambda_2 E)\vec{v} = \vec{0}$$

  ist der Eigenraum

  $$E_{-2} = \left\{ a \begin{pmatrix} 1 \\ -3 \end{pmatrix} \mid a \in \mathbb{C} \right\}.$$

  Eine Basis dieses Eigenaums ist der Vektor

  $$\vec{v}_2 = \begin{pmatrix} 1 \\ -3 \end{pmatrix}.$$

Eigenwerte symmetrischer Matrizen, Hauptachsentransformation

Definition 2.7 (Komplexes Skalarprodukt)
Das Skalarprodukt zweier komplexer Vektoren $\vec{x}, \vec{y} \in \mathbb{C}^n$ ist definiert durch

$$\begin{aligned} \vec{x} \circ \vec{y} &= \overline{\vec{x}}^{\mathrm{T}} \vec{y} \\ &= \sum_{i=1}^{n} \overline{x_i} y_i. \end{aligned}$$

Im Vergleich zum reellen Skalarprodukt wird der erste Faktor komplex konjugiert. Dies bewirkt, dass auch das komplexe Skalarprodukt positiv definit ist. Da

$$\begin{aligned} \vec{x} \circ \vec{x} &= \sum_{i=1}^{n} \overline{x_i} x_i \\ &= \sum_{i=1}^{n} |x_i|^2. \end{aligned}$$

ist $\vec{x} \circ \vec{x}$ reell. Weiterhin gilt $\vec{x} \circ \vec{x} \geq 0$ für alle $\vec{x} \in \mathbb{C}^n$ und $\vec{x} \circ \vec{x} = 0$ genau dann wenn $\vec{x} = \vec{0}$.

Theorem 2.8
Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch. Dann sind alle Eigenwerte von A reell.

Beweis. Sei $A \in \mathbb{R}^{n \times n}$ eine symmetrische Matrix und $\vec{v}$ ein Eigenvektor von A mit Eigenwert λ. Dann gilt $\overline{A} = A$, $A^{\mathrm{T}} = A$ und

$$A\vec{v} = \lambda \vec{v}.$$

Nimmt man auf beiden Seiten das Skalarprodukt mit $\vec{v}$ erhält man

$$\begin{aligned} (A\vec{v}) \circ \vec{v} &= (\lambda \vec{v}) \circ \vec{v} \\ \left(\overline{A\vec{v}}\right)^{\mathrm{T}} \vec{v} &= \left(\overline{\lambda \vec{v}}\right)^{\mathrm{T}} \vec{v} \\ \left(A\overline{\vec{v}}\right)^{\mathrm{T}} \vec{v} &= \left(\overline{\lambda}\, \overline{\vec{v}}\right)^{\mathrm{T}} \vec{v} \\ \overline{\vec{v}}^{\mathrm{T}} A^{\mathrm{T}} \vec{v} &= \left(\overline{\lambda}\, \overline{\vec{v}}^{\mathrm{T}}\right) \vec{v} \\ \overline{\vec{v}}^{\mathrm{T}} A \vec{v} &= \overline{\lambda} \left(\overline{\vec{v}}^{\mathrm{T}} \vec{v}\right) \\ \overline{\vec{v}}^{\mathrm{T}} \lambda \vec{v} &= \overline{\lambda} \left(\overline{\vec{v}}^{\mathrm{T}} \vec{v}\right) \\ \lambda \left(\overline{\vec{v}}^{\mathrm{T}} \vec{v}\right) &= \overline{\lambda} \left(\overline{\vec{v}}^{\mathrm{T}} \vec{v}\right) \\ \lambda &= \overline{\lambda}. \end{aligned}$$

Folglich ist $\lambda \in \mathbb{R}$. Im letzten Schritt wurde ausgenutzt, dass $\vec{v} \neq \vec{0}$ da $\vec{v}$ ein Eigenvektor ist.

Wenn die Eigenwerte λ_i reell sind, existieren auch zugehörige *reelle* Eigenvektoren $\vec{v}_i$ als Lösung des singulären, reellen LGS

$$(A - \lambda_i E)\vec{v}_i \quad = \quad \vec{0}.$$

Theorem 2.9
Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix. Dann sind die Eigenvektoren von A zu unterschiedlichen Eigenwerten orthogonal.

Beweis. Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix. Seien $\vec{v}_1, \vec{v}_2$ Eigenvektoren von A mit Eigenwerten λ_1, λ_2 wobei $\lambda_1 \neq \lambda_2$. Ausgehend von

$$A\vec{v}_1 \quad = \quad \lambda_1\vec{v}_1$$

wird auf beiden Seiten das Skalarprodukt mit $\vec{v}_2$ genommen.

$$\begin{aligned}
(A\vec{v}_1) \circ \vec{v}_2 &= (\lambda_1\vec{v}_1) \circ \vec{v}_2 \\
\left(\overline{A\vec{v}_1}\right)^{\mathrm{T}} \vec{v}_2 &= \left(\overline{\lambda_1\vec{v}_1}\right)^{\mathrm{T}} \vec{v}_2 \\
\left(A\overline{\vec{v}_1}\right)^{\mathrm{T}} \vec{v}_2 &= \left(\overline{\lambda_1}\ \overline{\vec{v}_1}\right)^{\mathrm{T}} \vec{v}_2 \\
\overline{\vec{v}_1}^{\mathrm{T}} A^{\mathrm{T}}\vec{v}_2 &= \left(\lambda_1\overline{\vec{v}_1}\right)^{\mathrm{T}} \vec{v}_2 \\
\overline{\vec{v}_1}^{\mathrm{T}} A\vec{v}_2 &= \left(\lambda_1\ \overline{\vec{v}_1}^{\mathrm{T}}\right) \vec{v}_2 \\
\overline{\vec{v}_1}^{\mathrm{T}} \lambda_2\vec{v}_2 &= \lambda_1\left(\overline{\vec{v}_1}^{\mathrm{T}}\vec{v}_2\right) \\
\lambda_2\left(\overline{\vec{v}_1}^{\mathrm{T}}\vec{v}_2\right) &= \lambda_1\left(\overline{\vec{v}_1}^{\mathrm{T}}\vec{v}_2\right) \\
\lambda_2\left(\vec{v}_1 \circ \vec{v}_2\right) &= \lambda_1\left(\vec{v}_1 \circ \vec{v}_2\right)
\end{aligned}$$

Da $\lambda_1 \neq \lambda_2$ folgt $\vec{v}_1 \circ \vec{v}_2 = 0$.

Theorem 2.10 (Hauptachsentransformation)
Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch mit n einfachen Eigenwerten $\lambda_1, \ldots, \lambda_n$. Dann existieren zugehörige Eigenvektoren $\vec{v}_1, \ldots, \vec{v}_n \in \mathbb{R}^n$, die paarweise orthogonal und normiert sind. Fasst man diese Vekoren spaltenweise zu einer Matrix T zusammen, d.h.

$$T = (\vec{v}_1, \ldots, \vec{v}_n)$$

gilt

$$T^{\mathrm{T}}AT = \mathrm{diag}(\lambda_1, \ldots, \lambda_n),$$

wobei $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ *die Diagonalmatrix mit Diagonalelementen* $\lambda_1, \ldots, \lambda_n$ *ist.*

Beweis. Dass die Eigenvektoren einer symmetrischen Matrix paarweise orthogonal sind, wurde bereits gezeigt. Da Eigenräume abgeschlossen sind unter skalarer Multiplikation, kann man die Eigenvektoren auch auf Länge 1 normieren. Mit

$$T = (\vec{v}_1, \ldots, \vec{v}_n)$$

gilt

$$\begin{aligned}
(T^{\mathrm{T}}AT)_{ij} &= (T^{\mathrm{T}}(A\vec{v}_1, \ldots, A\vec{v}_n))_{ij} \\
&= (T^{\mathrm{T}}A\vec{v}_j)_i \\
&= (T^{\mathrm{T}}\lambda_j\vec{v}_j)_i \\
&= \lambda_j(T^{\mathrm{T}}\vec{v}_j)_i \\
&= \lambda_j(\vec{v}_i^{\mathrm{T}}\vec{v}_j) \\
&= \lambda_j(\vec{v}_i \circ \vec{v}_j) \\
&= \begin{cases} \lambda_j & \text{falls } i = j \\ 0 & \text{sonst} \end{cases}
\end{aligned}$$

Damit ist

$$T^{-1}AT = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix}.$$

Ohne Beweis sei ergänzt, dass man die Bedingung, dass A nur einfache Eigenwerte hat, auch weglassen kann.

Die Hauptachsentransformation spielt u.a. eine wichtige Rolle in der Mustererkennung. Hier ist A die (symmetrische) Kovarianzmatrix von Merkmalen. Sie wird auch in Kapitel 3.7 verwendet bei der Berechnung von lokalen Maxima bzw. Minima von mehrstelligen Funktionen, wobei A dann die (symmetrische) Hessematrix ist.

Weiterhin wird die Diagonalisierung von (nicht notwendigerweise symmetrischen) Matrizen auch in Kapitel 2.7 angewandt zur Lösung von DGL Systemen.

2.4 Von Einzelgleichungen zu Systemen

Systeme mit einer Gleichung. Lineare Differentialgleichungen beschreiben häufig zeitabhängige Phänomene wie z.B. Schwingungen. Als Funktionsvariable verwenden wir daher im Folgenden t. Die gesuchte Funktion wird mit $x(t)$ bezeichnet. Eine homogene, lineare Differentialgleichung erster Ordnung mit konstanten Koeffizienten hat somit die Form

$$x'(t) = ax(t).$$

Bei DGL dieser Bauart kommt man immer mit dem Ansatz

$$x(t) = e^{\lambda t}$$

zum Ziel, wobei λ durch Einsetzen bestimmt wird. Ableiten des Ansatzes ergibt

$$x'(t) = \lambda e^{\lambda t}.$$

Einsetzen in die DGL ergibt

$$\lambda e^{\lambda t} = ae^{\lambda t}.$$

Kürzen auf beiden Seiten mit $e^{\lambda t}$ liefert

$$\lambda = a.$$

Damit ist eine Lösungsfunktion

$$x(t) = e^{at},$$

was man durch Einsetzen leicht überprüfen kann. Da die DGL linear und homogen ist, ist auch die Funktion

$$x(t) = ce^{at}$$

für jede Konstante c eine Lösungsfunktion. Die allgemeine Lösung der DGL ist somit

$$x(t) = ce^{at}, \quad c \in \mathbb{R}.$$

Systeme mit mehreren Gleichungen. Ein homogenes lineares DGL System erster Ordnung mit konstanten Koeffizienten, bestehend aus zwei Gleichungen mit zwei unbekannten Funktionen $x(t)$ und $y(t)$ hat die Form

$$\begin{aligned} x'(t) &= a_{11}x(t) + a_{12}y(t) \\ y'(t) &= a_{21}x(t) + a_{22}y(t). \end{aligned}$$

Wäre $a_{12} = a_{21} = 0$, dann hätte man zwei voneinander unabhängige Differentialgleichungen mit je einer unbekannten Funktion, die man wie oben beschrieben unabhängig voneinander lösen könnte. Tatsächlich beruhen viele Ansätze zur Lösung von DGL Systemen auf Verfahren um die Gleichungen voneinander zu entkoppeln.

Die allgemeine Form eines homogenen linearen DGL Systems erster Ordnung mit konstanten Koeffizienten bestehend aus n Gleichungen mit n unbekannten Funktionen $x_1(t), \ldots, x_n(t)$ ist

$$\begin{aligned} x_1'(t) &= a_{11}x_1(t) + a_{12}x_2(t) + \ldots + a_{1n}x_n(t) \\ x_2'(t) &= a_{21}x_1(t) + a_{22}x_2(t) + \ldots + a_{2n}x_n(t) \\ &\vdots \\ x_n'(t) &= a_{n1}x_1(t) + a_{n2}x_2(t) + \ldots + a_{nn}x_n(t). \end{aligned}$$

Fasst man die Funktionen $x_1(t), \ldots, x_n(t)$ zu einem Vektor $\vec{x}(t)$ zusammen und die Koeffizienten a_{ij} zu einer $n \times n$ Matrix A, erhält man die kompakte Darstellung

$$\underbrace{\begin{pmatrix} x_1'(t) \\ \vdots \\ x_n'(t) \end{pmatrix}}_{\vec{x}'(t)} = \underbrace{\begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{nn} \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix}}_{\vec{x}(t)}$$

bzw.

$$\vec{x}'(t) = A\vec{x}(t).$$

Lösungsverfahren. Wie bei allen homogenen linearen DGL ist auch hier die Lösungsmenge abgeschlossen unter Addition und unter skalarer Multiplikation und bildet daher einen Vektorraum.

- Seien $\vec{x}_1(t), \vec{x}_2(t)$ Lösungsfunktionen, d.h.

  $$\vec{x}_1'(t) = A\vec{x}_1(t), \qquad \vec{x}_2'(t) = A\vec{x}_2(t).$$

  Dann ist

  $$\begin{aligned} (\vec{x}_1(t) + \vec{x}_2(t))' &= \vec{x}_1'(t) + \vec{x}_2'(t) \\ &= A\vec{x}_1(t) + A\vec{x}_2(t) \\ &= A(\vec{x}_1(t) + \vec{x}_2(t)) \end{aligned}$$

  und somit ist $\vec{x}_1(t) + \vec{x}_2(t)$ eine Lösungsfunktion.

- Sei $\vec{x}(t)$ eine Lösungsfunktion, d.h.

  $$\vec{x}'(t) = A\vec{x}(t)$$

  und $u \in \mathbb{R}$. Dann ist

  $$\begin{aligned} (u\vec{x}(t))' &= u\vec{x}'(t) \\ &= uA\vec{x}(t) \\ &= A(u\vec{x}(t)) \end{aligned}$$

  und somit ist $u\vec{x}(t)$ eine Lösungsfunktion.

Es geht also darum, eine Basis von linear unabhängigen Lösungsfunktionen zu finden. Die allgemeine Lösung ist dann die Menge aller Linearkombinationen dieser Basisfunktionen.

Der Ansatz zur Lösung des DGL Systems

$$\vec{x}'(t) = A\vec{x}(t)$$

ist ähnlich wie im eindimensionalen Fall

$$\vec{x}(t) = \vec{v}e^{\lambda t}, \qquad \vec{x}'(t) = \lambda\vec{v}e^{\lambda t}.$$

Einsetzen in die DGL ergibt

$$\lambda\vec{v}e^{\lambda t} = A\vec{v}e^{\lambda t}.$$

Kürzen mit $e^{\lambda t}$ führt auf

$$\lambda\vec{v} = A\vec{v}.$$

Das heißt, dass $\vec{v}$ ein Eigenvektor zum Eigenwert λ der Matrix A sein muss! Durch Lösen des Eigenwertproblems erhält man Eigenwerte λ_i mit zugehörigen Basisvektoren $\vec{v}_i$ des Eigenraums und damit linear unabhängige Lösungsfunktionen

$$\vec{x}_i(t) = \vec{v}_i e^{\lambda_i t}.$$

Die allgemeine Lösung erhält man durch beliebige Linearkombinationen:

$$\vec{x}(t) = \sum_i c_i \vec{v}_i e^{\lambda_i t} \text{ für beliebige } c_i \in \mathbb{R}.$$

Beispiel 2.11 Sei

$$\vec{x}'(t) = \begin{pmatrix} 1 & 1 \\ 3 & -1 \end{pmatrix} \vec{x}(t).$$

Mit dem Ansatz

$$\vec{x}(t) = \vec{v}e^{\lambda t}, \qquad \vec{x}'(t) = \lambda \vec{v} e^{\lambda t}$$

erhält man durch Einsetzen

$$\lambda \vec{v} e^{\lambda t} = \begin{pmatrix} 1 & 1 \\ 3 & -1 \end{pmatrix} \vec{v} e^{\lambda t}.$$

Kürzen mit $e^{\lambda t}$ ergibt

$$\lambda \vec{v} = \begin{pmatrix} 1 & 1 \\ 3 & -1 \end{pmatrix} \vec{v}.$$

Lösen des Eigenwertproblems liefert

$$\lambda_1 = 2, \quad \vec{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{und} \quad \lambda_2 = -2, \quad \vec{v}_2 = \begin{pmatrix} 1 \\ -3 \end{pmatrix}.$$

Damit hat man zwei Lösungsfunktionen

$$\begin{aligned} \vec{x}_1(t) &= \begin{pmatrix} 1 \\ 1 \end{pmatrix} e^{2t} \\ \vec{x}_2(t) &= \begin{pmatrix} 1 \\ -3 \end{pmatrix} e^{-2t}. \end{aligned}$$

Da beliebige Linearkombinationen von Lösungsfunktionen wieder Lösungsfunktionen sind, ist die allgemeine Lösung

$$\vec{x}(t) = c_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} e^{2t} + c_2 \begin{pmatrix} 1 \\ -3 \end{pmatrix} e^{-2t}, \qquad c_1, c_2 \in \mathbb{R}.$$

Leider klappt dieses Verfahren nicht immer. Eine erste Schwierigkeit ergibt sich, wenn man komplexe Eigenwerte hat.

Beispiel 2.12 Sei

$$\vec{x}'(t) = \begin{pmatrix} 0 & -1 \\ 2 & 2 \end{pmatrix} \vec{x}(t).$$

Mit dem Ansatz

$$\vec{x}(t) = \vec{v}e^{\lambda t}, \qquad \vec{x}'(t) = \lambda \vec{v} e^{\lambda t}$$

erhält man durch Einsetzen

$$\lambda \vec{v} e^{\lambda t} = \begin{pmatrix} 0 & -1 \\ 2 & 2 \end{pmatrix} \vec{v} e^{\lambda t}.$$

Kürzen mit $e^{\lambda t}$ ergibt

$$\lambda \vec{v} = \begin{pmatrix} 0 & -1 \\ 2 & 2 \end{pmatrix} \vec{v}.$$

Lösen des Eigenwertproblems liefert

$$\lambda_1 = 1 + j, \quad \vec{v}_1 = \begin{pmatrix} -1 \\ 1+j \end{pmatrix} \quad \text{und} \quad \lambda_2 = 1 - j, \quad \vec{v}_2 = \begin{pmatrix} -1 \\ 1-j \end{pmatrix}.$$

Damit hat man zwei Lösungsfunktionen

$$\begin{aligned} \vec{x}_1(t) &= \begin{pmatrix} -1 \\ 1+j \end{pmatrix} e^{(1+j)t} \\ \vec{x}_2(t) &= \begin{pmatrix} -1 \\ 1-j \end{pmatrix} e^{(1-j)t}. \end{aligned}$$

Mit der selben Argumentation wie bei der Schwingungsgleichung erhält man auch hier ein System von reellen Lösungen, indem man von einer komplexen Lösung Real- und Imaginärteil betrachtet. Umformen ergibt

$$\begin{aligned} \vec{x}_1(t) &= e^t(\cos(t) + j\sin(t)) \begin{pmatrix} -1 \\ 1+j \end{pmatrix} \\ &= e^t \begin{pmatrix} -\cos(t) - j\sin(t) \\ \cos(t) - \sin(t) + j(\cos(t) + \sin(t)) \end{pmatrix} \\ &= e^t \begin{pmatrix} -\cos(t) \\ \cos(t) - \sin(t) \end{pmatrix} + je^t \begin{pmatrix} -\sin(t) \\ \cos(t) + \sin(t) \end{pmatrix}. \end{aligned}$$

Damit ist die allgemeine Lösung

$$\vec{x}(t) = c_1 \underbrace{e^t \begin{pmatrix} -\cos(t) \\ \cos(t) - \sin(t) \end{pmatrix}}_{\mathrm{re}(\vec{x}_1(t))} + c_2 \underbrace{e^t \begin{pmatrix} -\sin(t) \\ \cos(t) + \sin(t) \end{pmatrix}}_{\mathrm{im}(\vec{x}_1(t))}, \quad c_1, c_2 \in \mathbb{R}.$$

Komplikationen gibt's wenn mehrfache Eigenwerte auftreten, d.h. das charakteristische Polynom mehrfache Nullstellen hat. Man kann mit gutem Recht argumentieren, dass dieser Fall für Ingenieursanwendungen uninteressant ist. Aufgrund von Messfehlern oder Toleranzen werden die Koeffizienten der Matrix immer mit kleinen Fehlern behaftet sein und sobald man etwas an ihnen wackelt, werden aus einer doppelten Nullstelle zwei reelle oder komplexe Nullstellen.

Beispiel 2.13 Sei

$$\vec{x}'(t) = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \vec{x}(t).$$

Die beiden Differentialgleichungen sind hier enkoppelt und man könnte sie unabhängig voneinander lösen. Trotzdem wenden wir die vorgestellte Methode an. Mit dem Ansatz

$$\vec{x}(t) = \vec{v}e^{\lambda t}, \qquad \vec{x}'(t) = \lambda \vec{v} e^{\lambda t}$$

erhält man durch Einsetzen in die DGL nach Kürzen mit $e^{\lambda t}$ wieder das Eigenwertproblem

$$\lambda \vec{v} = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} \vec{v}.$$

Es gibt hier nur einen (doppelten) Eigenwert $\lambda = 3$, der zugehörige Eigenraum wird jedoch von zwei linear unabhängigen Vektoren aufgespannt, d.h. λ hat algebraische und geometrische Vielfachheit 2. Eine Basis des Eigenraums ist z.B.

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Damit hat man zwei linear unabhängige Lösungsfunktionen

$$\vec{x}_1(t) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} e^{3t}$$
$$\vec{x}_2(t) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} e^{3t}$$

und die allgemeine Lösung

$$\vec{x}(t) = e^{3t} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}, \quad c_1, c_2 \in \mathbb{R}.$$

Und nun noch ein Beispiel, in dem das Verfahren nicht funktioniert. Die Matrix hat einen Eigenwert mit algebraischer Vielfachheit zwei, die geometrische Vielfachheit ist jedoch nur eins. Somit erhält man nur eine Basislösung.

Beispiel 2.14 Sei

$$\vec{x}'(t) = \begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix} \vec{x}(t).$$

Der Ansatz

$$\vec{x}(t) = \vec{v}e^{\lambda t}$$

führt zum Eigenwertproblem

$$\lambda\vec{v} = \begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix} \vec{v}.$$

Die charakteristische Gleichung

$$\lambda^2 - 2\lambda + 1 = 0$$

hat die doppelte Nullstelle $\lambda = 1$. Der zugehörige Eigenraum ist jedoch nur eindimensional. Eine Basis ist

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

Damit erhält man nur *eine* Lösungsfunktion

$$\vec{x}_1(t) = \begin{pmatrix} 1 \\ -1 \end{pmatrix} e^t.$$

Tatsächlich existiert aber noch eine zweite Lösungsfunktion, die von der ersten linear unabhängig ist, aber mit diesem Verfahren nicht gefunden wird!

2.5 Lösung durch Laplace Transformation.

Um auch den Fall lösen zu können, wenn die Eigenvektoren von A keine Basis des $\mathbb{R}^n$ bilden, sind stärkere Geschütze erforderlich. Bereits in früheren Situationen hat sich die Laplace Transformation als Werkzeug zur Lösung von Differentialgleichungen erwiesen und wird nun für Systeme eingesetzt.

Die Lösungsfunktionen $x(t)$ sind e-Funktionen oder Schwingungen, die keine Laplace Transformierte haben. Damit man trotzdem die Laplace Transformation anwenden kann, wird nicht $x(t)$ berechnet sondern $\sigma(t)x(t)$. Man erhält somit genau genommen die Lösungsfunktion nur für $t \geq 0$.

Wichtig ist der Ableitungssatz der Laplace Transformation, da hierbei aus einer Ableitung im Zeitbereich im wesentlichen ein einfacher Faktor s im Bildbereich wird. Wenn

$$\sigma(t)x(t) \quad \circ\!\!-\!\!\bullet \quad X(s)$$

dann gilt

$$\sigma(t)x'(t) \quad \circ\!\!-\!\!\bullet \quad sX(s) - x(0^-).$$

Die Laplace Transformation eignet sich also insbesondere für Anfangswertprobleme, bei denen $x(0^-)$ gegeben ist. Auf Systeme kann die Laplace Transformation dadurch angewandt werden, dass jede Gleichung des Systems Laplace transformiert wird.

Ausgehend von dem DGL System

$$\vec{x}'(t) \quad = \quad A\vec{x}(t)$$

werden zunächst beide Seiten mit $\sigma(t)$ multipliziert

$$\sigma(t)\vec{x}'(t) \quad = \quad A\sigma(t)\vec{x}(t).$$

Wendet man die Laplace Transformation komponentenweise auf beiden Seiten an, erhält man mit

$$\begin{aligned} \sigma(t)\vec{x}(t) \quad &\circ\!\!-\!\!\bullet \quad \vec{X}(s) \\ \sigma(t)\vec{x}'(t) \quad &\circ\!\!-\!\!\bullet \quad s\vec{X}(s) - \vec{x}(0^-) \end{aligned}$$

das lineare Gleichungssystem mit den Unbekannten $X_1(x), \ldots, X_n(s)$

$$s\vec{X}(s) - \vec{x}(0^-) \quad = \quad A\vec{X}(s).$$

Umformen ergibt

$$\begin{aligned} s\vec{X}(s) - A\vec{x}(s) \quad &= \quad \vec{x}(0^-) \\ sE\vec{X}(s) - A\vec{x}(s) \quad &= \quad \vec{x}(0^-) \\ (sE - A)\vec{X}(s) \quad &= \quad \vec{x}(0^-) \\ \vec{X}(s) \quad &= \quad (sE - A)^{-1}\vec{x}(0). \end{aligned}$$

Im zweiten Schritt musste die Einheitsmatrix E eingebaut werden, damit $X(s)$ ausgeklammert werden kann, vgl. Vorgehensweise bei der Berechnung der Eigenwerte. Hat man die Lösung $\vec{X}(s)$ im Bildbereich berechnet, muss diese zurücktransformiert werden, d.h.

$$\sigma(t)\vec{x}(t) \quad \circ\!\!-\!\!\bullet \quad (sE - A)^{-1}\vec{x}(0).$$

Um die Lösung $x_i(t)$ für $t \geq 0$ zu erhalten, ist die inverse Laplace Transformation von $X_i(s)$ erforderlich. Da $X_i(s)$ eine rationale Funktion in s ist, ist das mit Partialbruchzerlegung immer möglich. Statt die inverse Matrix von $sE - A$ zu berechnen, kann das LGS auch z.B. mit dem Gauß Algorithmus gelöst werden.

Da die Lösungsfunktionen von homogenen linearen DGL mit konstanten Koeffizienten immer stetig sind, gilt

$$\vec{x}(0^-) \quad = \quad \vec{x}(0).$$

Es ist somit egal, ob man als Startwerte $\vec{x}(0^-)$ oder $\vec{x}(0)$ vorgibt. Um die Notation zu vereinfachen lässt man den Faktor $\sigma(t)$ oft weg und denkt ihn sich dazu.

Beispiel 2.15 Wir betrachten das DGL System aus Beispiel 2.11:

$$\vec{x}'(t) = \begin{pmatrix} 1 & 1 \\ 3 & -1 \end{pmatrix} \vec{x}(t)$$

bzw.

$$\begin{aligned} x_1'(t) &= x_1(t) + x_2(t) \\ x_2'(t) &= 3x_1(t) - x_2(t). \end{aligned}$$

Laplace Transformation liefert

$$\begin{aligned} sX_1(s) - x_1(0) &= X_1(s) + X_2(s) \\ sX_2(s) - x_2(0) &= 3X_1(s) - X_2(s) \end{aligned}$$

bzw.

$$\begin{aligned} (1-s)X_1(s) + X_2(s) &= -x_1(0) \\ 3X_1(s) - (1+s)X_2(s) &= -x_2(0). \end{aligned}$$

Aus dem ursprünglichen DGL System ist also ein lineares Gleichungssystem geworden, deren Lösung man z.B. mit dem Gauß Algorithmus berechnen kann:

$$\begin{aligned} X_1(s) &= \frac{(1+s)x_1(0) + x_2(0)}{s^2 - 4} \\ X_2(s) &= \frac{3x_1(0) + (s-1)x_2(0)}{s^2 - 4}. \end{aligned}$$

Der eigentlich schwierige Teil ist die Rücktransformation der Lösung in den Zeitbereich. Hier hilft in der Regel eine Partialbruchzerlegung. Faktorisierung des Nenners ergibt

$$s^2 - 4 = (s-2)(s+2).$$

Damit erhält man

$$\begin{aligned} X_1(s) &= \frac{3x_1(0) + x_2(0)}{4(s-2)} + \frac{x_1(0) - x_2(0)}{4(s+2)} \\ X_2(s) &= \frac{3x_1(0) + x_2(0)}{4(s-2)} + \frac{-3x_1(0) + 3x_2(0)}{4(s+2)} \end{aligned}$$

Mit der Laplace Korrespondenz

$$e^{at} \quad \circ\!\!-\!\!\bullet \quad \frac{1}{s-a}$$

erhält man die Lösung im Zeitbereich.

$$\begin{aligned} x_1(t) &= \frac{3x_1(0) + x_2(0)}{4} e^{2t} + \frac{x_1(0) - x_2(0)}{4} e^{-2t} \\ x_2(t) &= \frac{3x_1(0) + x_2(0)}{4} e^{2t} + \frac{-3x_1(0) + 3x_2(0)}{4} e^{-2t} \end{aligned}$$

bzw.

$$\vec{x}(t) = \underbrace{\frac{3x_1(0)+x_2(0)}{4}}_{c_1}\begin{pmatrix}1\\1\end{pmatrix}e^{2t} + \underbrace{\frac{x_1(0)-x_2(0)}{4}}_{c_2}\begin{pmatrix}1\\-3\end{pmatrix}e^{-2t}.$$

Im Vergleich zu der in Beispiel 2.11 hergeleiteten Lösung zeigt der Weg über die Laplace Transformation gleich noch wie die Konstanten c_1 und c_2 von den Anfangswerten $x_1(0)$ und $x_2(0)$ abhängen.

Beispiel 2.16 Versuchen wir Beispiel 2.14

$$\vec{x}'(t) = \begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix} \vec{x}(t),$$

bei dem wir ja zunächst gescheitert sind, nun mit der Laplace Transformation zu lösen. Die beiden Gleichungen sind

$$\begin{aligned} x_1'(t) &= 3x_1(t) + 2x_2(t) \\ x_2'(t) &= -2x_1(t) - x_2(t). \end{aligned}$$

Laplace Transformation ergibt

$$\begin{aligned} sX_1(s) - x_1(0) &= 3X_1(s) + 2X_2(s) \\ sX_2(s) - x_2(0) &= -2X_1(s) - X_2(s) \end{aligned}$$

bzw.

$$\begin{aligned} (3-s)X_1(s) + 2X_2(s) &= -x_1(0) \\ 2X_1(s) + (1+s)X_2(s) &= x_2(0). \end{aligned}$$

Die Lösung des linearen Gleichungssystems ist

$$\begin{aligned} X_1(s) &= \frac{(1+s)x_1(0) + 2x_2(0)}{(s-1)^2} \\ X_2(s) &= \frac{-2x_1(0) + (-3+s)x_2(0)}{(s-1)^2} \end{aligned}$$

Für die Rücktransformation in den Zeitbereich wird zunächst eine Partialbruchzerlegung durchgeführt.

$$\begin{aligned} X_1(s) &= \frac{x_1(0)}{s-1} + \frac{2x_1(0) + 2x_2(0)}{(s-1)^2} \\ X_2(s) &= \frac{x_2(0)}{s-1} - \frac{2x_1(0) + 2x_2(0)}{(s-1)^2}. \end{aligned}$$

Mit Hilfe der Korrespondenzen

$$\begin{aligned} e^t &\circ\!\!-\!\!\bullet \frac{1}{s-1} \\ te^t &\circ\!\!-\!\!\bullet \frac{1}{(s-1)^2} \end{aligned}$$

erhält man die Lösung im Zeitbereich

$$\begin{aligned} x_1(t) &= x_1(0)e^t + (x_1(0) + x_2(0))2te^t \\ x_2(t) &= x_2(0)e^t - (x_1(0) + x_2(0))2te^t. \end{aligned}$$

In vektorieller Notation erhält man

$$\vec{x}(t) = e^t \begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix} + 2te^t(x_1(0) + x_2(0)) \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Dies lässt sich umformen in

$$\vec{x}(t) \quad = \quad x_1(0)e^t \begin{pmatrix} 1 \\ -1 \end{pmatrix} + (x_1(0) + x_2(0))e^t \begin{pmatrix} 2t \\ 1-2t \end{pmatrix}.$$

In dieser Darstellung sieht man, dass es zwei Basislösungen

$$e^t \begin{pmatrix} 1 \\ -1 \end{pmatrix} \text{ und } e^t \begin{pmatrix} 2t \\ 1-2t \end{pmatrix}$$

gibt. Die allgemeine Lösung ist die Menge aller Linearkombinationen dieser beiden Lösungen. In Beispiel 2.14 wurde mit der Eigenwertmethode nur die erste Basislösung gefunden.

2.6 Lösung mit e^{At} Ansatz.

Die e-Funktion, die bei der Lösung von linearen Differentialgleichungen so hilfreich war, kann formal auf Matrizen erweitert werden. Betrachten wir die Taylor Entwicklung

$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \ldots = \sum_{i=0}^{\infty} \frac{1}{i!}x^i$$

und ersetzen die Variable x durch eine Matrix $A \in \mathbb{R}^{n\times n}$, erhalten wir

$$e^A = E + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \ldots = \sum_{i=0}^{\infty} \frac{1}{i!}A^i.$$

Da die Summanden aufgrund von $i!$ im Nenner sehr schnell kleiner werden, konvergiert die Summe tatsächlich für jedes $A \in \mathbb{R}^{n\times n}$.

Definition 2.17 (e-Funktion für Matrizen)
Für jede Matrix $A \in \mathbb{R}^{n\times n}$ ist die Matrix $e^A \in \mathbb{R}^{n\times n}$ definiert durch

$$e^A = \sum_{i=0}^{\infty} \frac{1}{i!}A^i.$$

Aufgrund dieser Definition übertragen sich alle bekannten Eigenschaften der e-Funktion für reelle Zahlen auf die e-Funktion für Matrizen.

Ersetzt man in der Definition die Matrix A durch die Matrix At, erhält man eine Matrix

$$e^{At} = \sum_{i=0}^{\infty} \frac{1}{i!}(At)^i,$$

deren Komponenten Funktionen von der Zeit sind.

Wie von der e-Funktion nicht anders zu erwarten, gilt auch hier die Kettenregel der Ableitung

$$\left(e^{At}\right)' = Ae^{At}.$$

wobei die Ableitung der Matrix e^{At} komponentenweise zu verstehen ist. Dies sieht man wie folgt:

$$\begin{aligned}
\left(e^{At}\right)' &= \left(E + At + \frac{1}{2!}(At)^2 + \frac{1}{3!}(At)^3 + \frac{1}{4!}(At)^4 + \dots\right)' \\
&= \left(E + At + \frac{1}{2!}A^2t^2 + \frac{1}{3!}A^3t^3 + \frac{1}{4!}A^4t^4 \dots\right) \\
&= A + \frac{1}{2!}A^2\, 2t + \frac{1}{3!}A^3\, 3t^2 + \frac{1}{4!}A^4\, 4t^3 \dots \\
&= A + \frac{2}{2!}A^2t + \frac{3}{3!}A^3t^2 + \frac{4}{4!}A^4t^3 + \dots \\
&= A + A^2t + \frac{1}{2!}A^3t^2 + \frac{1}{3!}A^4t^3 + \dots \\
&= A\left(E + At + \frac{1}{2!}A^2t^2 + \frac{1}{3!}A^3t^3 + \dots\right) \\
&= Ae^{At}.
\end{aligned}$$

Sei nun $\vec{c}$ ein beliebiger, konstanter Vektor und

$$\vec{x}(t) = e^{At}\vec{c}.$$

Dann ist

$$\begin{aligned}
\vec{x}'(t) &= \left(e^{At}\vec{c}\right)' \\
&= Ae^{At}\vec{c} \\
&= A\vec{x}(t).
\end{aligned}$$

Damit ist $e^{At}\vec{c}$ für beliebiges $\vec{c}$ eine Lösung des DGL Systems

$$\vec{x}'(t) = A\vec{x}(t).$$

Wählt man für $\vec{c}$ die kanonischen Basisvektoren, erkennt man, dass jede Spalte von e^{At} eine Lösung des DGL Systems ist.

Sind Anfangswerte $\vec{x}_0$ gegeben, so ist

$$\vec{x}(t) = e^{At}\vec{x}_0$$

die Lösung des Anfangswertproblems, da

$$\vec{x}(0) = e^{A0}\vec{x}_0 = E\vec{x}_0 = \vec{x}_0.$$

Bleibt nur noch die Frage, wie man die Matrix e^{At} berechnen kann. In Kapitel 2.5 wurde gezeigt, dass

$$\vec{x}(t) \quad \circ\!\!-\!\!\bullet \quad (sE - A)^{-1}\vec{x}_0$$

die Lösung des Anfangswertproblems ist. Wie gerade gezeigt wurde, gilt aber auch

$$\vec{x}(t) \quad = \quad e^{At}\,\vec{x}_0.$$

Folglich muss

$$e^{At}\,\vec{x}_0 \quad \circ\!\!-\!\!\bullet \quad (sE - A)^{-1}\,\vec{x}_0$$

gelten. Da dies für jeden Vektor $\vec{x}_0$ der Fall ist, gilt die Matrix Korrespondenz

$$e^{At} \quad \circ\!\!-\!\!\bullet \quad (sE - A)^{-1},$$

die ganz ähnlich aussieht wie die bekannte Korrespondenz

$$e^{at} \quad \circ\!\!-\!\!\bullet \quad (s - a)^{-1}.$$

Die Matrix e^{At} kann also durch inverse Laplace Transformation von $(sE-A)^{-1}$ berechnet werden.

Man kann sich aber auch auf andere Weise davon überzeugen, dass

$$e^{At} \quad \circ\!\!-\!\!\bullet \quad (sE - A)^{-1}.$$

Mit der Korrespondenz

$$t^n \quad \circ\!\!-\!\!\bullet \quad \frac{n!}{s^{n+1}}$$

folgt

$$\begin{aligned} e^{At} &= E + At + \frac{1}{2!}A^2t^2 + \frac{1}{3!}A^3t^3 \\ &\circ\!\!-\!\!\bullet \quad E\frac{1}{s} + A\frac{1}{s^2} + \frac{1}{2!}A^2\frac{2!}{s^3} + \frac{1}{3!}A^3\frac{3!}{s^4} + \dots \\ &= \frac{E}{s} + \frac{A}{s^2} + \frac{A^2}{s^3} + \frac{A^3}{s^4} + \dots . \end{aligned}$$

Um zu sehen, dass dies die inverse Matrix von $sE - A$ ist, multipliziert man mit $sE - A$ und überzeugt sich, dass dabei die Einheitsmatrix herauskommt.

$$\begin{aligned} &(sE - A)\left(\frac{E}{s} + \frac{A}{s^2} + \frac{A^2}{s^3} + \frac{A^3}{s^4} + \dots\right) \\ &= s\left(\frac{E}{s} + \frac{A}{s^2} + \frac{A^2}{s^3} + \frac{A^3}{s^4} + \dots\right) - A\left(\frac{E}{s} + \frac{A}{s^2} + \frac{A^2}{s^3} + \frac{A^3}{s^4} + \dots\right) \\ &= E + \frac{A}{s} + \frac{A^2}{s^2} + \frac{A^3}{s^3} + \dots - \frac{A}{s} - \frac{A^2}{s^2} - \frac{A^3}{s^3} - \dots \\ &= E. \end{aligned}$$

Die Berechnung der Inversen Matrix $(sE - A)^{-1}$ mit dem Gauß Algorithmus ist manchmal etwas mühsam, da rationale Funktionen von s auftreten. Es gibt aber eine einfache Formel für die Invertierung für 2×2 Matrizen.

Sei

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Dann ist

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Man muss also nur die Diagonalelemente vertauschen, die Gegendiagonalelemente negieren und durch die Determinante teilen. Dass dies stimmt, sieht man wie folgt.

$$\begin{aligned} \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} &= \frac{1}{ad-bc} \begin{pmatrix} da-bc & db-bd \\ -ca+ac & -cb+ad \end{pmatrix} \\ &= \frac{1}{ad-bc} \begin{pmatrix} ad-bc & 0 \\ 0 & ad-bc \end{pmatrix} \\ &= E. \end{aligned}$$

Beispiel 2.18 Lösen wir Beispiel 2.16 mit der e^{At} Methode. In diesem Beispiel ist

$$\begin{aligned} A &= \begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix} \\ sE - A &= \begin{pmatrix} s-3 & -2 \\ 2 & s+1 \end{pmatrix} \\ (sE-A)^{-1} &= \frac{1}{(s-3)(s+1)+4} \begin{pmatrix} s+1 & 2 \\ -2 & s-3 \end{pmatrix} \\ &= \frac{1}{s^2-2s+1} \begin{pmatrix} s+1 & 2 \\ -2 & s-3 \end{pmatrix} \\ &= \frac{1}{(s-1)^2} \begin{pmatrix} s+1 & 2 \\ -2 & s-3 \end{pmatrix}. \end{aligned}$$

Die doppelte Nullstelle bei $s = 1$ im Nenner entspricht dem doppelten Eigenwert $\lambda = 1$ von A. Statt alle vier Komponenten der Matrix separat zurückzutransformieren, ist es geschickter, zunächst

$$\frac{1}{(s-1)^2} \quad \text{und} \quad \frac{s}{(s-1)^2}$$

zurückzutransformieren und dann Linearität auszunutzen. Aus der Formelsammlung entnimmt man

$$\frac{1}{(s-1)^2} \quad \bullet\!\!-\!\!\circ \quad t\,e^t.$$

Mit der Ableitung im Zeitbereich erhält man

$$\frac{s}{(s-1)^2} \quad \bullet\!\!-\!\!\circ \quad e^t + t\,e^t.$$

Der Übersichtlichkeit halber wurde der Faktor $\sigma(t)$ weggelassen. Damit gilt

$$\begin{aligned} \frac{1}{(s-1)^2} \begin{pmatrix} s+1 & 2 \\ -2 & s-3 \end{pmatrix} \quad \bullet\!\!-\!\!\circ \quad & \begin{pmatrix} e^t + t\,e^t + t\,e^t & 2t\,e^t \\ -2t\,e^t & e^t + t\,e^t - 3t\,e^t \end{pmatrix} \\ = \quad & e^t \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + te^t \begin{pmatrix} 2 & 2 \\ -2 & -2 \end{pmatrix} \end{aligned}$$

Für die Lösung des DGL Systems erhält man

$$\begin{aligned} (sE-A)^{-1}\vec{x}(0) \quad \bullet\!\!-\!\!\circ \quad & e^t \begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix} + 2te^t \begin{pmatrix} x_1(0)+x_2(0) \\ -x_1(0)-x_2(0) \end{pmatrix} \\ = \quad & e^t \begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix} + 2te^t(x_1(0)+x_2(0) \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \end{aligned}$$

2.7 Lösung durch Entkopplung.

Wir betrachten wieder das DGL System

$$\vec{x}'(t) = A\vec{x}(t).$$

Wenn die Matrix A eine Diagonalmatrix wäre, dann hätte man n unabhängige Differentialgleichungen, die man leicht lösen könnte. Im Folgenden wird gezeigt, wie man den allgemeinen Fall auf diesen trivialen Fall reduzieren kann.

Definition 2.19 (Diagonalisierbare Matrix)
Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt diagonalisierbar wenn es eine Matrix $T \in \mathbb{C}^{n\times n}$ gibt so dass

$$T^{-1}AT$$

eine Diagonalmatrix ist.

Die Transformation einer Matrix A auf eine Diagonalmatrix mit Hilfe einer Matrix T durch $T^{-1}AT$ wird als Hauptachsentransformation bezeichnet.

Nehmen wir einmal an, wir haben eine Matrix T gefunden, so dass

$$T^{-1}AT = D$$

Diagonalmatrix ist. Auflösen nach A ergibt

$$A = TDT^{-1}.$$

Ersetzt man A im DGL System durch TDT^{-1} erhält man

$$\begin{aligned} \vec{x}'(t) &= TDT^{-1}\vec{x}(t) \\ T^{-1}\vec{x}'(t) &= DT^{-1}\vec{x}(t). \end{aligned}$$

Mit der Substitution

$$\begin{aligned} \vec{y}(t) &= T^{-1}\vec{x}(t) \\ \vec{y}'(t) &= T^{-1}\vec{x}'(t) \end{aligned}$$

erhält man das DGL System

$$\vec{y}'(t) = D\vec{y}(t).$$

Da D Diagonalmatrix ist, sind in diesem System die Gleichungen voneinander unabhängig und lassen sich leicht lösen. Rücksubstituion liefert dann

$$\vec{x}(t) = T\vec{y}(t).$$

Bleibt das Problem, wie man zu einer gegebenen Matrix A eine Matrix T berechnet so dass $T^{-1}AT$ Diagonalmatrix ist.

Theorem 2.20

Eine Matix $A \in \mathbb{R}^{n\times n}$ ist diagonalisierbar genau dann wenn sie n linear unabhängige Eigenvektoren hat.

Die Gleichung

$$T^{-1}AT \;=\; D$$

ist genau dann erfüllt wenn die Spalten von T linear unabhängige Eigenvektoren von A sind und die Diagonalelemente von D die zugehörigen Eigenwerte.

Beweis. Nach Definition 2.19 ist A diagonalisierbar, wenn es eine Matrix $T \in \mathbb{C}^{n\times n}$ und eine Diagonalmatrix $D \in \mathbb{C}^{n\times n}$ gibt so dass

$$T^{-1}AT \;=\; D.$$

Diese Gleichung ist genau dann erfüllt wenn T regulär ist und

$$AT \;=\; TD.$$

Seien $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n$ die Spalten von T und $\lambda_1, \lambda_2, \ldots, \lambda_n$ die Diagonalelemente von D, d.h.

$$\begin{aligned} T &= (\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n) \\ D &= \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots \lambda_n & \end{pmatrix} = (\lambda_1\vec{e}_1,\ \lambda_2\vec{e}_2,\ \ldots,\ \lambda_n\vec{e}_n)\,. \end{aligned}$$

Für die Spalten von AT und TD gilt

$$\begin{aligned} AT &= (A\vec{v}_1,\ A\vec{v}_2,\ \ldots,\ A\vec{v}_n) \\ TD &= (T\lambda_1\vec{e}_1,\ T\lambda_2\vec{e}_2,\ \ldots,\ T\lambda_n\vec{e}_n) \;=\; (\lambda_1\vec{v}_1,\ \lambda_2\vec{v}_2,\ \ldots,\ \lambda_n\vec{v}_n)\,. \end{aligned}$$

Aus $AT = TD$ folgt somit

$$(A\vec{v}_1,\ A\vec{v}_2,\ \ldots,\ A\vec{v}_n) \;=\; (\lambda_1\vec{v}_1,\ \lambda_2\vec{v}_2,\ \ldots,\ \lambda_n\vec{v}_n)\,.$$

Diese Gleichung ist genau dann erfüllt wenn die Spalten $\vec{v}_1, \ldots, \vec{v}_n$ von T Eigenvektoren von A sind und die Diagonalelemente $\lambda_1, \ldots, \lambda_n$ von D die zugehörigen Eigenwerte. Damit T^{-1} existiert, muss T regulär sein, was genau dann der Fall ist wenn die Spalten von T linear unabhängig sind.

Schauen wir uns die auf diese Weise berechnete Lösung des DGL Systems

$$\vec{x}'(t) \quad = \quad A\vec{x}(t)$$

an. Das diagonalisierte System

$$\vec{y}'(t) \quad = \quad \underbrace{T^{-1}AT}_{\text{diag}(\lambda_1,\ldots,\lambda_n)} \quad \vec{y}(t)$$

ist einfach lösbar, da alle Gleichungen entkoppelt sind. Die i-te Gleichung hat die Form

$$y_i'(t) \quad = \quad \lambda_i y_i(t)$$

und die allgemeine Lösung ist

$$y_i(t) \quad = \quad c_i e^{\lambda_i t} \qquad \text{für beliebiges } c_i \in \mathbb{R}.$$

Die allgemeine Lösung des Systems $\vec{y}'(t) = T^{-1}AT\vec{y}(t)$ ist folglich

$$\vec{y}(t) \quad = \quad \begin{pmatrix} c_1 e^{\lambda_1 t} \\ c_2 e^{\lambda_2 t} \\ \vdots \\ c_n e^{\lambda_n t} \end{pmatrix}$$

Die allgemeine Lösung des ursprünglichen Systems $\vec{x}'(t) = A\vec{x}(t)$ ist damit

$$\begin{aligned} \vec{x}(t) &= T\vec{y}(t) \\ &= (\vec{v}_1 \quad \vec{v}_2 \quad \ldots \quad \vec{v}_n) \begin{pmatrix} c_1 e^{\lambda_1 t} \\ c_2 e^{\lambda_2 t} \\ \vdots \\ c_n e^{\lambda_n t} \end{pmatrix} \\ &= c_1\vec{v}_1 e^{\lambda_1 t} + c_2\vec{v}_2 e^{\lambda_2 t} + \ldots + c_n\vec{v}_n e^{\lambda_n t}. \end{aligned}$$

Das ist genau die gleiche Lösung, wie man sie beim Eigenvektoransatz erhält.

Bleibt die Frage, ob Diagonalisierbarkeit eine Ausnahmeeigenschaft von quadratischen Matrizen ist oder der Normalfall. Wenn Matrizen nur in Sonderfällen diagonalisierbar wären, würde sich der Aufwand nicht lohnen. Sie vermuten also richtig, dass man schon sehr viel Pech haben muss, um auf eine nicht diagonalisierbare Matrix zu treffen. Die Begründung ist wie folgt.

Ein Polynom n-ten Grades hat im "Normalfall" genau n einfache Nullstellen. Mehrfache Nullstellen sind eine Ausnahme. Sollte ein Polynom eine mehrfache Nullstelle haben, und man wackelt beliebig wenig an den Koeffizienten, ist die Wahrscheinlichkeit 1, dass es danach nur noch einfache Nullstellen hat. Da dies auch für das charakteristische Polynom gilt, ist es der Normalfall, dass eine $n \times n$ Matrix n Eigenwerte mit algebraischer Vielfachheit 1 hat.

Theorem 2.21
Sei $A \in \mathbb{R}^{n\times n}$ mit n einfachen Eigenwerten $\lambda_1, \ldots, \lambda_n$ und zugehörigen Eigenvektoren $\vec{v}_1, \ldots, \vec{v}_n$. Dann ist $(\vec{v}_1, \ldots, \vec{v}_n)$ linear unabhängig.

Beweis. Unter den Voraussetzungen des Theorems ist zu zeigen, dass $(\vec{v}_1, \ldots, \vec{v}_n)$ linear unabhängig ist. Wir verwenden hierfür Induktion.

- Das 1-Tupel $(\vec{v}_1)$ ist linear unabhängig, da $\vec{v}_1$ ein Eigenvektor und somit $\neq \vec{0}$ ist.
- Angenommen $(\vec{v}_1, \ldots, \vec{v}_{k-1})$ ist linear unabhängig.
- Zu zeigen ist, dass dann auch $(\vec{v}_1, \ldots, \vec{v}_k)$ linear unabhängig ist. Wäre $(\vec{v}_1, \ldots, \vec{v}_k)$ linear abhängig, dann müsste

$$\vec{v}_k = \sum_{\ell=1}^{k-1} c_\ell \vec{v}_\ell \qquad \text{für bestimmte } c_\ell$$

gelten. Laut Annahme ist ja $(\vec{v}_1, \ldots, \vec{v}_{k-1})$ linear unabhängig.
Dann ist

$$\begin{aligned} A\vec{v}_k &= \lambda_k \vec{v}_k \\ &= \lambda_k \sum_{\ell=1}^{k-1} c_\ell \vec{v}_\ell \\ &= \sum_{\ell=1}^{k-1} c_\ell \lambda_k \vec{v}_\ell \\ A\vec{v}_k &= A \sum_{\ell=1}^{k-1} c_\ell \vec{v}_\ell \\ &= \sum_{\ell=1}^{k-1} c_\ell A\vec{v}_\ell \\ &= \sum_{\ell=1}^{k-1} c_\ell \lambda_\ell \vec{v}_\ell. \end{aligned}$$

Folglich ist

$$\begin{aligned} \sum_{\ell=1}^{k-1} c_\ell \lambda_k \vec{v}_\ell &= \sum_{\ell=1}^{k-1} c_\ell \lambda_\ell \vec{v}_\ell \\ \sum_{\ell=1}^{k-1} c_\ell (\lambda_k - \lambda_\ell) \vec{v}_\ell &= \vec{0}. \end{aligned}$$

Da alle Eigenwerte einfach sind, ist $\lambda_k - \lambda_\ell \neq 0$. Da $\vec{v}_1, \ldots, \vec{v}_{k-1}$ linear unabhängig ist, muss somit $c_\ell = 0$ sein für $\ell = 1, \ldots, k-1$. In diesem Fall wäre aber $\vec{v}_k = \vec{0}$, was nicht sein kann, da $\vec{v}_k$ ein Eigenvektor ist.

Der Normalfall ist also, dass eine $n \times n$ Matrix genau n einfache Eigenwerte hat. In diesem Fall sind die zugehörigen Eigenvektoren $\vec{v}_1, \ldots, \vec{v}_n$ linear unabhängig. Die Matrix

$$T = (\vec{v}_1 \quad \vec{v}_2 \quad \ldots \quad \vec{v}_n)$$

ist dann invertierbar und

$$T^{-1} A T = D$$

eine Diagonalmatrix. Normalerweise sind quadratische Matrizen somit diagonalisierbar.

Man kann sogar noch etwas präzisieren: Eine $n \times n$ Matrix ist diagonalisierbar genau dann wenn für jeden Eigenwert die geometrische und die algebraische Vielfachheit gleich ist.

Weitere Eigenschaften der Hauptachsentransformation. Tatsächlich kann man mit der oben berechneten Matrix T nicht nur die Matrix A diagonalisieren sondern auch beliebige Potenzen von A.

Theorem 2.22
Wenn

$$T^{-1}AT = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

dann gilt auch

$$T^{-1}A^iT = \operatorname{diag}(\lambda_1^i, \ldots, \lambda_n^i)$$

für alle i.

Beweis. Sei

$$T^{-1}AT = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

Da $TT^{-1} = E$ und die Matrix Multiplikation assoziativ ist, gilt

$$\begin{aligned} T^{-1}A^iT &= T^{-1}A\,A\,A\cdots AT \\ &= T^{-1}A\,(TT^{-1})\,A\,(TT^{-1})\,A\cdots(TT^{-1})\,AT \\ &= (T^{-1}AT)\,(T^{-1}AT)\cdots(T^{-1}AT) \\ &= \operatorname{diag}(\lambda_1, \ldots, \lambda_n)\operatorname{diag}(\lambda_1, \ldots, \lambda_n)\cdots\operatorname{diag}(\lambda_1, \ldots, \lambda_n) \\ &= \operatorname{diag}(\lambda_1^i, \ldots, \lambda_n^i). \end{aligned}$$

Mit der gleichen Matrix T kann man auch e^{At} diagonalisieren.

Theorem 2.23
Wenn

$$T^{-1}AT = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

dann gilt auch

$$T^{-1}e^{At}T = \operatorname{diag}(e^{\lambda_1 t}, \ldots, e^{\lambda_n t}).$$

Beweis. Im vorigen Theorem wurde gezeigt, dass

$$T^{-1}A^iT = \operatorname{diag}(\lambda_1^i, \ldots, \lambda_n^i).$$

Auflösen nach A^i ergibt

$$A^i = T\operatorname{diag}(\lambda_1^i, \ldots, \lambda_n^i)T^{-1}.$$

Damit ist

$$\begin{aligned}
e^{At} &= \sum_{i=0}^{\infty} \frac{1}{i!}(At)^i \\
&= \sum_{i=0}^{\infty} \frac{t^i}{i!} A^i \\
&= \sum_{i=0}^{\infty} \frac{t^i}{i!} T\mathrm{diag}(\lambda_1^i, \ldots, \lambda_n^i) T^{-1} \\
&= T \sum_{i=0}^{\infty} \frac{t^i}{i!} \mathrm{diag}(\lambda_1^i, \ldots, \lambda_n^i) T^{-1} \\
&= T \sum_{i=0}^{\infty} \mathrm{diag}\left(\frac{1}{i!}(\lambda_1 t)^i, \ldots, \frac{1}{i!}(\lambda_n t)^i\right) T^{-1} \\
&= T\mathrm{diag}\left(\sum_{i=0}^{\infty} \frac{1}{i!}(\lambda_1 t)^i, \ldots, \sum_{i=0}^{\infty} \frac{1}{i!}(\lambda_n t)^i\right) T^{-1} \\
&= T\mathrm{diag}(e^{\lambda_1 t}, \ldots, e^{\lambda_n t}) T^{-1}.
\end{aligned}$$

Umformen liefert

$$T^{-1} e^{At} T = \mathrm{diag}(e^{\lambda_1 t}, \ldots, e^{\lambda_n t}).$$

Damit kann man e^{At} ohne Laplace Transformation berechnen, allerdings nur wenn die Matrix n linear unabhängige Eigenvektoren hat.

2.8 Lösung des inhomogenen DGL Systems

Einzelgleichungen. Wir betrachten zunächst das Anfangswertproblem mit einer Gleichung und einer unbekannten Funktion $x(t)$:

$$x'(t) \quad = \quad ax(t) + u(t), \quad x(0^-) = x_0.$$

Beim Lösungsverfahren mit Laplace Transformation tritt die Funktion e^{at} auf. Da diese keine Laplace Transformierte hat, müssen zunächst beide Seiten mit $\sigma(t)$ multipliziert werden.

$$\sigma(t)x'(t) \quad = \quad a\sigma(t)x(t) + \sigma(t)u(t).$$

Mit

$$\begin{aligned} \sigma(t)x(t) \quad &\circ\!\!-\!\!\bullet \quad X(s) \\ \sigma(t)u(t) \quad &\circ\!\!-\!\!\bullet \quad U(s) \end{aligned}$$

erhält man im Bildbereich

$$sX(s) - x_0 \quad = \quad aX(s) + U(s).$$

Auflösen nach $X(s)$ liefert die Lösung im Bildbereich.

$$X(s) \quad = \quad \frac{1}{s-a}x_0 + \frac{1}{s-a}U(s)$$

Mit den Laplace Korrespondenzen

$$\begin{aligned} \sigma(t)e^{at} \quad &\circ\!\!-\!\!\bullet \quad \frac{1}{s-a} \\ \sigma(t)e^{at} * \sigma(t)u(t) \quad &\circ\!\!-\!\!\bullet \quad \frac{1}{s-a}U(s) \end{aligned}$$

und

$$\begin{aligned} \sigma(t)e^{at} * \sigma(t)u(t) \quad &= \quad \int_{-\infty}^{\infty} \sigma(t-\tau)e^{a(t-\tau)}\sigma(\tau)u(\tau)d\tau \\ &= \quad \sigma(t)\int_0^t e^{a(t-\tau)}u(\tau)d\tau \end{aligned}$$

erhält man im Zeitbereich

$$\sigma(t)x(t) \quad = \quad \sigma(t)e^{at}x_0 + \sigma(t)\int_0^t e^{a(t-\tau)}u(\tau)d\tau.$$

Aufgrund des Faktors $\sigma(t)$ liefert dies keine Aussage über $x(t)$ für $t < 0$. Folglich gilt

$$x(t) \quad = \quad e^{at}x_0 + \int_0^t e^{a(t-\tau)}u(\tau)d\tau \quad \text{für } t \geq 0.$$

Gleichungssysteme. Die Vorgehensweise lässt sich auf die Lösung inhomogener DGL Systeme übertragen. Sei

$$\vec{x}'(t) \quad = \quad A\vec{x}(t) + \vec{u}(t), \quad \vec{x}(0^-) = \vec{x}_0.$$

Multiplikation mit $\sigma(t)$ liefert

$$\sigma(t)\vec{x}'(t) \quad = \quad A\ \sigma(t)\vec{x}(t) + \sigma(t)\vec{u}(t).$$

Mit der Laplace Transformation erhält man

$$s\vec{X}(s) - \vec{x}_0 \quad = \quad A\vec{X}(s) + \vec{U}(s).$$

Auflösen nach $\vec{X}(s)$ liefert die Lösung im Bildbereich.

$$\begin{aligned} \vec{X}(s) &= (sE - A)^{-1}(\vec{x}_0 + \vec{U}(s)) \\ &= (sE - A)^{-1}\vec{x}_0 \ + \ (sE - A)^{-1}\vec{U}(s). \end{aligned}$$

Mit den Korrespondenzen

$$\begin{aligned} \sigma(t)e^{At} &\quad \circ\!\!-\!\!\bullet \quad (sE - A)^{-1} \\ \sigma(t)e^{At} * \sigma(t)\vec{u}(t) &\quad \circ\!\!-\!\!\bullet \quad (sE - A)^{-1}\vec{U}(s) \end{aligned}$$

und

$$\begin{aligned} \sigma(t)e^{At} * \sigma(t)\vec{u}(t) &= \int_{-\infty}^{\infty} \sigma(t-\tau)e^{A(t-\tau)}\sigma(\tau)\vec{u}(\tau)d\tau \\ &= \sigma(t)\int_0^t e^{A(t-\tau)}\vec{u}(\tau)d\tau \end{aligned}$$

erhält man im Zeitbereich

$$\sigma(t)\vec{x}(t) \quad = \quad \sigma(t)e^{At}\vec{x}_0 \ + \ \sigma(t)\int_0^t e^{A(t-\tau)}\vec{u}(\tau)d\tau.$$

Aufgrund des Faktors $\sigma(t)$ liefert dies keine Aussagen über $\vec{x}(t)$ für $t < 0$. Folglich gilt

$$\vec{x}(t) \quad = \quad e^{At}\vec{x}_0 \ + \ \int_0^t e^{A(t-\tau)}\vec{u}(\tau)d\tau \quad \text{für } t \geq 0.$$

Beispiel 2.24 Sei

$$\vec{x}'(t) = \underbrace{\begin{pmatrix} 1 & 1 \\ 4 & -2 \end{pmatrix}}_{A} \vec{x}(t) + \underbrace{\begin{pmatrix} 1 \\ e^{-t} \end{pmatrix}}_{\vec{u}(t)}, \qquad \vec{x}_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Um den Rechenweg abzukürzen, kann man den Faktor $\sigma(t)$ weglassen und sich merken, dass die berechnete Lösung nur für $t \geq 0$ gilt.[1] Zunächst wird die Matrix e^{At} aus

$$e^{At} \quad \circ\!\!-\!\!\bullet \quad (sE - A)^{-1}$$

bestimmt. Es gilt

$$\begin{aligned} sE - A &= \begin{pmatrix} s-1 & -1 \\ -4 & s+2 \end{pmatrix} \\ (sE-A)^{-1} &= \frac{1}{s^2+s-6} \begin{pmatrix} s+2 & 1 \\ 4 & s-1 \end{pmatrix} \\ &= \frac{1}{s^2+s-6} \begin{pmatrix} 2 & 1 \\ 4 & -1 \end{pmatrix} + \frac{s}{s^2+s-6} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{aligned}$$

Partialbruchzerlegung ergibt

$$\begin{aligned} \frac{1}{s^2+s-6} &= \frac{1/5}{s-2} - \frac{1/5}{s+3} \\ \frac{s}{s^2+s-6} &= \frac{2/5}{s-2} + \frac{3/5}{s+3}. \end{aligned}$$

Damit ist

$$\begin{aligned} &(sE-A)^{-1} \\ &= \left(\frac{1/5}{s-2} - \frac{1/5}{s+3}\right) \begin{pmatrix} 2 & 1 \\ 4 & -1 \end{pmatrix} + \left(\frac{2/5}{s-2} + \frac{3/5}{s+3}\right) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \frac{1/5}{s-2} \begin{pmatrix} 2 & 1 \\ 4 & -1 \end{pmatrix} + \frac{2/5}{s-2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &\quad -\frac{1/5}{s+3} \begin{pmatrix} 2 & 1 \\ 4 & -1 \end{pmatrix} + \frac{3/5}{s+3} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \frac{1}{5(s-2)} \begin{pmatrix} 4 & 1 \\ 4 & 1 \end{pmatrix} + \frac{1}{5(s+3)} \begin{pmatrix} 1 & -1 \\ -4 & 4 \end{pmatrix} \\ e^{At} &= \begin{pmatrix} 4 & 1 \\ 4 & 1 \end{pmatrix} \frac{e^{2t}}{5} + \begin{pmatrix} 1 & -1 \\ -4 & 4 \end{pmatrix} \frac{e^{-3t}}{5}. \end{aligned}$$

[1] Tatsächlich gilt die Lösung in diesem Beispiel für alle $t \in \mathbb{R}$, aber das gibt die Laplace Transformation nicht her.

Die weiteren Terme in der Lösung werden wie folgt berechnet:

$$\begin{aligned}
e^{At}\vec{x}_0 &= \begin{pmatrix} 4 \\ 4 \end{pmatrix} \frac{e^{2t}}{5} + \begin{pmatrix} 1 \\ -4 \end{pmatrix} \frac{e^{-3t}}{5} \\
e^{A(t-\tau)}u(\tau) &= \left(\begin{pmatrix} 4 & 1 \\ 4 & 1 \end{pmatrix} \frac{e^{2(t-\tau)}}{5} + \begin{pmatrix} 1 & -1 \\ -4 & 4 \end{pmatrix} \frac{e^{-3(t-\tau)}}{5} \right) \begin{pmatrix} 1 \\ e^{-\tau} \end{pmatrix} \\
&= \begin{pmatrix} 4+e^{-\tau} \\ 4+e^{-\tau} \end{pmatrix} \frac{e^{2(t-\tau)}}{5} + \begin{pmatrix} 1-e^{-\tau} \\ -4+4e^{-\tau} \end{pmatrix} \frac{e^{-3(t-\tau)}}{5} \\
&= \begin{pmatrix} 4 \\ 4 \end{pmatrix} \frac{e^{2t}e^{-2\tau}}{5} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \frac{e^{2t}e^{-3\tau}}{5} + \\
&\quad \begin{pmatrix} 1 \\ -4 \end{pmatrix} \frac{e^{-3t}e^{3\tau}}{5} + \begin{pmatrix} -1 \\ 4 \end{pmatrix} \frac{e^{-3t}e^{2\tau}}{5} \\
\int_0^t e^{A(t-\tau)}u(\tau)d\tau &= \begin{pmatrix} 4 \\ 4 \end{pmatrix} \frac{e^{2t}-1}{10} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \frac{e^{2t}-e^{-t}}{15} + \\
&\quad \begin{pmatrix} 1 \\ -4 \end{pmatrix} \frac{1-e^{-3t}}{15} + \begin{pmatrix} -1 \\ 4 \end{pmatrix} \frac{e^{-t}-e^{-3t}}{10} \\
&= \frac{7}{15} \begin{pmatrix} 1 \\ 1 \end{pmatrix} e^{2t} + \frac{1}{6} \begin{pmatrix} -1 \\ 2 \end{pmatrix} e^{-t} + \\
&\quad \frac{1}{30} \begin{pmatrix} 1 \\ -4 \end{pmatrix} e^{-3t} - \frac{1}{3} \begin{pmatrix} 1 \\ 2 \end{pmatrix}
\end{aligned}$$

Nach dieser Zwischenrechnung kann nun die Lösung bestimmt werden.

$$\begin{aligned}
\vec{x}(t) &= e^{At}\vec{x}_0 + \int_0^t e^{A(t-\tau)}u(\tau)d\tau \\
&= \frac{19}{15} \begin{pmatrix} 1 \\ 1 \end{pmatrix} e^{2t} + \frac{1}{6} \begin{pmatrix} -1 \\ 2 \end{pmatrix} e^{-t} + \\
&\quad \frac{7}{30} \begin{pmatrix} 1 \\ -4 \end{pmatrix} e^{-3t} - \frac{1}{3} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \text{für } t \geq 0.
\end{aligned}$$

Alternativer Rechenweg. Statt e^{At} zu berechnen und mit $\vec{u}(t)$ im Zeitbereich zu falten, hätte man auch im Bildbereich multiplizieren und das Produkt zurücktransformieren können. Die Laplace Transformierte von $\vec{u}(t)$ ist komponentenweise

$$\begin{aligned}
\vec{u}(t) &= \begin{pmatrix} 1 \\ e^{-t} \end{pmatrix} \\
&\circ\!\!-\!\!\bullet \begin{pmatrix} \frac{1}{s} \\ \frac{1}{s+1} \end{pmatrix} = \vec{U}(s).
\end{aligned}$$

Damit ist

$$\begin{aligned}
\vec{x}_0 + \vec{U}(s) &= \begin{pmatrix} \frac{1}{s}+1 \\ \frac{1}{s+1} \end{pmatrix} \\
&= \frac{1}{s(s+1)} \begin{pmatrix} (s+1)^2 \\ s \end{pmatrix}
\end{aligned}$$

Multiplikation im Bildbereich.

$$\begin{aligned}(sE-A)^{-1}(\vec{x}_0+\vec{U}(s)) &= \left(\frac{1}{5(s-2)}\begin{pmatrix}4&1\\4&1\end{pmatrix}+\frac{1}{5(s+3)}\begin{pmatrix}1&-1\\-4&4\end{pmatrix}\right)\\ &\quad \frac{1}{s(s+1)}\begin{pmatrix}(s+1)^2\\ s\end{pmatrix}\\ &= \frac{4s^2+9s+4}{5(s-2)s(s+1)}\begin{pmatrix}1\\1\end{pmatrix}+\frac{s^2+s+1}{5(s+3)s(s+1)}\begin{pmatrix}1\\-4\end{pmatrix}\end{aligned}$$

Rücktransformation.

$$\begin{aligned}\frac{4s^2+9s+4}{5(s-2)s(s+1)} &= \frac{c_1}{s-2}+\frac{c_2}{s}+\frac{c_3}{s+1}\\ 4s^2+9s+4 &= 5c_1s(s+1)+5c_2(s-2)(s+1)+5c_3(s-2)s.\end{aligned}$$

Spezialfall $s=2$ liefert $c_1=19/15$
Spezialfall $s=0$ liefert $c_2=-2/5$
Spezialfall $s=-1$ liefert $c_3=-1/15$.

$$\begin{aligned}\frac{s^2+s+1}{5(s+3)s(s+1)} &= \frac{c_1}{s+3}+\frac{c_2}{s}+\frac{c_3}{s+1}\\ s^2+s+1 &= 5c_1s(s+1)+5c_2(s+3)(s+1)+5c_3(s+3)s.\end{aligned}$$

Spezialfall $s=-3$ liefert $c_1=7/30$
Spezialfall $s=0$ liefert $c_2=1/15$
Spezialfall $s=-1$ liefert $c_3=-1/10$.

Damit ist

$$\begin{aligned}&(sE-A)^{-1}(\vec{x}_0+\vec{U}(s))\\ &= \left(\frac{19}{15}\frac{1}{s-2}-\frac{2}{5}\frac{1}{s}-\frac{1}{15}\frac{1}{s+1}\right)\begin{pmatrix}1\\1\end{pmatrix}\\ &+ \left(\frac{7}{30}\frac{1}{s+3}+\frac{1}{15}\frac{1}{s}-\frac{1}{10}\frac{1}{s+1}\right)\begin{pmatrix}1\\-4\end{pmatrix}\\ &= \frac{19}{15}\frac{1}{s-2}\begin{pmatrix}1\\1\end{pmatrix}-\frac{1}{3}\frac{1}{s}\begin{pmatrix}1\\2\end{pmatrix}+\frac{1}{6}\frac{1}{s+1}\begin{pmatrix}-1\\2\end{pmatrix}+\frac{7}{30}\frac{1}{s+3}\begin{pmatrix}1\\-4\end{pmatrix}\\ &\bullet\!\!-\!\!\circ\ \frac{19}{15}e^{2t}\begin{pmatrix}1\\1\end{pmatrix}-\frac{1}{3}\begin{pmatrix}1\\2\end{pmatrix}+\frac{1}{6}e^{-t}\begin{pmatrix}-1\\2\end{pmatrix}+\frac{7}{30}e^{-3t}\begin{pmatrix}1\\-4\end{pmatrix}\\ &= \vec{x}(t) \quad \text{für } t\geq 0.\end{aligned}$$

3 Differentialrechnung mehrstelliger Funktionen

Eine Funktion $f \in \mathbb{R} \to \mathbb{R}$ kann man leicht veranschaulichen. Der Graph von f ist die Menge von zweistelligen Vektoren

$$G = \left\{ \begin{pmatrix} x \\ f(x) \end{pmatrix} \mid x \in \mathbb{R} \right\}.$$

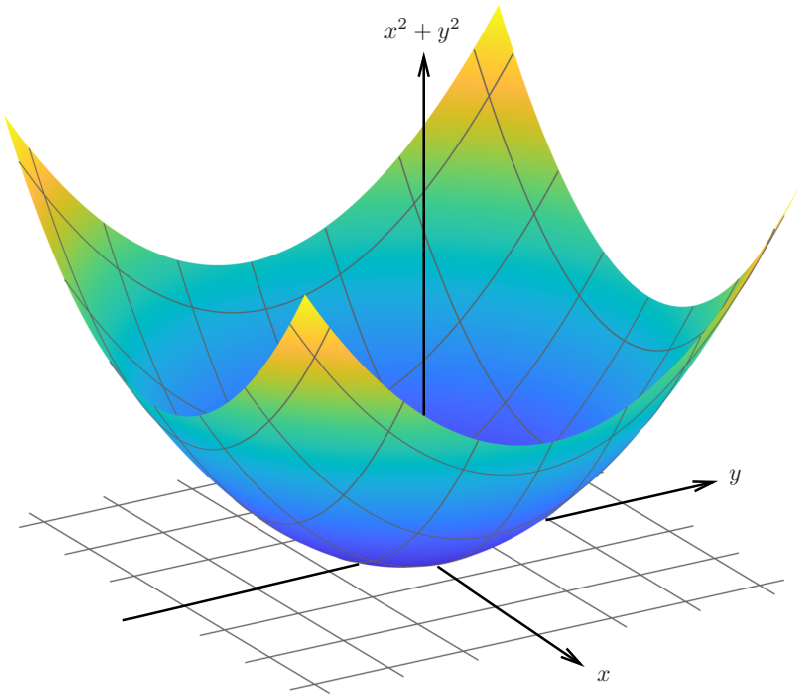
Jedes Element von G_f wird als Punkt in ein zweidimensionales Koordinatensystem einzeichnet.

Für Funktionen $f \in \mathbb{R}^2 \to \mathbb{R}$ kann man genauso vorgehen. Der Graph ist dann eine Menge von dreistelligen Vektoren

$$G = \left\{ \begin{pmatrix} x \\ y \\ f(x,y) \end{pmatrix} \mid x, y \in \mathbb{R} \right\}.$$

Um die Elemente von G_f als Punkte darzustellen, benötigt man ein dreidimensionales Koordinatensystem. Am einfachsten stellt man sich so eine Funktion als Gebirge über der xy-Ebene vor, das im Punkt (x, y) die Höhe $f(x, y)$ hat.

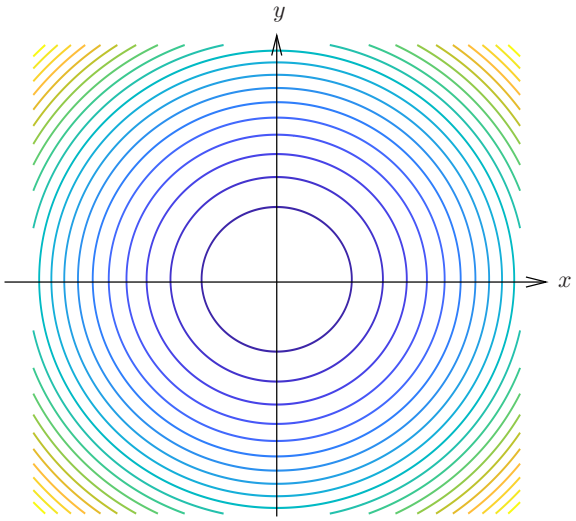
Für die Funktion $f(x, y) = x^2 + y^2$ ergibt sich folgendes Bild. Man nennt diese Funktion daher auch Rotationsparaboloid.



Eine andere Darstellungsmöglichkeit sind die Höhenlinien wie auf einer Landkarte. Eine Höhenlinie zur Höhe c ist die Menge

$$G_c = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \mid f(x,y) = c \right\}.$$

Zu unterschiedlichen Werten c zeichnet man die Elemente von G_c als Punkte in ein Koordinatensystem ein. Für die Funktion $f(x,y) = x^2 + y^2$ sind die Höhenlinien konzentrische Kreise. Der Abstand der Höhenlinien ist ein Maß für die Steilheit der Funktion.



3.1 Partielle Ableitung und Gradient

Bisher haben wir uns ausschließlich mit der Ableitung von einstelligen Funktionen $f \in \mathbb{R} \to \mathbb{R}$ beschäftigt. In diesem Kapitel wird der Begriff der Ableitung auf mehrstellige Funktionen $f \in \mathbb{R}^n \to \mathbb{R}$ verallgemeinert. Auch hier soll die Ableitung dem entsprechen was man anschaulich unter der Steigung versteht.

Stellt man sich eine zweistellige Funktion $f(x, y)$ als höhe eins Gebirges über der xy-Ebene vor und fragt sich wie steil das Gebirge an einer bestimmten Stelle $(\hat{x}, \hat{y})$ ist, so hängt die Antwort davon ab in welcher Richtung man gehen möchte. Geht man z.B. am Berg entlang, so ist die Steigung Null, geht man hingegen auf den Gipfel zu, ist die Steigung positiv. Besonders einfach ist die Steigung in x- und y-Richtung zu berechnen. Hierzu ein Beispiel:

Beispiel 3.1 Sei

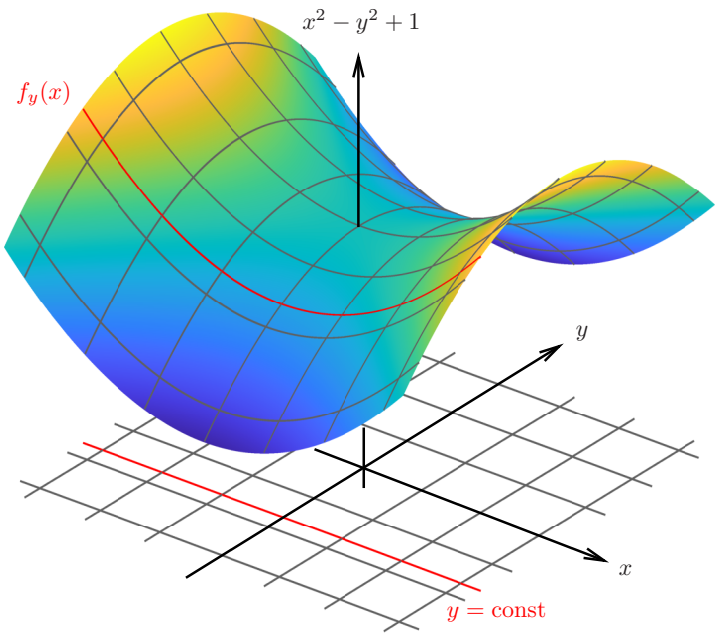
$$f(x, y) = x^2 - y^2 + 1.$$

- Die Steigung in x-Richtung erhält man, indem man y als konstanten Parameter betrachtet und sich nur noch in x-Richtung bewegt. Man erhält auf diese Weise eine einstellige Funktion $f_y(x)$, die man wie gewohnt ableitet.

$$\begin{aligned} f_y(x) &= x^2 - y^2 + 1 \\ f_y'(x) &= 2x. \end{aligned}$$

  Das Ergebnis heißt partielle Ableitung von f nach x

$$\frac{\partial}{\partial x} f(x, y) = 2x.$$

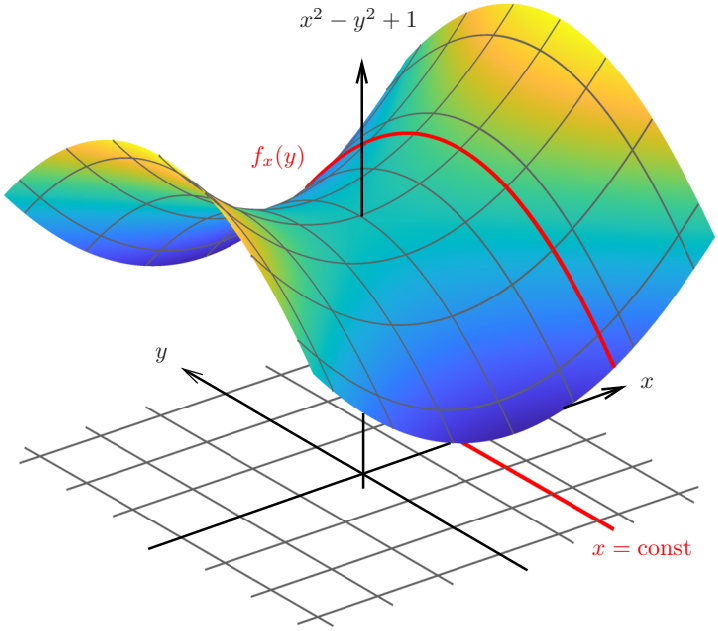


- In gleicher Weise erhält man die Steigung in y-Richtung, indem man x als konstanten Parameter betrachtet und sich nur noch in y-Richtung bewegt. Man erhält auf diese Weise eine einstellige Funktion $f_x(y)$, die man wie gewohnt ableitet, wobei die Funktionsvariable y ist.

$$\begin{aligned} f_x(y) &= x^2 - y^2 + 1 \\ f'_x(y) &= -2y. \end{aligned}$$

Das Ergebnis heißt partielle Ableitung von f nach y

$$\frac{\partial}{\partial y} f(x,y) = -2y.$$



Eine andere Möglichkeit, sich die partiellen Ableitungen zu veranschaulichen ist unter Verwendung der Höhenlinien.

- Um die Steigung in x-Richtung im Punkt $(\hat{x}, \hat{y})$ zu berechnen, läuft man ein kleines Stück Δx in x-Richtung. Die Sekantensteigung in x-Richtung ist dann die Höhendifferenz zwischen Zielpunkt $(\hat{x}+\Delta x, \hat{y})$ und Startpunkt $(\hat{x}, \hat{y})$ geteilt durch Schrittweite Δx, d.h.

  $$\frac{f(\hat{x} + \Delta x, \hat{y}) - f(\hat{x}, \hat{y})}{\Delta x}.$$

  Die Steigung im Punkt $(\hat{x}, \hat{y})$ ist der Grenzwert der Sekantensteigung für $\Delta x \to 0$, d.h.

  $$\frac{\partial}{\partial x} f(\hat{x}, \hat{y}) \quad = \quad \lim_{\Delta x \to 0} \frac{f(\hat{x} + \Delta x, \hat{y}) - f(\hat{x}, \hat{y})}{\Delta x}$$

- Analog erhält man im Punkt $(\hat{x}, \hat{y})$ die Sekantensteigung in y-Richtung. Man läuft ein kleines Stück Δy in y-Richtung und teilt die Höhendifferenz im Punkt $(\hat{x}, \hat{y} + \Delta y)$ und $(\hat{x}, \hat{y})$ durch die Schrittweite, d.h.

  $$\frac{f(\hat{x}, \hat{y} + \Delta y) - f(\hat{x}, \hat{y})}{\Delta y}.$$

  Die Steigung im Punkt $(\hat{x}, \hat{y})$ ist der Grenzwert der Sekantensteigung für $\Delta y \to 0$, d.h.

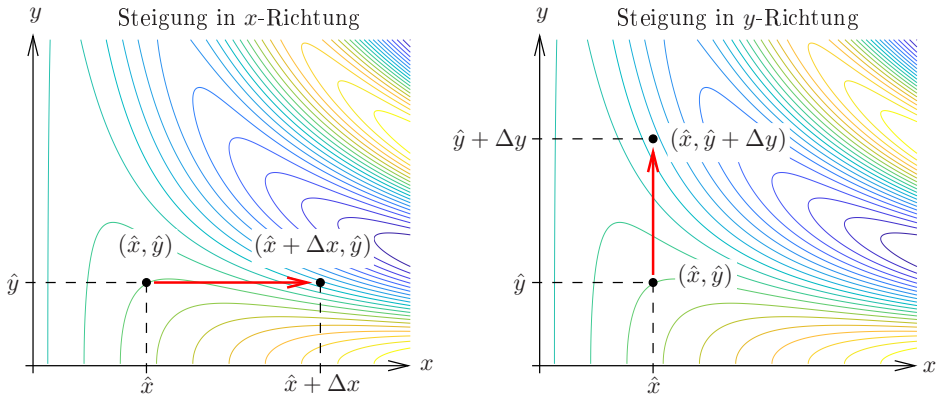
  $$\frac{\partial}{\partial y} f(\hat{x}, \hat{y}) \quad = \quad \lim_{\Delta y \to 0} \frac{f(\hat{x}, \hat{y} + \Delta y) - f(\hat{x}, \hat{y})}{\Delta y}.$$

Beispiel 3.2 Sei

$$f(x, y) \quad = x \cos(xy).$$

Dann ist

$$\begin{aligned} \frac{\partial}{\partial x} f(x, y) &= \cos(xy) - xy \sin(xy) \\ \frac{\partial}{\partial y} f(x, y) &= -x^2 \sin(xy). \end{aligned}$$



Allgemein kann man von einer n-stelligen Funktion auf diese Weise die partielle Ableitungen nach jeder Variablen berechnen.

Definition 3.3 partielle Ableitung

Die partielle Ableitung einer n-stelligen Funktion $f(x_1, x_2, \ldots, x_n)$ nach x_i erhält man indem man alle Variablen außer x_i als Konstanten betrachtet und die so entstehende einstellige Funktion $f(x_i)$ ableitet. Man schreibt hierfür

$$\frac{\partial}{\partial x_i} f(x_1, x_2, \ldots, x_n).$$

Wie im einstelligen Fall wird dies als Grenzwert der Sekantensteigung in x_i-Richtung definiert.

$$\begin{aligned} &\frac{\partial}{\partial x_i} f(x_1, x_2, \ldots, x_n) \\ &\quad = \lim_{\Delta x \to 0} \frac{f(x_1, \ldots, x_i + \Delta x, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{\Delta x}. \end{aligned}$$

Definition 3.4 Gradient

Der Gradient einer n-stelligen Funktion $f(x_1, x_2, \ldots, x_n)$ ist der n-stellige Vektor bestehend aus den partiellen Ableitungen

$$\nabla f(x_1, \ldots, x_n) \quad = \quad \begin{pmatrix} \partial/\partial x_1 f(x_1, \ldots, x_n) \\ \partial/\partial x_2 f(x_1, \ldots, x_n) \\ \vdots \\ \partial/\partial x_n f(x_1, \ldots, x_n) \end{pmatrix}$$

In Beispiel 3.2 ist somit

$$\nabla f(x, y) \quad = \quad \begin{pmatrix} \cos(xy) - xy \sin(xy) \\ -x^2 \sin(xy). \end{pmatrix}$$

Beispiel 3.5 Sei $\vec{b} \in \mathbb{R}^n$ und

$$f \in \mathbb{R}^n \to \mathbb{R}, \qquad f(\vec{x}) = \vec{b} \circ \vec{x}.$$

Gesucht ist der Gradient von f. Umformen ergibt

$$f(\vec{x}) \quad = \quad b_1 x_1 + b_2 x_2 + \ldots + b_n x_n.$$

Damit ist

$$\frac{\partial}{\partial x_i} f(\vec{x}) \quad = \quad b_i$$

und folglich

$$\nabla f(\vec{x}) \quad = \quad \begin{pmatrix} \partial/\partial x_1 f(\vec{x}) \\ \vdots \\ \partial/\partial x_n f(\vec{x}) \end{pmatrix} \quad = \quad \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \quad = \quad \vec{b}.$$

Beispiel 3.6 Sei $A \in \mathbb{R}^{n \times n}$, $\vec{b} \in \mathbb{R}^n$ und

$$f \in \mathbb{R}^n \to \mathbb{R}, \qquad f(\vec{x}) = \vec{b} \circ (A\vec{x}).$$

Gesucht ist der Gradient von f. Das Skalarprodukt kann durch eine Matrix Multiplikation ersetzt werden

$$\vec{x} \circ \vec{y} \quad = \quad \vec{x}^{\mathrm{T}} \vec{y}.$$

Unter Verwendung der Rechengesetze für Matrizen erhält man

$$\begin{aligned} f(\vec{x}) &= \vec{b}^{\mathrm{T}}(A\vec{x}) \\ &= (\vec{b}^{\mathrm{T}} A)\vec{x} \\ &= (\vec{b}^{\mathrm{T}} A)^{\mathrm{T}} \circ \vec{x} \\ &= (A^{\mathrm{T}} \vec{b}) \circ \vec{x}. \end{aligned}$$

Mit dem vorigen Beispiel gilt somit

$$\nabla f(\vec{x}) \quad = \quad A^{\mathrm{T}} \vec{b}.$$

Beispiel 3.7 Sei $A \in \mathbb{R}^{n\times n}$ und

$$f(\vec{x}) \quad = \quad \vec{x}^{\,\mathrm{T}} A\vec{x}.$$

Gesucht ist der Gradient von f. Umformen ergibt

$$\begin{aligned} f(\vec{x}) &= \vec{x} \circ (x_1\vec{a}_1 + \ldots + x_n\vec{a}_n) \\ &= \vec{x} \circ \sum_{\ell=1}^{n} x_\ell \vec{a}_\ell \\ &= \sum_{\ell=1}^{n} x_\ell (\vec{x} \circ \vec{a}_\ell) \\ &= \sum_{\ell=1}^{n} x_\ell \left(\sum_{k=1}^{n} a_{k\ell} x_k \right) \\ &= \sum_{\ell=1}^{n} \sum_{k=1}^{n} a_{k\ell} x_k x_\ell. \end{aligned}$$

Um die partielle Ableitung nach x_i zu berechnen, wird die Summe zerlegt.

$$\begin{aligned} \sum_{\ell=1}^{n} \sum_{k=1}^{n} a_{k\ell} x_k x_\ell &= \sum_{\substack{\ell\neq i \\ k\neq i}} a_{k\ell} x_k x_\ell + \sum_{\substack{\ell= i \\ k\neq i}} a_{k\ell} x_k x_\ell + \sum_{\substack{\ell\neq i \\ k= i}} a_{k\ell} x_k x_\ell + \sum_{\substack{\ell= i \\ k= i}} a_{k\ell} x_k x_\ell \\ &= \sum_{\substack{\ell\neq i \\ k\neq i}} a_{k\ell} x_k x_\ell + x_i \sum_{k\neq i} a_{ki} x_k + x_i \sum_{\ell\neq i} a_{i\ell} x_\ell + a_{ii} x_i^2. \end{aligned}$$

Damit ist

$$\begin{aligned} \frac{\partial}{\partial x_i} f(\vec{x}) &= \sum_{k\neq i} a_{ki} x_k + \sum_{\ell\neq i} a_{i\ell} x_\ell + 2x_i a_{ii} \\ &= \sum_{k=1}^{n} a_{ki} x_k + \sum_{\ell=1}^{n} a_{i\ell} x_\ell \\ &= (A^{\mathrm{T}}\vec{x})_i + (A\vec{x})_i \\ &= (A^{\mathrm{T}}\vec{x} + A\vec{x})_i \end{aligned}$$

Folglich ist

$$\begin{aligned} \nabla f(\vec{x}) &= A^{\mathrm{T}}\vec{x} + A\vec{x} \\ &= (A^{\mathrm{T}} + A)\vec{x}. \end{aligned}$$

Falls A eine symmetrische Matrix ist, gilt $A^{\mathrm{T}} = A$ und in diesem Fall

$$\nabla f(\vec{x}) \quad = \quad 2A\vec{x}.$$

Der Gradient einer Funktion hat ein paar sehr nützliche Eigenschaften, die in den folgenden Kapiteln bewiesen werden:

Merkregel 3.8

- *Der Gradient von* $f \in \mathbb{R}^n \to \mathbb{R}$ *ist der* n*-stellige Vektor*

$$\nabla f(x_1, \ldots, x_n) = \begin{pmatrix} \partial/\partial x_1 f(x_1, \ldots, x_n) \\ \vdots \\ \partial/\partial x_n f(x_1, \ldots, x_n) \end{pmatrix}.$$

- *Die Steigung von* f *an der Stelle* $(\hat{x}_1, \ldots, \hat{x}_n)$ *in Richtung der* i*-ten Koordinate ergibt sich aus der partiellen Ableitung von* f *nach* x_i

$$\frac{\partial}{\partial x_i} f(\hat{x}_1, \ldots, \hat{x}_n).$$

- *Die Steigung von* f *an der Stelle* $(\hat{x}_1, \ldots, \hat{x}_n)$ *in Richtung eines beliebigen Vektors* $\vec{r} \neq \vec{0}$ *ist*

$$\nabla f(\hat{x}_1, \ldots, \hat{x}_n) \circ \frac{\vec{r}}{\|\vec{r}\|}.$$

- *Der Gradient von* f *an der Stelle* $(\hat{x}_1, \ldots, \hat{x}_n)$ *zeigt in die Richtung des steilsten Anstiegs von* f *an der Stelle* $(\hat{x}_1, \ldots, \hat{x}_n)$
- *Hat* f *an der Stelle* $(\hat{x}_1, \ldots, \hat{x}_n)$ *ein lokales Extremum, so ist der Gradient dort gleich dem Nullvektor.*

3.2 Richtungsableitungen

Die Steigung einer zweistelligen Funktion in x- und y-Richtung kann durch die partiellen Ableitungen einfach berechnet werden. Nachfolgend soll die Steigung von $f(x, y)$ in einer beliebigen Richtung $\vec{r}$ berechnet werden.

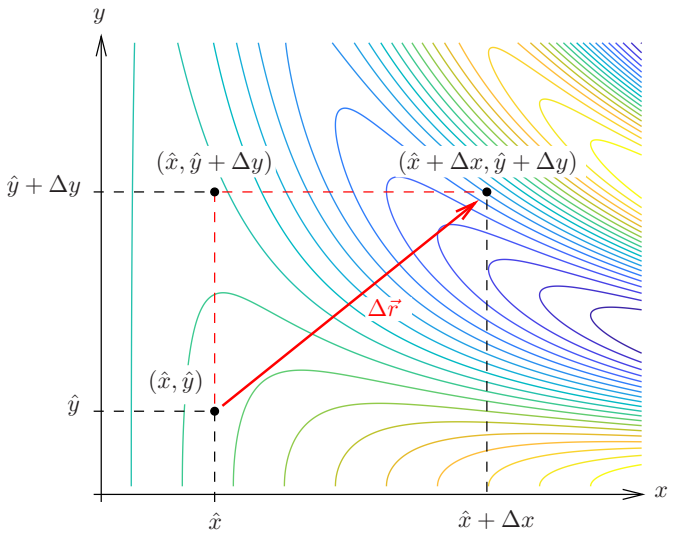
Sei

$$\Delta\vec{r} \quad = \quad \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$$

ein Vektor mit gleicher Richtung wie $\vec{r}$.

Dann ist die Sekantensteigung von f zwischen $(\hat{x}, \hat{y})$ und $(\hat{x} + \Delta x, \hat{y} + \Delta y)$

$$\frac{f(\hat{x} + \Delta x, \hat{y} + \Delta y) - f(\hat{x}, \hat{y})}{\|\Delta\vec{r}\|}.$$



Ein infinitesimal kleiner Schritt in Richtung $\vec{r}$ wird mit $d\vec{r}$ beichnet, d.h.

$$d\vec{r} \quad = \quad \begin{pmatrix} dx \\ dy \end{pmatrix}.$$

Die Richtungsableitung in Richtung $\vec{r}$ ist dann

$$\begin{aligned}
&\frac{d}{d\vec{r}} f(\hat{x},\hat{y}) \\
&= \frac{f(\hat{x}+dx,\hat{y}+dy)-f(\hat{x},\hat{y})}{\|d\vec{r}\|} \\
&= \frac{f(\hat{x}+dx,\hat{y}+dy)\overbrace{-f(\hat{x},\hat{y}+dy)+f(\hat{x},\hat{y}+dy)}^{=0}-f(\hat{x},\hat{y})}{\|d\vec{r}\|} \\
&= \frac{f(\hat{x}+dx,\hat{y}+dy)-f(\hat{x},\hat{y}+dy)}{\|d\vec{r}\|} + \frac{f(\hat{x},\hat{y}+dy)-f(\hat{x},\hat{y})}{\|d\vec{r}\|} \\
&= \left(\frac{f(\hat{x}+dx,\hat{y}+dy)-f(\hat{x},\hat{y}+dy)}{dx}\right)\frac{dx}{\|d\vec{r}\|} + \left(\frac{f(\hat{x},\hat{y}+dy)-f(\hat{x},\hat{y})}{dy}\right)\frac{dy}{\|d\vec{r}\|} \\
&= \frac{\partial}{\partial x} f(\hat{x},\hat{y}+dy)\frac{dx}{\|d\vec{r}\|} + \frac{\partial}{\partial y} f(\hat{x},\hat{y})\frac{dy}{\|d\vec{r}\|} \\
&= \frac{\partial}{\partial x} f(\hat{x},\hat{y})\frac{dx}{\|d\vec{r}\|} + \frac{\partial}{\partial y} f(\hat{x},\hat{y})\frac{dy}{\|d\vec{r}\|} \\
&= \begin{pmatrix} \partial/\partial x f(\hat{x},\hat{y}) \\ \partial/\partial y f(\hat{x},\hat{y}) \end{pmatrix} \circ \begin{pmatrix} dx/\|d\vec{r}\| \\ dy/\|d\vec{r}\| \end{pmatrix} \\
&= \nabla f(\hat{x},\hat{y}) \circ \frac{d\vec{r}}{\|d\vec{r}\|} \\
&= \nabla f(\hat{x},\hat{y}) \circ \frac{\vec{r}}{\|\vec{r}\|}.
\end{aligned}$$

Der Vektor $\vec{r}/\|\vec{r}\|$ ist der normierte Richtungsvektor von $\vec{r}$. Es ist klar, dass die Steigung in Richtung $\vec{r}$ unabhängig von der Länge von $\vec{r}$ sein muss.

Theorem 3.9 (Richtungsableitung)
Sei $f \in \mathbb{R}^n \to \mathbb{R}$ und $\vec{r} \in \mathbb{R}^n$ mit $\vec{r} \neq \vec{0}$. Für die Steigung von f in Richtung $\vec{r}$ gilt

$$\frac{f(\vec{x}+d\vec{r})-f(\vec{x})}{\|d\vec{r}\|} \quad = \quad \nabla f(\vec{x}) \circ \frac{\vec{r}}{\|\vec{r}\|}.$$

Insbesondere gilt für die Steigung in Richtung der kanonischen Basisvektoren $\vec{e}_i$

$$\nabla f(\vec{x}) \circ \vec{e}_i \quad = \quad \frac{\partial}{\partial x_i} f(\vec{x}).$$

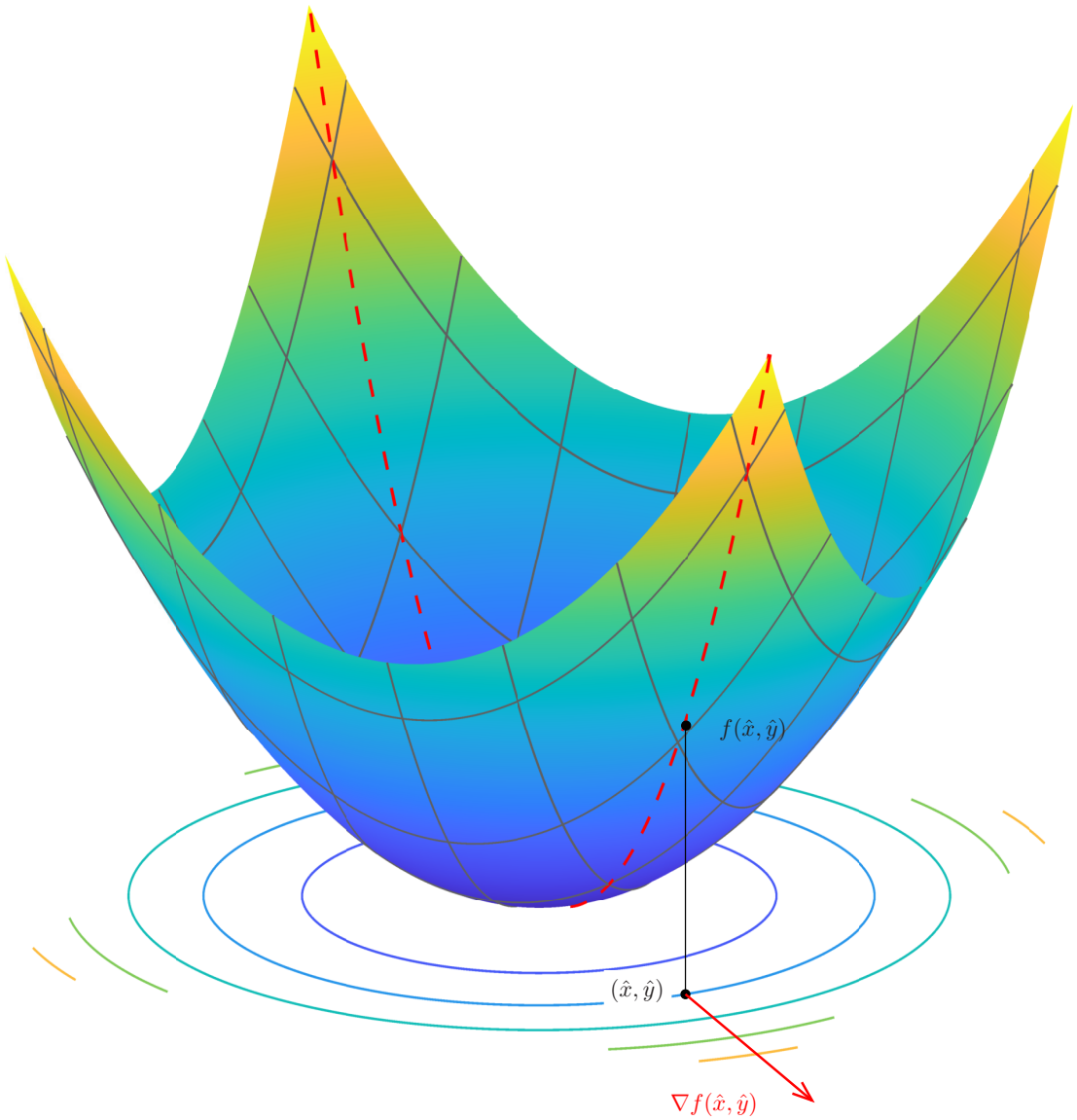
Für den Winkel α zwischen zwei Vektoren $\vec{x}, \vec{y}$ gilt

$$\cos(\alpha) \quad = \quad \frac{\vec{x} \circ \vec{y}}{\|\vec{x}\|\|\vec{y}\|}.$$

Folglich gilt für die Steigung von $f(x, y)$ an der Stelle $(\hat{x}, \hat{y})$ in Richtung $\vec{r}$

$$\nabla f(\hat{x}, \hat{y}) \circ \frac{\vec{r}}{\|\vec{r}\|} \quad = \quad \|\nabla f(\hat{x}, \hat{y})\| \cos(\alpha)$$

wobei α der Winkel zwischen Gradient und $\vec{r}$ ist. Die maximale Steigung erhält man für $\cos(\alpha) = 1$ bzw. $\alpha = 0$, d.h. wenn $\vec{r}$ die gleiche Richtung wie der Gradient hat. Der Gradient zeigt somit immer in die Richtung des steilsten Anstiegs von f.



Diese Beobachtung ist die Grundlage von sog. Gradientenverfahren. Sucht man ein lokales Maximum einer n-stelligen Funktion, muss man immer nur kleine Schritte in Richtung des Gradienten machen. Für ein lokales Minimum läuft man in die entgegengesetzte Richtung. So in etwa würde auch ein Bergsteiger vorgehen, der bei völliger Dunkelheit und Orientierungslosigkeit das Tal sucht. Auf dieser recht simplen Idee beruhen u.a. auch Lernverfahren für neuronale Netze.

Die Höhenlinien verlaufen in der Richtung, in der f konstant ist, d.h. keine Steigung hat. Zeigt der Richtungsvektor in Richtung der Höhenlinien, muss folglich für den Winkel α zwischen Gradient und Richtungsvektor gelten $\cos(\alpha) = 0$ bzw. $\alpha = \pi/2$. Der Gradient steht daher senkrecht zu den Höhenlinien.

Theorem 3.10 (Steilster Anstieg, Höhenlinien)
Sei $f \in \mathbb{R}^n \to \mathbb{R}$. Der Gradient $\nabla f(\vec{x})$ von f im Punkt $\vec{x}$ zeigt in die Richtung, in die f an der Stelle $\vec{x}$ am steilsten ansteigt.

Der Gradient steht senkrecht zu den Höhenlinien.

Beispiel 3.11 Sei $f(x,y) = x^2 + y^2$. Die partiellen Ableitungen sind

$$\begin{aligned} \frac{\partial}{\partial x} f(x,y) &= 2x \\ \frac{\partial}{\partial y} f(x,y) &= 2y \end{aligned}$$

und somit

$$\nabla f(x,y) = \begin{pmatrix} 2x \\ 2y \end{pmatrix}.$$

- Der Gradient im Punkt $(1,2)$ ist

$$\nabla f(1,2) = \begin{pmatrix} 2 \\ 4 \end{pmatrix}.$$

und somit ist f im Punkt $(1,2)$ am steilsten wenn man in Richtung des Vektors

$$\vec{r} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$

läuft. Die Steigung in dieser Richtung ist

$$\begin{aligned} \nabla f(1,2) \circ \frac{\vec{r}}{\|\vec{r}\|} &= \nabla f(1,2) \circ \frac{\nabla f(1,2)}{\|\nabla f(1,2)\|} \\ &= \frac{\nabla f(1,2) \circ \nabla f(1,2)}{\|\nabla f(1,2)\|} \\ &= \frac{\|\nabla f(1,2)\|^2}{\|\nabla f(1,2)\|} \\ &= \|\nabla f(1,2)\| \\ &= \sqrt{20} \\ &\approx 4.47. \end{aligned}$$

- Die Steigung im Punkt $(1,2)$ in alle anderen Richtungen ist kleiner. So erhält man z.B. für die Richtung

$$\vec{r} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$$

nur eine Steigung von

$$\nabla f(1,2) \circ \frac{\vec{r}}{\|\vec{r}\|} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} \circ \frac{1}{5} \begin{pmatrix} 3 \\ 4 \end{pmatrix} = \frac{22}{5} = 4.4.$$

In Richtung des Vektors

$$\vec{r} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

ist f an der Stelle $(1,2)$ sogar eben und man erhält

$$\nabla f(1,2) \circ \frac{\vec{r}}{\|\vec{r}\|} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} \circ \frac{1}{\sqrt{5}} \begin{pmatrix} 2 \\ -1 \end{pmatrix} = 0.$$

Beispiel 3.12 Sei $f(x,y,z) = \sin(xy) + z^2$. Die partiellen Ableitungen sind

$$\begin{aligned}\frac{\partial}{\partial x} f(x,y,z) &= y\cos(xy)\\ \frac{\partial}{\partial y} f(x,y,z) &= x\cos(xy)\\ \frac{\partial}{\partial z} f(x,y,z) &= 2z.\end{aligned}$$

und somit

$$\nabla f(x,y,z) = \begin{pmatrix} y\cos(xy)\\ x\cos(xy)\\ 2z\end{pmatrix}.$$

- Der Gradient im Punkt $(1,\pi,2)$ ist

$$\nabla f(1,\pi,2) = \begin{pmatrix} -\pi\\ -1\\ 4\end{pmatrix}$$

und somit ist f im Punkt $(1,\pi,2)$ am steilsten, wenn man in Richtung des Vektors

$$\vec{r} = \begin{pmatrix} -\pi\\ -1\\ 4\end{pmatrix}$$

läuft. Die Steigung in dieser Richtung ist

$$\begin{aligned}\nabla f(1,\pi,2) \circ \frac{\vec{r}}{\|\vec{r}\|} &= \nabla f(1,\pi,2) \circ \frac{\nabla f(1,\pi,2)}{\|\nabla f(1,\pi,2)\|}\\ &= \frac{\|\nabla f(1,\pi,2)\|^2}{\|\nabla f(1,\pi,2)\|}\\ &= \|\nabla f(1,\pi,2)\|\\ &= \sqrt{\pi^2+17}\\ &\approx 5.18.\end{aligned}$$

- Die Steigung im Punkt $(1,\pi,2)$ in alle anderen Richtungen ist kleiner. So erhält man z.B. für die Richtung

$$\vec{r} = \begin{pmatrix} 3\\ 4\\ -5\end{pmatrix}$$

nur eine Steigung von

$$\begin{aligned}\nabla f(1,\pi,2) \circ \frac{\vec{r}}{\|\vec{r}\|} &= \begin{pmatrix} -\pi\\ -1\\ 4\end{pmatrix} \circ \frac{1}{\sqrt{50}} \begin{pmatrix} 3\\ 4\\ -5\end{pmatrix}\\ &= \frac{-3\pi - 24}{\sqrt{50}}\\ &\approx -4.73\end{aligned}$$

3.3 Extremwertberechnung

Hat f an der Stelle $(\hat{x}_1, \ldots, \hat{x}_n)$ einen lokalen Extremwert, dann ist dort die Steigung in jede Richtung gleich Null. Dies ist genau dann der Fall, wenn $\nabla f(\hat{x}_1, \ldots, \hat{x}_n) = \vec{0}$ bzw.

$$\frac{\partial}{\partial x_i} f(\hat{x}_1, \ldots, \hat{x}_n) \quad = \quad 0 \quad \text{für } i = 1, \ldots, n.$$

Die Extremwertberechnung von mehrstelligen Funktionen wird also analog zur Extremwertberechnung von einstelligen Funktionen dadurch bewerkstelligt, dass man die gemeinsamen Nullstellen aller partiellen Ableitungen berechnet. Dies führt zur Lösung eines Gleichungssystems mit n Gleichungen und n Unbekannte. Wie bei einstelligen Funktionen gilt auch hier, dass ein verschwindender Gradient eine notwendige Bedingung für ein lokales Extremum ist aber keine hinreichende! Ein Punkt, an dem der Gradient verschwindet heißt auch stationärer Punkt. Es gibt also stationäre Punkte, die keine lokalen Extremwerte von f sind.

Beispiel 3.13 Sei $f(x, y, z) = \sin(xy) + z^2$. Die partiellen Ableitungen sind

$$\begin{aligned} \frac{\partial}{\partial x} f(x, y, z) &= y \cos(xy) \\ \frac{\partial}{\partial y} f(x, y, z) &= x \cos(xy) \\ \frac{\partial}{\partial z} f(x, y, z) &= 2z. \end{aligned}$$

Damit die partielle Ableitung nach z Null ist, muss $z = 0$ sein.

- Wenn $y = 0$ ist, ist die partielle Ableitung nach x Null. Da dann aber $\cos(xy) \neq 0$ ist, muss auch $x = 0$ sein damit die partielle Ableitung nach y Null ist.
- Wenn $y \neq 0$ ist, muss $\cos(xy) = 0$ sein, damit die partielle Ableitung nach x Null ist. In diesem Fall ist dann auch die partielle Ableitung nach x Null.

Die Menge der stationären Punkte ist somit

$$\left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right\} \cup \left\{ \begin{pmatrix} x \\ (\pi/2 + k\pi)/x \\ 0 \end{pmatrix} \mid x \neq 0, \;\; k \in \mathbb{Z} \right\}.$$

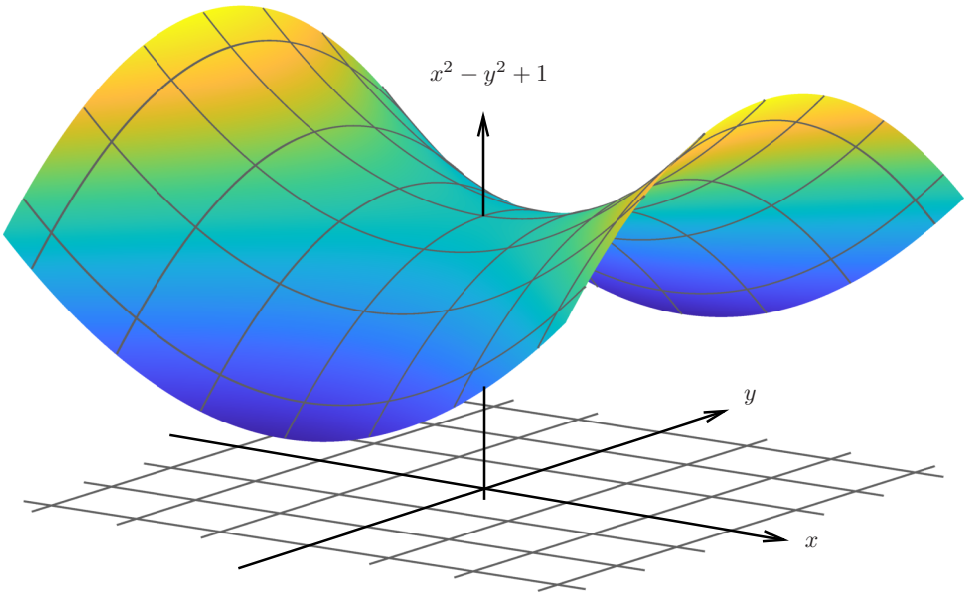
Beispiel 3.14 Sei

$$f(x,y) = x^2 - y^2.$$

Der Gradient von f ist

$$\nabla f(x,y) = \begin{pmatrix} 2x \\ -2y \end{pmatrix}.$$

Die Funktion hat einen Stationären Punkt bei $(\hat{x},\hat{y}) = (0,0)$. Dies ist jedoch weder ein lokales Maximum noch ein lokales Minimum sondern ein Sattelpunkt. Weicht man in x-Richtung ab, wird der Funktionswert größer, weicht man in y-Richtung ab, wird der Funktionswert kleiner.



3.4 Tangentialebenen

Tangente an $f \in \mathbb{R} \to \mathbb{R}$

Häufig hat man es mit komplizierten, nichtlinearen Funktionn $f \in \mathbb{R} \to \mathbb{R}$ zu tun, wobei jedoch nur die Umgebung um einen Arbeitspunkt $\hat{x}$ interessant ist. Ein Beispiel ist die nichtlinear Kennline eines Transistors, die nur im Bereich eines Basisstroms von wenigen mA relevant ist. In diesem Fall kann man $f(x)$ in der Nähe von $\hat{x}$ durch eine einfache Gerade $\ell(x)$ approximieren. Damit $\ell(x)$ eine Approximation an $f(x)$ in der Umgebung von $\hat{x}$ ist, wird verlangt, dass der Funktionswert und die Steigung von ℓ und f im Arbeitspunkt gleich sind, d.h.

$$\begin{aligned} \ell(\hat{x}) &= f(\hat{x}) \\ \ell'(\hat{x}) &= f'(\hat{x}). \end{aligned}$$

Da $\ell(x)$ eine Gerade ist, existieren Konstanten a, b so dass

$$\ell(x) = ax + b$$

für alle x. Setzt man diesen Ansatz in die o.g. Bedingungen ein, erhält man zwei Gleichungen

$$\begin{aligned} a\hat{x} + b &= f(\hat{x}) \\ a &= f'(\hat{x}). \end{aligned}$$

Hieraus können die Parameter a, b berechnet werden.

$$\begin{aligned} a &= f'(\hat{x}) \\ b &= f(\hat{x}) - f'(\hat{x})\hat{x}. \end{aligned}$$

Durch Einsetzen von a und b in die Geradengleichung erhält man die Lösung

$$\ell(x) = f(\hat{x}) + f'(\hat{x})(x - \hat{x}).$$

Somit ist die gesuchte Gerade $\ell(x)$ gleich der Tangente an $f(x)$ im Punkt $\hat{x}$.

Tangentialebene an $f \in \mathbb{R}^2 \to \mathbb{R}$

Das selbe Prinzip lässt sich auch für zweistellige Funktionen anwenden. Die Linearisierung von $f(x, y)$ im Punkt $(\hat{x}, \hat{y})$ ist hierbei eine zweistellige affin lineare Funktion $\ell(x, y)$, welche im Punkt $(\hat{x}, \hat{y})$ den gleichen Funktionswert und die gleichen partiellen Ableitungen wie $f(x, y)$ hat, d.h.

$$\begin{aligned} \ell(\hat{x}, \hat{y}) &= f(\hat{x}, \hat{y}) \\ \frac{\partial}{\partial x}\ell(\hat{x}, \hat{y}) &= \frac{\partial}{\partial x} f(\hat{x}, \hat{y}) \\ \frac{\partial}{\partial y}\ell(\hat{x}, \hat{y}) &= \frac{\partial}{\partial y} f(\hat{x}, \hat{y}). \end{aligned}$$

Mit der allgemeinen Form einer zweistelligen affin linearen Funktion

$$\ell(x, y) = a + bx + cy$$

erhält man die Bedingungen

$$\begin{aligned} a + b\hat{x} + c\hat{y} &= f(\hat{x}, \hat{y}) \\ b &= \frac{\partial}{\partial x} f(\hat{x}, \hat{y}) \\ c &= \frac{\partial}{\partial y} f(\hat{x}, \hat{y}) \end{aligned}$$

mit der Lösung

$$\begin{aligned} a &= f(\hat{x}, \hat{y}) - \frac{\partial}{\partial x} f(\hat{x}, \hat{y})\hat{x} - \frac{\partial}{\partial y} f(\hat{x}, \hat{y})\hat{y} \\ b &= \frac{\partial}{\partial x} f(\hat{x}, \hat{y}) \\ c &= \frac{\partial}{\partial y} f(\hat{x}, \hat{y}). \end{aligned}$$

Setzt man diese Werte in die allgemeine Form ein, erhält man

$$\begin{aligned} \ell(x, y) &= f(\hat{x}, \hat{y}) + \frac{\partial}{\partial x} f(\hat{x}, \hat{y})(x - \hat{x}) + \frac{\partial}{\partial y} f(\hat{x}, \hat{y})(y - \hat{y}) \\ &= f(\hat{x}, \hat{y}) + \nabla f(\hat{x}, \hat{y}) \circ \begin{pmatrix} x - \hat{x} \\ y - \hat{y} \end{pmatrix}. \end{aligned}$$

Beispiel 3.15 Gesucht ist die Linearisierung von

$$f(x,y) = x^2 + y^2$$

an der Stelle $(\hat{x}, \hat{y}) = (-1, -2)$.

Hierzu werden zunächst die partiellen Ableitungen von f berechnet:

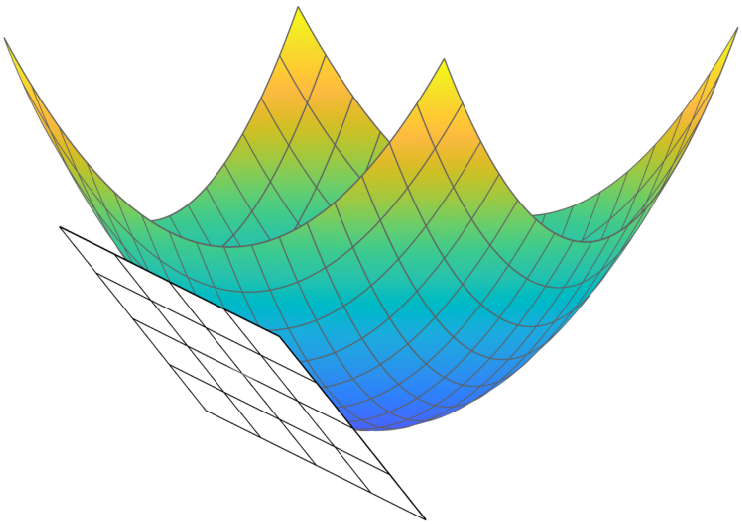
$$\begin{aligned} \frac{\partial}{\partial x} f(x,y) &= 2x \\ \frac{\partial}{\partial y} f(x,y) &= 2y. \end{aligned}$$

Durch Auswerten an der Stelle $(-1,-2)$ erhält man

$$\begin{aligned} f(-1,-2) &= 5 \\ \frac{\partial}{\partial x} f(-1,-2) &= -2 \\ \frac{\partial}{\partial y} f(-1,-2) &= -4. \end{aligned}$$

Somit ist die Tangentialebene an $f(x,y)$ im Punkt $(-1,-2)$ gegeben durch

$$\begin{aligned} \ell(x,y) &= f(1,2) + \frac{\partial}{\partial x} f(-1,-2)(x+1) + \frac{\partial}{\partial y} f(-1,-2)(y+2) \\ &= 5 - 2(x+1) - 4(y+2) \\ &= -5 - 2x - 4y. \end{aligned}$$



Tangentialebene an $f \in \mathbb{R}^n \to \mathbb{R}$

Für allgemeine n-stellige Funktionen ist die Linearisierung von $f(x_1, \ldots, x_n)$ im Punkt $(\hat{x}_1, \ldots, \hat{x}_n)$ die n-stellige affin lineare Funktion $\ell(x_1, \ldots, x_n)$, die im Punkt $(\hat{x}_1, \ldots, \hat{x}_n)$ den gleichen Funktionswert und die gleichen partiellen Ableitungen wie $f(x_1, \ldots, x_n)$ hat, d.h.

$$\begin{aligned} \ell(\hat{x}_1, \ldots, \hat{x}_n) &= f(\hat{x}_1, \ldots, \hat{x}_n) \\ \frac{\partial}{\partial x_i}\ell(\hat{x}_1, \ldots, \hat{x}_n) &= \frac{\partial}{\partial x_i} f(\hat{x}_1, \ldots, \hat{x}_n), \qquad i = 1, \ldots, n. \end{aligned}$$

Die allgemeine Form einer affin linearen Funktion ist

$$\ell(x_1, \ldots, x_n) = a + b_1 x_1 + \ldots + b_n x_n.$$

Daraus ergeben sich die Bedingungen

$$\begin{aligned} a + b_1\hat{x}_1 + \ldots + b_n\hat{x}_n &= f(\hat{x}_1, \ldots, \hat{x}_n) \\ b_i &= \frac{\partial}{\partial x_i} f(\hat{x}_1, \ldots, \hat{x}_n), \qquad i = 1, \ldots, n \end{aligned}$$

mit der Lösung

$$\begin{aligned} a &= f(\hat{x}_1, \ldots, \hat{x}_n) - \frac{\partial}{\partial x_1} f(\hat{x}_1, \ldots, \hat{x}_n) - \ldots - \frac{\partial}{\partial x_n} f(\hat{x}_1, \ldots, \hat{x}_n) \\ b_i &= \frac{\partial}{\partial x_i} f(\hat{x}_1, \ldots, \hat{x}_n), \qquad i = 1, \ldots, n. \end{aligned}$$

Eingesetzt in die allgemeine Form erhält man

$$\begin{aligned} \ell(x_1, \ldots, x_n) &= f(\hat{x}_1, \ldots, \hat{x}_n) + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(\hat{x}_1, \ldots, \hat{x}_n)(x_i - \hat{x}_i) \\ &= f(\hat{x}_1, \ldots, \hat{x}_n) + \begin{pmatrix} \partial/\partial x_1 f(\hat{x}_1, \ldots, \hat{x}_n) \\ \vdots \\ \partial/\partial x_n f(\hat{x}_1, \ldots, \hat{x}_n) \end{pmatrix} \circ \begin{pmatrix} x_1 - \hat{x}_1 \\ \vdots \\ x_n - \hat{x}_n \end{pmatrix} \\ &= f(\hat{\vec{x}}) + \nabla f(\hat{\vec{x}}) \circ (\vec{x} - \hat{\vec{x}}). \end{aligned}$$

Die Formel hat eine ähnliche Sturkur wie die Tangente von einstelligen Funktionen auf Seite 100

$$\ell(x) = f(\hat{x}) + f'(\hat{x})(x - \hat{x}).$$

Beispiel 3.16 Gesucht ist die Linearisierung von

$$f(x,y,z) = x^2 \sin(yz) + 3z$$

an der Stelle $(\hat{x},\hat{y},\hat{z}) = (1,0,-2)$. Hierzu werden zunächst die partiellen Ableitungen von f berechnet:

$$\begin{aligned}
\frac{\partial}{\partial x} f(x,y,z) &= 2x\sin(yz) \\
\frac{\partial}{\partial y} f(x,y,z) &= x^2 z\cos(yz) \\
\frac{\partial}{\partial z} f(x,y,z) &= x^2 y\cos(yz) + 3.
\end{aligned}$$

Durch Auswerten an der Stelle $(1,0,-2)$ erhält man

$$\begin{aligned}
f(1,0,-2) &= -6 \\
\frac{\partial}{\partial x} f(1,0,-2) &= 0 \\
\frac{\partial}{\partial y} f(1,0,-2) &= -2 \\
\frac{\partial}{\partial z} f(1,0,-2) &= 3.
\end{aligned}$$

Somit ist die Tangentialebene an $f(x,y,z)$ im Punkt $(1,0,-2)$ gegeben durch

$$\begin{aligned}
\ell(x,y,z) &= f(1,0,-2) + \frac{\partial}{\partial x} f(1,0,-2)(x-1) + \\
&\quad \frac{\partial}{\partial y} f(1,0,-2)(y-0) + \frac{\partial}{\partial z} f(1,0,-2)(z+2) \\
&= -6 + 0(x-1) - 2(y-0) + 3(z+2) \\
&= 3z - 2y.
\end{aligned}$$

Merkregel 3.17

Die Linearisierung von $f(x_1,\ldots,x_n)$ im Punkt $(\hat{x}_1,\ldots,\hat{x}_n)$ ist die affin lineare Funktion

$$\ell(x_1,\ldots,x_n) = f(\hat{x}_1,\ldots,\hat{x}_n) + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(\hat{x}_1,\ldots,\hat{x}_n)(x_i - \hat{x}_i)$$

bzw.

$$\ell(\vec{x}) = f(\hat{\vec{x}}) + \nabla f(\hat{\vec{x}}) \circ (\vec{x} - \hat{\vec{x}}).$$

3.5 Mehrstellige Taylor Polynome

Einstellige Taylor Polynome

Definition 3.18 (Taylor Polynom)
Das Taylor Polynom vom Grad m an $f \in \mathbb{R} \to \mathbb{R}$ im Arbeitspunkt $\hat{x}$ ist definiert durch

$$\begin{aligned} p(x) &= f(\hat{x}) + f'(\hat{x})(x-\hat{x}) + \frac{1}{2!}f''(\hat{x})(x-\hat{x})^2 + \\ &\quad \ldots + \frac{1}{m!}f^{(m)}(\hat{x})(x-\hat{x})^m \\ &= \sum_{k=0}^{m} \frac{1}{k!} f^{(k)}(\hat{x})(x-\hat{x})^k. \end{aligned}$$

Das Taylor Polynom $p(x)$ ist eine gute Approximation an f in der Umgebung von $\hat{x}$, da alle Ableitungen von f und p bis zur m-tem im Punkt $\hat{x}$ übereinstimmen.

Theorem 3.19
Sei $p(x)$ das Taylor Polynom vom Grad m an $f(x)$ im Arbeitspunkt $\hat{x}$. Dann gilt

$$p^{(i)}(\hat{x}) = f^{(i)}(\hat{x}) \qquad \text{für } i = 0, \ldots, m.$$

Beweis. Für die i-te Ableitung des Taylor Polynoms gilt

$$\begin{aligned} p^{(i)}(x) &= \frac{d^i}{dx^i} \sum_{k=0}^{m} \frac{1}{k!} f^{(k)}(\hat{x})(x-\hat{x})^k \\ &= \sum_{k=0}^{m} \frac{1}{k!} f^{(k)}(\hat{x}) \frac{d^i}{dx^i}(x-\hat{x})^k. \end{aligned}$$

Es gilt

$$\frac{d^i}{dx^i}(x-\hat{x})^k = \begin{cases} 0 & \text{falls } k < i \\ i! & \text{falls } k = i \end{cases}.$$

Für $k > i$ ist

$$\frac{d^i}{dx^i}(x-\hat{x})^k = 0$$

an der Stelle $x = \hat{x}$. Somit bleibt vom Taylor Polynom ausgewertet bei $x = \hat{x}$ nur der Summand mit $k = i$ übrig, d.h.

$$p^{(i)}(\hat{x}) = \frac{1}{i!} f^{(i)}(\hat{x}) i! = f^{(i)}(\hat{x}).$$

Zweistellige Taylor Polynome

Höhere partielle Ableitungen sind für mehrstellige Funktionen gleich definiert wie im einstelligen Fall. Mit $f^{(i,j)}$ wird die i-te partielle Ableitung nach x und die j-te partielle Ableitung nach y von $f(x,y)$ bezeichnet, d.h.

$$f^{(i,j)}(x,y) \quad = \quad \left(\frac{\partial}{\partial x}\right)^i \left(\frac{\partial}{\partial y}\right)^j f(x,y).$$

Die Reihenfolge, in der man die Ableitungen durchführt, spielt dabei keine Rolle sofern alle dabei auftretenden partiellen Ableitungen stetig sind (Satz von Schwarz). Dies wird im Folgenden vorausgesetzt.

Beispiel 3.20 Sei

$$f(x,y) \quad = \quad e^x \sin(x+3y).$$

Dann ist

$$\begin{aligned}
\frac{\partial}{\partial x} f(x,y) &= e^x \sin(x+y) + e^x \cos(x+3y) \\
&= e^x(\sin(x+3y) + \cos(x+3y)) \\
\frac{\partial}{\partial y} f(x,y) &= 3e^x \cos(x+3y) \\
\frac{\partial}{\partial x}\frac{\partial}{\partial y} f(x,y) &= \frac{\partial}{\partial x}(3e^x \cos(x+3y)) \\
&= 3(e^x \cos(x+3y) - e^x \sin(x+3y)) \\
&= 3e^x(\cos(x+3y - \sin(x+3y)) \\
\frac{\partial}{\partial y}\frac{\partial}{\partial x} f(x,y) &= \frac{\partial}{\partial y}(e^x(\sin(x+3y) + \cos(x+3y))) \\
&= e^x(3\cos(x+3y) - 3\sin(x+3y)) \\
&= 3e^x(\cos(x+3y) - \sin(x+3y)).
\end{aligned}$$

Definition 3.21 (Zweistelliges Taylor Polynom)
Das Taylor Polynom vom Grad m an $f \in \mathbb{R}^2 \to \mathbb{R}$ im Arbeitspunkt $(\hat{x}, \hat{y})$ ist definiert durch

$$p(x,y) = \sum_{\substack{k,\ell \geq 0 \\ k+\ell \leq m}} \frac{1}{k!\ell!} f^{(k,\ell)}(\hat{x},\hat{y})(x-\hat{x})^k (y-\hat{y})^\ell.$$

Theorem 3.22
Sei $p(x,y)$ das Taylor Polynom vom Grad m an $f(x,y)$ im Arbeitspunkt $(\hat{x},\hat{y})$. Dann gilt

$$p^{(i,j)}(\hat{x},\hat{y}) = f^{(i,j)}(\hat{x},\hat{y}) \qquad \text{für alle } i,j \text{ mit } i+j \leq m$$

Beweis. Für die i,j-te Ableitung des Taylor Polynoms gilt

$$\begin{aligned} p^{(i,j)}(x,y) &= \left(\frac{\partial}{\partial x}\right)^i \left(\frac{\partial}{\partial y}\right)^j \sum_{\substack{k,\ell \geq 0 \\ k+\ell \leq m}} \frac{1}{k!\ell!} f^{(k,\ell)}(\hat{x},\hat{y})(x-\hat{x})^k (y-\hat{y})^\ell \\ &= \sum_{\substack{k,\ell \geq 0 \\ k+\ell \leq m}} \frac{1}{k!\ell!} f^{(k,\ell)}(\hat{x},\hat{y}) \left(\left(\frac{\partial}{\partial x}\right)^j (x-\hat{x})^k\right) \left(\left(\frac{\partial}{\partial y}\right)^j (y-\hat{y})^\ell\right). \end{aligned}$$

Es gilt

$$\left(\left(\frac{\partial}{\partial x}\right)^i (x-\hat{x})^k\right) \left(\left(\frac{\partial}{\partial y}\right)^j (y-\hat{y})^\ell\right) = \begin{cases} 0 & \text{falls } k < i \text{ oder } \ell < j \\ i!j! & \text{falls } k = i \text{ und } \ell = j \end{cases}.$$

Für $k > i$ oder $\ell > j$ ist

$$\left(\left(\frac{\partial}{\partial x}\right)^i (x-\hat{x})^k\right) \left(\left(\frac{\partial}{\partial y}\right)^j (y-\hat{y})^\ell\right) = 0$$

für $x = \hat{x}$ und $y = \hat{y}$. Somit bleibt vom Taylor Polynom ausgewertet bei $(\hat{x},\hat{y})$ nur der Summand mit $k = i$ und $\ell = j$ übrig, d.h.

$$\begin{aligned} p^{(i,j)}(\hat{x},\hat{y}) &= \frac{1}{i!j!} f^{(i,j)}(\hat{x},\hat{y}) i!j! \\ &= f^{(i,j)}(\hat{x},\hat{y}). \end{aligned}$$

Beispiel 3.23 Sei

$$f(x,y) \;=\; e^{x^2+y}.$$

Gesucht ist das Taylor Polynom vom Grad $m = 2$ zum Entwicklungspunkt $(\hat{x}, \hat{y}) = (1, 0)$. Die partiellen Ableitungen sind

$$\begin{aligned}
f^{(1,0)}(x,y) &= \frac{\partial}{\partial x} f(x,y) \;=\; 2xe^{x^2+y} \\
f^{(0,1)}(x,y) &= \frac{\partial}{\partial y} f(x,y) \;=\; e^{x^2+y} \\
f^{(1,1)}(x,y) &= \frac{\partial}{\partial x}\frac{\partial}{\partial y} f(x,y) \;=\; 2xe^{x^2+y} \\
f^{(2,0)}(x,y) &= \left(\frac{\partial}{\partial x}\right)^2 f(x,y) \;=\; 2e^{x^2+y} + 4xe^{x^2+y} \\
f^{(0,2)}(x,y) &= \left(\frac{\partial}{\partial y}\right)^2 f(x,y) \;=\; e^{x^2+y}.
\end{aligned}$$

Auswerten im Entwicklungspunkt gibt

$$\begin{aligned}
f(\hat{x},\hat{y}) &= e \\
f^{(1,0)}(\hat{x},\hat{y}) &= 2e \\
f^{(0,1)}(\hat{x},\hat{y}) &= e \\
f^{(1,1)}(\hat{x},\hat{y}) &= 2e \\
f^{(2,0)}(\hat{x},\hat{y}) &= 6e \\
f^{(0,2)}(\hat{x},\hat{y}) &= e.
\end{aligned}$$

Damit ist

$$\begin{aligned}
p(x,y) &= f(\hat{x},\hat{y}) + f^{(1,0)}(\hat{x},\hat{y})(x-\hat{x}) + f^{(0,1)}(\hat{x},\hat{y})(y-\hat{y}) \\
&+ f^{(1,1)}(\hat{x},\hat{y})(x-\hat{x})(y-\hat{y}) \\
&+ {}^1\!/_2 f^{(2,0)}(\hat{x},\hat{y})(x-\hat{x})^2 + {}^1\!/_2 f^{(0,2)}(\hat{x},\hat{y})(y-\hat{y})^2 \\
&= e + 2e(x-1) + ey + 2e(x-1)y + {}^{6e(x-1)^2}\!/_2 + {}^{ey^2}\!/_2 \\
&= e + 2ex - 2e + ey + 2exy - 2ey + 3ex^2 - 6ex + 3e + {}^{ey^2}\!/_2 \\
&= 2e - 4ex - ey + 2exy + 3ex^2 + {}^{ey^2}\!/_2.
\end{aligned}$$

n-stellige Taylor Polynome

Bei n-stelligen Taylor Polynomen tritt eine Summe über n Indizes auf, was etwas unübersichtlich wird. Daher zur Vereinfachung der Notation ein paar Abkürzungen:

$$\begin{aligned}
\underline{i} &= (i_1, i_2, \ldots, i_n) \\
f^{(\underline{i})} &= f^{(i_1,i_2,\ldots,i_n)} = \left(\frac{\partial}{\partial x_1}\right)^{i_1}\left(\frac{\partial}{\partial x_2}\right)^{i_2}\cdots\left(\frac{\partial}{\partial x_n}\right)^{i_n} f \\
\underline{i}! &= i_1! i_2! \ldots i_n! \\
(\vec{x}-\hat{\vec{x}})^{\underline{i}} &= (x_1-\hat{x}_1)^{i_1}(x_2-\hat{x}_2)^{i_2}\cdots(x_n-\hat{x}_n)^{i_n}.
\end{aligned}$$

Weiterhin steht $\underline{i} \leq m$ für

$$i_1 + i_2 + \ldots + i_n \quad \leq \quad m.$$

Definition 3.24 (n-stelliges Taylor Polynom)
Das Taylor Polynom vom Grad m an $f \in \mathbb{R}^n \to \mathbb{R}$ im Arbeitspunkt $\hat{\vec{x}}$ ist definiert durch

$$p(\vec{x}) \quad = \quad \sum_{\underline{k}\leq m} \frac{1}{\underline{k}!} f^{(\underline{k})}(\hat{\vec{x}})(\vec{x}-\hat{\vec{x}})^{\underline{k}}$$

Theorem 3.25
Sei $p(\vec{x})$ das Taylor Polynom vom Grad m an $f(\vec{x})$ im Arbeitspunkt $\hat{\vec{x}}$. Dann gilt

$$p^{(\underline{i})}(\hat{\vec{x}}) \quad = \quad f^{(\underline{i})}(\hat{\vec{x}}) \qquad \text{für alle } \underline{i} \leq m.$$

Beweis. Für die $\underline{i}$-te Ableitung des Taylor Polynoms gilt

$$\begin{aligned}
p^{(\underline{i})}(\vec{x}) &= \left(\frac{\partial}{\partial x_1}\right)^{i_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{i_n} \sum_{\underline{k}\leq m} \frac{1}{\underline{k}!} f^{(\underline{k})}(\hat{\vec{x}})(x_1-\hat{x}_1)^{k_1}\cdots(x_n-\hat{x}_n)^{k_n} \\
&= \sum_{\underline{k}\leq m} \frac{1}{\underline{k}!} f^{(\underline{k})}(\hat{\vec{x}}) \left(\left(\frac{\partial}{\partial x_1}\right)^{i_1}(x_1-\hat{x}_1)^{k_1}\right)\cdots\left(\left(\frac{\partial}{\partial x_n}\right)^{i_n}(x_n-\hat{x}_n)^{k_n}\right) \\
&= \sum_{\underline{k}\leq m} \frac{1}{\underline{k}!} f^{(\underline{k})}(\hat{\vec{x}}) \prod_{\ell=1}^{n}\left(\frac{\partial}{\partial x_\ell}\right)^{i_\ell}(x_\ell-\hat{x}_\ell)^{k_\ell}.
\end{aligned}$$

Es gilt

$$\prod_{\ell=1}^{n}\left(\frac{\partial}{\partial x_\ell}\right)^{i_\ell}(x_\ell-\hat{x}_\ell)^{k_\ell} \quad = \quad \begin{cases} 0 & \text{falls } k_\ell < i_\ell \text{ für ein } \ell = 1,\ldots,n \\ \underline{i}! & \text{falls } k_\ell = i_\ell \text{ für alle } \ell = 1,\ldots,n. \end{cases}$$

Falls $k_\ell > i_\ell$ für ein ℓ gilt

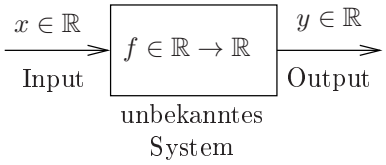
$$\prod_{\ell=1}^{n}\left(\frac{\partial}{\partial x_\ell}\right)^{i_\ell}(x_\ell-\hat{x}_\ell)^{k_\ell} \quad = \quad 0 \qquad \text{für } x_\ell = \hat{x}_\ell.$$

Somit bleibt vom Taylor Polynom ausgewertet bei $\hat{\vec{x}}$ nur der Summand mit $\underline{k} = \underline{i}$ übrig, d.h.

$$p^{(\underline{i})}(\hat{\vec{x}}) \quad = \quad \frac{1}{\underline{i}!} f^{(\underline{i})}(\hat{\vec{x}}) \underline{i}! \quad = \quad f^{(\underline{i})}(\hat{\vec{x}}).$$

3.6 Ausgleichsrechnung

In der Praxis werden häufig Probleme folgender Art gestellt: Von einem unbekannten System, das eine Eingangsgröße $x \in \mathbb{R}$ auf eine Ausgangsgröße y abblidet, sind Messerte $x^{(i)}, y^{(i)}$, $i = 1, \ldots, m$ gegeben.



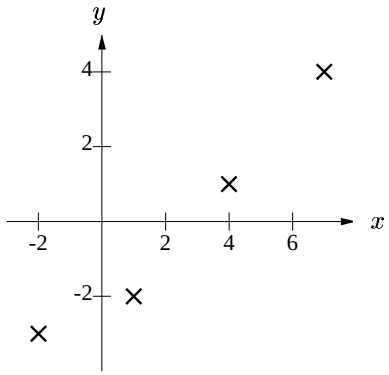
Gesucht ist eine einfache Funktion $f \in \mathbb{R} \to \mathbb{R}$, die das System möglichst gut modelliert, d.h. dass

$$f(x^{(i)}) \quad \approx \quad y^{(i)} \text{ für alle Messwerte } i = 1, \ldots, m$$

wobei der Fehler möglichst klein sein soll.

Beispiel 3.26

i	$x^{(i)}$	$y^{(i)}$
1	-2	-3
2	1	-2
3	4	1
4	7	4



Das Problem ist in dieser Form natürlich nicht eindeutig lösbar. Es ist nicht klar, was man unter einer "einfachen" Funktion oder einem "möglichst kleinen Fehler" zu verstehen hat. Letzteres wird in der Regel dadurch konkretisiert, dass man fordert, dass der quadratische Fehler

$$\sum_{i=1}^{m} \left(f(x^{(i)}) - y^{(i)} \right)^2$$

minimal sein soll. Als "einfache" Funktionen werden in der Regel Polynome mit niedrigem Grad herangezogen, also z.B.

$$f(x) \quad = \quad ax + b.$$

Gesucht sind dann die Parameter a und b so dass

$$\sum_{i=1}^{m} \left(ax^{(i)} + b - y^{(i)} \right)^2$$

minimal ist. In dieser Form ist das o.g. Beispiel eindeutig lösbar und man erhält

$$a = 0.8, \quad b = -2, \quad f(x) = 0.8x - 2.$$

Diese Funktion nennt man Ausgleichsgerade.

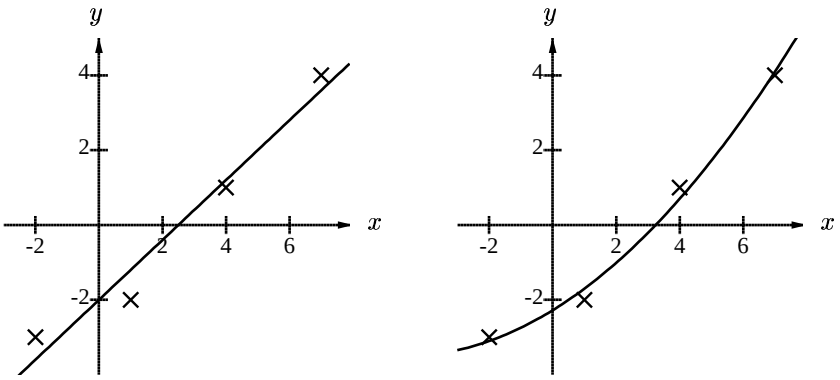
Eine genauere Approximation kann man erreichen, wenn man ein Polynom höheren Grades nimmt, z.B.

$$f(x) \quad = \quad ax^2 + bx + c.$$

die Lösung ist in diesem Fall

$$a = 0.056, \quad b = 0.52, \quad c = -2.28, \quad f(x) = 0.056x^2 + 0.52x - 2.28.$$

Diese Funktion nennt man Ausgleichsparabel.



Allgemein setzt man voraus, dass $f(x)$ eine parametrische Funktion ist, d.h.

$$f(x) \quad = \quad c_1 f_1(x) + c_2 f_2(x) + \ldots + c_n f_n(x)$$

wobei $f_i(x)$ vorgegebene, i.a. einfache Funktionen sind. Das Optimierungsproblem

$$\sum_{i=1}^{m} \left(f(x^{(i)}) - y^{(i)} \right)^2 \quad = \quad \min$$

läuft dann auf die Berechnung der Parameter $c_1, \ldots, c_n$ hinaus.

Im Fall der Ausgleichsgeraden ist $n = 2$ und

$$f_1(x) = x, \quad f_2(x) = 1, \quad f(x) = c_1 x + c_2.$$

Im Fall der Ausgleichsparabel ist $n = 3$ und

$$f_1(x) = x^2, \quad f_2(x) = x, \quad f_3(x) = 1, \quad f(x) = c_1 x^2 + c_2 x + c_3.$$

Das Verfahren funktioniert übrigens völlig analog wenn f eine mehrstellige Funktion ist, d.h.

$$f(\vec{x}) \quad = \quad c_1 f_1(\vec{x}) + c_2 f_2(\vec{x}) + \ldots + c_n f_n(\vec{x}).$$

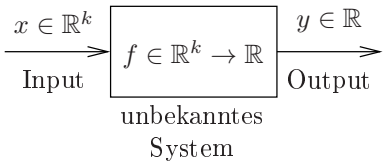
Im allgemeinen Fall hat man Messwerte

$$\vec{x}^{(i)} \in \mathbb{R}^k, \quad y^{(i)} \in \mathbb{R}, \quad i = 1, \ldots, m$$

und eine parametrische Funktion

$$y \quad = \quad c_1 f_1(\vec{x}) + c_2 f_2(\vec{x}) + \ldots + c_n f_n(\vec{x}).$$

wobei die $f_j(\vec{x}) \in \mathbb{R}^k \to \mathbb{R}$, $j = 1, \ldots, n$ fest gewählte Funktionen sind.



Normalerweise hat man mehr Messwerte als unbekannte Parameter, d.h. $m > n$. Setzt man die Messwerte in die Gleichung ein, erhält man ein überbestimmtes *lineares* Gleichungssystem mit den Unbekannten $c_1, \ldots, c_n$:

$$\begin{aligned}
y^{(1)} &= c_1 f_1(\vec{x}^{(1)}) + c_2 f_2(\vec{x}^{(1)}) + \ldots + c_n f_n(\vec{x}^{(1)}) \\
y^{(2)} &= c_1 f_1(\vec{x}^{(2)}) + c_2 f_2(\vec{x}^{(2)}) + \ldots + c_n f_n(\vec{x}^{(2)}) \\
&\vdots \\
y^{(m)} &= c_1 f_1(\vec{x}^{(m)}) + c_2 f_2(\vec{x}^{(m)}) + \ldots + c_n f_n(\vec{x}^{(m)}).
\end{aligned}$$

Dies lässt sich kompakt in vektorieller Schreibweise darstellen:

$$\underbrace{\begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix}}_{\vec{b}} = \underbrace{\begin{pmatrix} f_1\left(\vec{x}^{(1)}\right) & f_2\left(\vec{x}^{(1)}\right) & \ldots & f_n\left(\vec{x}^{(1)}\right) \\ f_1\left(\vec{x}^{(2)}\right) & f_2\left(\vec{x}^{(2)}\right) & \ldots & f_n\left(\vec{x}^{(2)}\right) \\ \vdots & \vdots & \ddots & \vdots \\ f_1\left(\vec{x}^{(m)}\right) & f_2\left(\vec{x}^{(m)}\right) & \ldots & f_n\left(\vec{x}^{(m)}\right) \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}}_{\vec{x}}$$

Um die Notation zu vereinfachen, verwenden wir ab dieser Stelle wieder die in der linearen Algebra üblichen Symbole. Im Folgenden ist also $A \in \mathbb{R}^{m \times n}$ die Koeffizientenmatrix, $\vec{b} \in \mathbb{R}^m$ die rechte Seite und $\vec{x} \in \mathbb{R}^n$ ist der gesuchte Parametervektor.

Das Optimierungsproblem, bei dem die Parameter $c_1, \ldots, c_n$ gesucht sind mit

$$\sum_{i=1}^{m} \left(c_1 f_1(\vec{x}^{(i)}) + c_2 f_2(\vec{x}^{(i)}) + \ldots + c_n f_n(\vec{x}^{(i)}) \; - \; y^{(i)} \right)^2 \quad = \quad \min$$

ist somit äquivalent zu der Bestimmung des Vektors $\vec{x} \in \mathbb{R}^n$ mit

$$\|A\vec{x} - \vec{b}\|^2 \quad = \quad \min.$$

Umformen ergibt

$$\begin{aligned}
\|A\vec{x} - \vec{b}\|^2 &= (A\vec{x} - \vec{b}) \circ (A\vec{x} - \vec{b}) \\
&= (A\vec{x}) \circ (A\vec{x}) - (A\vec{x}) \circ \vec{b} - \vec{b} \circ (A\vec{x}) + \vec{b} \circ \vec{b} \\
&= (A\vec{x})^{\mathrm{T}} A\vec{x} - 2\vec{b} \circ (A\vec{x}) + \|\vec{b}\|^2 \\
&= \vec{x}^{\mathrm{T}} A^{\mathrm{T}} A\vec{x} - 2\vec{b}^{\mathrm{T}} A\vec{x} + \|\vec{b}\|^2.
\end{aligned}$$

In einem globalen Minimum verschwindet der Gradient, d.h.

$$\begin{aligned}\nabla\left(\vec{x}^{\,\mathrm{T}}A^{\mathrm{T}}A\vec{x}-2\vec{b}^{\,\mathrm{T}}A\vec{x}+\|\vec{b}\|\right) &= \vec{0}\\ \nabla\left(\vec{x}^{\,\mathrm{T}}A^{\mathrm{T}}A\vec{x}\right)-2\nabla\left(\vec{b}^{\,\mathrm{T}}A\vec{x}\right) &= \vec{0}\\ \nabla\left(\vec{x}^{\,\mathrm{T}}A^{\mathrm{T}}A\vec{x}\right) &= 2\nabla\left(\vec{b}^{\,\mathrm{T}}A\vec{x}\right).\end{aligned}$$

Die Gradienten wurden in den Beispielen auf Seite 89 und 90 berechnet. Da $A^{\mathrm{T}}A$ symmetrisch ist, gilt

$$\begin{aligned}\nabla\left(\vec{x}^{\,\mathrm{T}}A^{\mathrm{T}}A\vec{x}\right) &= 2A^{\mathrm{T}}A\vec{x}\\ \nabla\left(\vec{b}^{\,\mathrm{T}}A\vec{x}\right) &= A^{\mathrm{T}}\vec{b}.\end{aligned}$$

Damit kommt man auf die sog. Normalgleichungen

$$A^{\mathrm{T}}A\vec{x} = A^{\mathrm{T}}\vec{b}.$$

Ist $A^{\mathrm{T}}A$ regulär, existiert genaue eine Lösung

$$\vec{x} = \left(A^{\mathrm{T}}A\right)^{-1}A^{\mathrm{T}}\vec{b}.$$

Für die Berechnung der Ausgleichsparabel zu Beispiel 3.26 erhält man das überbestimmte LGS

$$\underbrace{\begin{pmatrix}4 & -2 & 1\\ 1 & 1 & 1\\ 16 & 4 & 1\\ 49 & 7 & 1\end{pmatrix}}_{A}\underbrace{\begin{pmatrix}a\\ b\\ c\end{pmatrix}}_{\vec{x}} = \underbrace{\begin{pmatrix}-3\\ -2\\ 1\\ 4\end{pmatrix}}_{\vec{b}}.$$

Die Normalgleichungen sind

$$\underbrace{\begin{pmatrix}26740 & 400 & 70\\ 400 & 70 & 10\\ 70 & 10 & 4\end{pmatrix}}_{A^TA}\underbrace{\begin{pmatrix}a\\ b\\ c\end{pmatrix}}_{\vec{x}} = \underbrace{\begin{pmatrix}198\\ 36\\ 0\end{pmatrix}}_{A^T\vec{b}}$$

mit der Lösung

$$a = 0.0556, \quad b = 0.522, \quad c = -2.28.$$

3.7 Hesse Matrix

Eine notwendige Bedingung, dass eine Funktion $f(\vec{x})$ im Punkt $\hat{\vec{x}}$ einen lokalen Extremwert hat, ist

$$\nabla f(\hat{\vec{x}}) \quad = \quad \vec{0}.$$

In diesem Abschnitt wird untersucht, ob es sich hierbei um ein Maximum, Minimum oder einen Sattelpunkt handelt. Die Idee ist, die Funktion $f(\vec{x})$ im Punkt $\hat{\vec{x}}$ durch ihr Taylor Polynom $p(\vec{x})$ zweiten Grades zu approximieren. Da sich $f(\vec{x})$ in der Nähe von $\hat{\vec{x}}$ gleich verhält wie $p(\vec{x})$, haben $f(\vec{x})$ und $p(\vec{x})$ den selben Typ von Extermwert bei $\hat{\vec{x}}$. Für Polynome kann man jedoch sehr einfach entscheiden, was für ein Extremwert vorliegt.

Einstellige Funktionen

Zunächst wird wieder der Fall $n = 1$ untersucht. Das Taylor Polynom vom Grad 2 an $f(x)$ im Punkt $\hat{x}$ ist

$$p(x) \quad = \quad f(\hat{x}) + f'(\hat{x})(x - \hat{x}) + \frac{1}{2} f''(\hat{x})(x - \hat{x})^2.$$

In einem stationären Punkt, d.h. $f'(\hat{x}) = 0$ gilt

$$p(x) \quad = \quad f(\hat{x}) + \frac{1}{2} f''(\hat{x})(x - \hat{x})^2.$$

Ist nun $f''(\hat{x}) > 0$, dann gilt für alle $x \neq \hat{x}$

$$f''(\hat{x})(x - \hat{x})^2 \quad > \quad 0$$

und somit

$$p(x) \quad > \quad p(\hat{x}).$$

Folglich hat p bei $\hat{x}$ ein Minimum. Für f bedeutet dies, dass für alle x in einer Umgebung von $\hat{x}$ gilt

$$f(x) \quad > \quad f(\hat{x})$$

und somit hat auch f bei $\hat{x}$ ein lokales Minimum. Eine analoge Überlegung gilt für den Fall $f''(\hat{x}) < 0$ und lokale Minima.

Mehrstellige Funktionen

Das Taylor Polynom vom Grad 2 an $f(\vec{x})$ im Punkt $\hat{\vec{x}}$ ist

$$\begin{aligned} p(\vec{x}) &= \sum_{\underline{k}\leq 2} \frac{1}{\underline{k}!} f^{(k)}(\hat{\vec{x}})(\vec{x}-\hat{\vec{x}})^{(\underline{k})} \\ &= f(\hat{\vec{x}}) + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(\hat{\vec{x}})(x_i - \hat{x}_i) + \sum_{i,j=1}^{n} \frac{1}{2} \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} (x_i - \hat{x}_i)(x_j - \hat{x}_j). \end{aligned}$$

Die zweiten partiellen Ableitungen werden nun zu der sog. Hesse Matrix $H_f(\vec{x})$ zusammengefasst, d.h.

$$H_f(\vec{x}) \in \mathbb{R}^{n\times n}, \qquad H_f(\vec{x})_{ij} = \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} f(\vec{x}).$$

Es wird wieder vorausgesetzt, dass die Ableitungen stetig sind. Somit ist $H_f(\vec{x})$ symmetrisch (Satz von Schwarz).

Auf Seite 90 wurde gezeigt, dass

$$\sum_{i,j=1}^{n} a_{ij} x_i x_j = \vec{x}^{\,\mathrm{T}} A \vec{x}.$$

Ersetzt man A durch $H_f(\hat{\vec{x}})$ und $\vec{x}$ durch $\vec{x} - \hat{\vec{x}}$, kann man das Taylor Polynom kompakt darstellen durch

$$p(\vec{x}) = f(\hat{\vec{x}}) + \nabla f(\hat{\vec{x}}) \circ (\vec{x} - \hat{\vec{x}}) + \frac{1}{2} (\vec{x} - \hat{\vec{x}})^{\mathrm{T}} H_f(\hat{\vec{x}}) (\vec{x} - \hat{\vec{x}}).$$

Wenn $\hat{\vec{x}}$ ein stationärer Punkt ist, dann ist $\nabla f(\hat{\vec{x}}) = \vec{0}$ und das Taylor Polynom vereinfacht sich zu

$$p(\vec{x}) = f(\hat{\vec{x}}) + \frac{1}{2} (\vec{x} - \hat{\vec{x}})^{\mathrm{T}} H_f(\hat{\vec{x}}) (\vec{x} - \hat{\vec{x}}).$$

Da $H_f(\hat{\vec{x}})$ symmetrisch ist, sind alle Eigenwerte $\lambda_1, \ldots, \lambda_n$ reell. Weiterhin existieren n zugehörige, reelle und paarweise orthogonale Eigenvektoren $\vec{v}_1, \ldots, \vec{v}_n$. Diese können so skaliert werden, dass jeder die Norm 1 hat. Sei nun

$$T = (\vec{v}_1, \ldots, \vec{v}_n), \qquad D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

Dann gilt mit Theorem 2.10 auf Seite 51

$$T^{\mathrm{T}} H_f(\hat{\vec{x}}) T = D.$$

Da T eine orthogonale Matrix ist und folglich $T^{-1} = T^{\mathrm{T}}$ folgt

$$H_f(\hat{\vec{x}}) = T D T^{\mathrm{T}}.$$

Das Taylor Polynom lässt sich damit wie folgt umformen.

$$\begin{aligned} p(\vec{x}) &= f(\hat{\vec{x}}) + \frac{1}{2} (\vec{x} - \hat{\vec{x}})^{\mathrm{T}} T D T^{\mathrm{T}} (\vec{x} - \hat{\vec{x}}) \\ &= f(\hat{\vec{x}}) + \frac{1}{2} \left(T^{\mathrm{T}} (\vec{x} - \hat{\vec{x}}) \right)^{\mathrm{T}} D \left(T^{\mathrm{T}} (\vec{x} - \hat{\vec{x}}) \right). \end{aligned}$$

Mit der Abkürzung

$$\vec{y} \quad = \quad T^{\mathrm{T}}(\vec{x} - \hat{\vec{x}})$$

vereinfacht sich dies zu

$$\begin{aligned} p(\vec{x}) &= f(\hat{\vec{x}}) + \frac{1}{2}\vec{y}^{\,\mathrm{T}} D\vec{y} \\ &= f(\hat{\vec{x}}) + \frac{1}{2}\sum_{i=1}^{n} y_i^2 \lambda_i. \end{aligned}$$

Sind alle Eigenwerte λ_i positiv, dann ist

$$p(\vec{x}) \quad > \quad p(\hat{\vec{x}}) \quad \text{für alle } \vec{x} \neq \hat{\vec{x}}$$

und damit ist $\hat{x}$ ein Minimum. Für die Funtion f bedeutet dies, dass

$$f(\vec{x}) \quad > \quad f(\hat{\vec{x}}) \quad \text{für alle } \vec{x} \text{ in einer Umgebung von } \hat{\vec{x}}.$$

Somit ist $\vec{x}$ ein lokales Minimum von f. Eine analoge Überlegung gilt wenn alle Eigenwerte λ_i negativ sind. In diesem Fall ist $\hat{x}$ ein lokales Maximum von f.

Hat $H_f(\hat{x})$ sowohl positive als auch negative Eigenwerte, gibt es Vektoren $\vec{x}$ in der Umgebung von $\hat{\vec{x}}$ für die $f(\vec{x}) > f(\hat{\vec{x}})$ und andere, für die $f(\vec{x}) < f(\hat{\vec{x}})$. In diesem Fall ist $\hat{x}$ ein Sattelpunkt von f.

Misslich wird es, wenn einige Eigenwerte Null sind und alle anderen positiv oder alle anderen negativ. In diesem Fall kann man mit der Hessematrix allein keine Aussage treffen. Dies entspricht dem Fall $f''(\hat{x}) = 0$ bei einstelligen Funktionen.

Abschließend werden die Ergebnisse dieses Kapitels zusammengefasst.

Definition 3.27 (Definitheit einer Matrix)
Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt

- *positiv definit, wenn*
$$\vec{x}^{\mathrm{T}} A\vec{x} \;>\; 0 \quad \text{für alle } \vec{x} \neq \vec{0}$$
- *negativ definit, wenn*
$$\vec{x}^{\mathrm{T}} A\vec{x} \;<\; 0 \quad \text{für alle } \vec{x} \neq \vec{0}$$
- *positiv semidefinit, wenn*
$$\vec{x}^{\mathrm{T}} A\vec{x} \;\geq\; 0 \quad \text{für alle } \vec{x} \neq \vec{0}$$
- *negativ semidefinit,*
$$\vec{x}^{\mathrm{T}} A\vec{x} \;\leq\; 0 \quad \text{für alle } \vec{x} \neq \vec{0}$$

und indefinit sonst.

Theorem 3.28
Eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ mit Eigenwerten $\lambda_1, \ldots, \lambda_n$ ist

- *positiv definit gdw $\lambda_i > 0$ für alle i*
- *negativ definit gdw $\lambda_i < 0$ für alle i*
- *positiv semidefinit gdw $\lambda_i \geq 0$ für alle i*
- *negativ semidefinit gdw $\lambda_i \leq 0$ für alle i*
- *indefinit gdw sie positive und negative Eigenwerte hat.*

Beweis. Der Beweis kann wie oben für die Hesse Matrix geführt werden. Da A symmetrisch ist, hat A lauter relle Eigenwerte $\lambda_1, \ldots, \lambda_n$ und zugehörige paarweise orthogonale und normierte Eigenvektoren $\vec{v}_1, \ldots, \vec{v}_n$. Mit

$$T = (\vec{v}_1, \ldots, \vec{v}_n), \qquad D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

gilt somit

$$T^{-1}AT \;=\; D.$$

Da die Spalten von T paarweise orthonormal sind, gilt $T^{-1} = T^{\mathrm{T}}$. Damit ist

$$\begin{aligned} \vec{x}^{\mathrm{T}} A\vec{x} &= \vec{x}^{\mathrm{T}} T D T^{\mathrm{T}} \vec{x} \\ &= \left(T^{\mathrm{T}}\vec{x}\right)^{\mathrm{T}} D \left(T^{\mathrm{T}}\vec{x}\right). \end{aligned}$$

Mit $\vec{y} = T^{\mathrm{T}}\vec{x}$ gilt

$$\vec{x}^{\mathrm{T}} A\vec{x} \;=\; \sum_{i=1}^{n} y_i^2 \lambda_i.$$

Folglich ist

$$\vec{x}^{\mathrm{T}} A \vec{x} \quad > \quad 0 \quad \text{für alle } \vec{x} \neq \vec{0}$$

genau dann wenn $\lambda_i > 0$ für alle i. Die übrigen Fälle gelten analog.

Wie oben gezeigt kann anhand der Definitheit der Hessematrix entschieden werden, ob ein stationärer Punkt ein lokaler Extremwert oder ein Sattelpunkt ist.

Theorem 3.29
Sei $\nabla f(\hat{\vec{x}}) = \vec{0}$.

- *Ist $H_f(\hat{\vec{x}})$ positiv definit, dann hat f an der Stelle $\hat{\vec{x}}$ ein lokales Minimum.*
- *Ist $H_f(\hat{\vec{x}})$ negativ definit, dann hat f an der Stelle $\hat{\vec{x}}$ ein lokales Maximum.*
- *Ist $H_f(\hat{\vec{x}})$ indefinit, dann hat f an der Stelle $\hat{\vec{x}}$ einen Sattelpunkt.*

Ist $H_f(\hat{\vec{x}})$ semidefinit, müssen wie im einstelligen Fall höhere Ableitungen untersucht werden.

Der Vollständigkeit halber sei erwähnt, dass die Definitheit von nicht symmetrischen Matrizen leicht auf die Definitheit von symmetrische Matrizen reduziert werden kann.

Ist $A \in \mathbb{R}^{n \times n}$ eine beliebige Matrix, dann ist

$$A + A^{\mathrm{T}}$$

eine symmetrische Matrix. A ist genau dann positiv bzw. negativ (semi)definit, wenn $A + A^{\mathrm{T}}$ positiv bzw. negativ (semi)definit ist. Dies folgt aus

$$\begin{aligned} \vec{x}^{\mathrm{T}} A \vec{x} &= \sum_{i,j} a_{ij} x_i x_j \\ &= \sum_{i,j} a_{ji} x_i x_j \\ &= \vec{x}^{\mathrm{T}} A^{\mathrm{T}} \vec{x} \end{aligned}$$

und

$$\begin{aligned} \vec{x}^{\mathrm{T}} A \vec{x} &= \frac{1}{2} \left(\vec{x}^{\mathrm{T}} A \vec{x} + \vec{x}^{\mathrm{T}} A \vec{x} \right) \\ &= \frac{1}{2} \left(\vec{x}^{\mathrm{T}} A \vec{x} + \vec{x}^{\mathrm{T}} A^{\mathrm{T}} \vec{x} \right) \\ &= \frac{1}{2} \left(\vec{x}^{\mathrm{T}} (A + A^{\mathrm{T}}) \vec{x} \right). \end{aligned}$$

Das Vorzeichen von $\vec{x}^{\mathrm{T}} A \vec{x}$ und $\vec{x}^{\mathrm{T}} (A + A^{\mathrm{T}}) \vec{x}$ ist somit immer gleich. Daher hat A die gleiche Definitheit wie die symmetrische Matrix $A + A^{\mathrm{T}}$.

3.8 Nichtlineare Gleichungssysteme, Newton Verfahren

vorigen Kapitel haben wir den Begriff der Linearisierung von einstelligen Funktionen auf mehrstellige Funktionen erweitert. Eine wichtige Anwendung von Linearisierungen ist das Newton Verfahren, das wir nun ebenfalls auf mehrstellige Funktionen erweitern werden.

Einstelliges Newton Verfahren

Die Aufgabe ist, eine Nullstelle einer Funktion $f \in \mathbb{R} \to \mathbb{R}$ zu finden. Falls f eine nichtlineare Funktion ist, führt dies auf die Gleichung $f(x) = 0$, die oft nicht geschlossen lösbar ist. Die Vorgehensweise des Newton Verfahrens ist dann wie folgt:

Schritt 1 (Startwert): Wähle einen Startwert $\hat{x}$, der möglichst in der Nähe der Nullstelle liegen sollte.

Schritt 2 (Linearisieren): Berechne die Linearisierung $\ell(x)$ an $f(x)$ im Punkt $\hat{x}$, d.h.

$$\ell(x) \quad = \quad f(\hat{x}) + f'(\hat{x})(x - \hat{x}).$$

Schritt 3 (Lösen): Da die Linearisierung eine gute Approximation an $f(x)$ ist, ist auch die Nullstelle der Linearisierung eine gute Approximation an die Nullstelle an $f(x)$. Diese berechnet sich wie folgt.

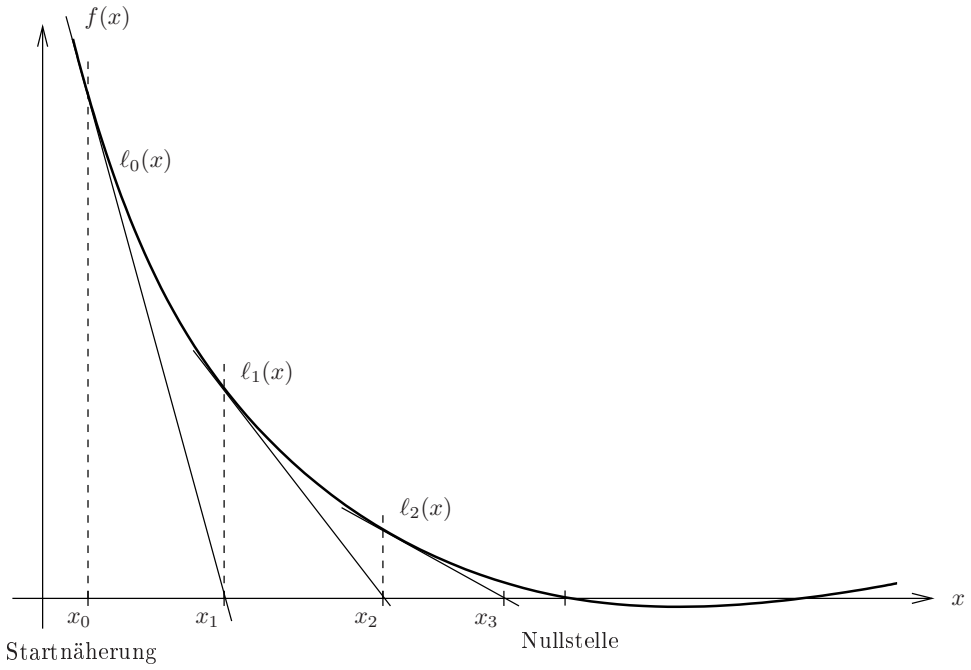
$$\begin{aligned} f(\hat{x}) + f'(\hat{x})(x - \hat{x}) &= 0 \\ f'(\hat{x})(x - \hat{x}) &= -f(\hat{x}) \\ x - \hat{x} &= -\frac{f(\hat{x})}{f'(\hat{x})} \\ x &= \hat{x} - \frac{f(\hat{x})}{f'(\hat{x})}. \end{aligned}$$

Schritt 4 (Iterieren): Nehme die Nullstelle der Linearisierung als neuen Nährungswert $\hat{x}$ und gehe damit zurück zu **Schritt 2**.

Man erhält auf diese Weise eine Folge von Näherungen

$$x_{n+1} \quad = \quad x_n - \frac{f(x_n)}{f'(x_n)}.$$

Leider ist nicht garantiert, dass diese Folge auch tatsächlich zu einer Nullstelle von f konvergiert. Wenn die Startnäherung jedoch hinreichend nah bei einer Nullstelle von f ist, tritt tatsächlich Konvergenz ein.



Beispiel 3.30 Mit dem Newton Verfahren kann man eine gute Näherung an $\sqrt{2}$ berechnen. Dies ist eine Nullstelle von

$$f(x) \quad = \quad x^2 - 2.$$

Mit $f'(x) = 2x$ und dem Startwert $x_0 = 1$ erhält man folgende Näherungen.

$$\begin{aligned}
x_1 &= x_0 - \frac{f(x_0)}{f'(x_0)} = 1 - \frac{-1}{2} = \frac{3}{2} = 1.5 \\
x_2 &= x_1 - \frac{f(x_1)}{f'(x_1)} = \frac{3}{2} - \frac{9/4 - 2}{3} = \frac{17}{12} \approx 1.417 \\
x_3 &= \frac{577}{408} \approx 1.41422 \\
&\vdots
\end{aligned}$$

Mehrstelliges Newton Verfahren

Im mehrstelligen Fall geht es darum, eine gemeinsame Nullstelle von n Funktionen in n Variablen zu finden, d.h. eine Lösung des Gleichungssystems

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0 \\ f_2(x_1, x_2, \ldots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0. \end{aligned}$$

Beispiel 3.31 Sei

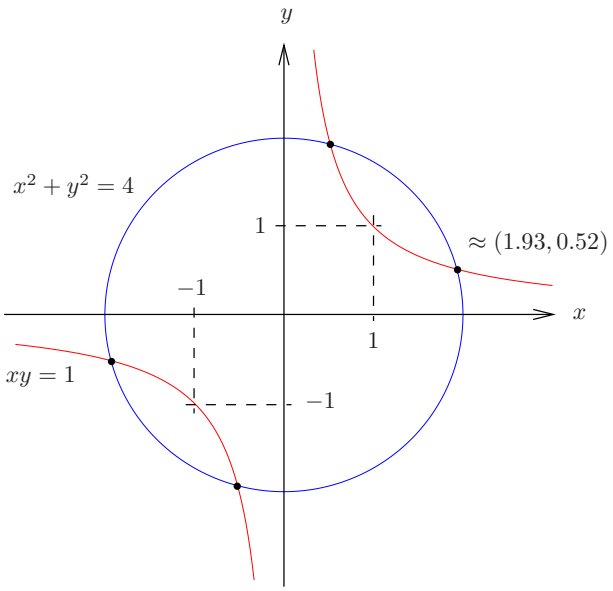
$$\begin{aligned} f_1(x, y) &= x^2 + y^2 - 4 \\ f_2(x, y) &= xy - 1. \end{aligned}$$

Gesucht ist eine Lösung des Gleichungssystems

$$\begin{aligned} f_1(x, y) &= 0 \\ f_2(x, y) &= 0, \end{aligned}$$

d.h. ein Punkt $(\hat{x}, \hat{y})$ so dass

$$\begin{aligned} \hat{x}^2 + \hat{y}^2 - 4 &= 0 \\ \hat{x}\hat{y} - 1 &= 0. \end{aligned}$$



Tatsächlich existieren 4 Lösungen $\hat{x} = \pm\sqrt{2 \pm \sqrt{3}},\ \hat{y} = 1/\hat{x}$.

Wenn die Funktionen $f_1, f_2, \ldots, f_n$ affin linear wären, so könnte man dieses Problem z.B. mit dem Gauß Algorithmus lösen. Im Folgenden lassen wir jedoch allgemeine nichtlineare Funktionen zu, von denen lediglich gefordert ist,

dass alle partiellen Ableitungen existieren. Dadurch wird das Problem erheblich schwieriger — anders als bei linearen Systemen existiert noch nicht einmal ein einfaches Kriterium um zu entscheiden ob Lösungen existieren bzw. wie viele. Die Vorgehensweise zur Lösung eines solchen nichtlinearen Gleichungssystems nach dem Newton Verfahren ist die selbe wie im einstelligen Fall:

Schritt 1 (Startwert): Wähle einen Startwert $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$.

Schritt 2 (Linearisieren): Berechne zu jeder Funktion

$$f_i(x_1, x_2, \ldots, x_n)$$

die Linearisierung

$$\ell_i(x_1, x_2, \ldots, x_n)$$

im Punkt $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ für alle $i = 1, \ldots, n$.

Schritt 3 (Lösen): Anstatt des ursprünglichen nichtlinearen Gleichungssystems löst man nun das linearisierte System

$$\begin{aligned} \ell_1(x_1, x_2, \ldots, x_n) &= 0 \\ \ell_2(x_1, x_2, \ldots, x_n) &= 0 \\ &\vdots \\ \ell_n(x_1, x_2, \ldots, x_n) &= 0 \end{aligned}$$

z.B. mit dem Gauß Algorithmus.

Schritt 4 (Iterieren): Nehme die Lösung dieses linearen Gleichungssystems als neue Näherungslösung $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ und gehe zu **Schritt 2**.

Man erhält somit eine Folge von Näherungslösungen und wie im eindimensionalen Fall ist leider durchaus nicht garantiert, dass diese Folge auch konvergiert. Ist man mit dem Startwert jedoch hinreichend nahe an einer Lösung, so tritt tatsächlich Konvergenz zu der Lösung ein. Ein großer Vorteil des Newton Verfahrens ist, dass man sich dann sehr schnell der Lösung nähert.

Beispiel 3.32 Wir suchen nun eine Lösung des Gleichungssystems von Beispiel 3.31 mit dem Newton Verfahren. Da f_1 und f_2 in jeder Iteration linearisiert werden müssen, berechnen wir diese vorab.

$$\begin{aligned} \frac{\partial}{\partial x} f_1(x, y) &= 2x & \frac{\partial}{\partial y} f_1(x, y) &= 2y \\ \frac{\partial}{\partial x} f_2(x, y) &= y & \frac{\partial}{\partial y} f_2(x, y) &= x \end{aligned}$$

$$\begin{aligned} \ell_1(x, y) &= f_1(\hat{x}, \hat{y}) + \frac{\partial}{\partial x} f_1(\hat{x}, \hat{y})(x - \hat{x}) + \frac{\partial}{\partial y} f_1(\hat{x}, \hat{y})(y - \hat{y}) \\ &= \hat{x}^2 + \hat{y}^2 - 4 + 2\hat{x}(x - \hat{x}) + 2\hat{y}(y - \hat{y}) \\ \ell_2(x, y) &= f_2(\hat{x}, \hat{y}) + \frac{\partial}{\partial x} f_2(\hat{x}, \hat{y})(x - \hat{x}) + \frac{\partial}{\partial y} f_2(\hat{x}, \hat{y})(y - \hat{y}) \\ &= \hat{x}\hat{y} - 1 + \hat{y}(x - \hat{x}) + \hat{x}(y - \hat{y}) \end{aligned}$$

Erste Iteration. Als Startnäherung wird der Punkt $(\hat{x}, \hat{y}) = (2, 1)$ gewählt. Die Linearisierungen von f_1 und f_2 bei $(2, 1)$ sind

$$\begin{aligned} \ell_1(x, y) &= 1 + 4(x - 2) + 2(y - 1) \\ &= -9 + 4x + 2y \\ \ell_2(x, y) &= 1 + 1(x - 2) + 2(y - 1) \\ &= -3 + x + 2y. \end{aligned}$$

Als nächstes muss das lineare Gleichungssystem

$$\begin{aligned} \ell_1(x, y) &= 0 \\ \ell_2(x, y) &= 0 \end{aligned}$$

gelöst werden, d.h. in diesem Fall

$$\begin{aligned} 4x + 2y &= 9 \\ x + 2y &= 3. \end{aligned}$$

Die Lösung ist $x = 2$, $y = 1/2$.

Zweite Iteration. Mit der neuen Näherung $(\hat{x}, \hat{y}) = (2, 1/2)$ beginnt man nun die nächste Iteration des Newton Verfahrens. Die Linearisierungen von f_1 und f_2 bei $(2, 1/2)$ sind

$$\begin{aligned} \ell_1(x, y) &= 1/4 + 4(x - 2) + 1(y - 1/2) \\ &= -33/4 + 4x + y \\ \ell_2(x, y) &= 0 + 1/2(x - 2) + 2(y - 1/2) \\ &= -2 + x/2 + 2y. \end{aligned}$$

Das zu lösende lineare Gleichungssystem

$$\begin{aligned} \ell_1(x, y) &= 0 \\ \ell_2(x, y) &= 0 \end{aligned}$$

ist nun

$$\begin{aligned} 4x + y &= 33/4 \\ x/2 + 2y &= 2. \end{aligned}$$

Die Lösung ist $x = 29/15 \approx 1.93$, $y = 31/60 \approx 0.52$. Damit ist man nach nur zwei Iterationen schon sehr nah bei einer Lösung des ursprünglichen nichtlinearen Gleichungssystems.

Das eindimensionale Newton Verfahren ließ sich kompakt in der Form

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

schreiben. Eine ähnliche Notation ist auch für das mehrstellige Newton Verfahren möglich. Hierzu wird zunächst die vektorielle Schreibweise eingeführt:

$$\vec{f}(\vec{x}) = \begin{pmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_n(x_1, \ldots, x_n) \end{pmatrix}.$$

Zu lösen ist also das Gleichungssystem

$$\vec{f}(\vec{x}) = \vec{0}.$$

Als nächstes werden die partiellen Ableitungen von $f_1, f_2, \ldots, f_n$ zu einer Matrix zusammengefaßt, wobei in der i-ten Zeile die partiellen Ableitungen von f_i stehen. Diese Matrix heißt Jacobi Matrix von $\vec{f}$.

Definition 3.33 (Jacobi Matrix)
Die Jacobi Matrix von $\vec{f}$ ist

$$J_{\vec{f}}(\vec{x}) = \begin{pmatrix} \nabla f_1(\vec{x})^{\mathrm{T}} \\ \vdots \\ \nabla f_n(\vec{x})^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} \partial/\partial x_1 f_1(\vec{x}) & \ldots & \partial/\partial x_n f_1(\vec{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} f_n(\vec{x}) & \ldots & \frac{\partial}{\partial x_n} f_n(\vec{x}) \end{pmatrix}.$$

Die i-te Zeile der Jacobi Matrix ist also nichts anderes als der Gradient von f_i als Zeilenvektor. Die Linearisierung $\ell_i(\vec{x})$ von $f_i(\vec{x})$ an der Stelle $\hat{\vec{x}}$ lässt sich unter Verwendung des Gradienten schreiben als

$$\ell_i(\vec{x}) = f_i(\hat{\vec{x}}) + \nabla f_i(\hat{\vec{x}}) \circ (\vec{x} - \hat{\vec{x}}).$$

Fasst man nun auch die Linearisierungen zu einem Vektor zusammen, erhält man

$$\begin{aligned}
\vec{\ell}(\vec{x}) &= \begin{pmatrix} \ell_1(\vec{x}) \\ \vdots \\ \ell_n(\vec{x}) \end{pmatrix} \\
&= \begin{pmatrix} f_1(\hat{\vec{x}}) + \nabla f_1(\hat{\vec{x}}) \circ (\vec{x} - \hat{\vec{x}}) \\ \vdots \\ f_n(\hat{\vec{x}}) + \nabla f_n(\hat{\vec{x}}) \circ (\vec{x} - \hat{\vec{x}}) \end{pmatrix} \\
&= \begin{pmatrix} f_1(\hat{\vec{x}}) \\ \vdots \\ f_n(\hat{\vec{x}}) \end{pmatrix} + \begin{pmatrix} \nabla f_i(\hat{\vec{x}})^{\mathrm{T}}(\vec{x} - \hat{\vec{x}}) \\ \vdots \\ \nabla f_n(\hat{\vec{x}})^{\mathrm{T}}(\vec{x} - \hat{\vec{x}}) \end{pmatrix} \\
&= \begin{pmatrix} f_1(\hat{\vec{x}}) \\ \vdots \\ f_n(\hat{\vec{x}}) \end{pmatrix} + \begin{pmatrix} \nabla f_i(\hat{\vec{x}})^{\mathrm{T}} \\ \vdots \\ \nabla f_n(\hat{\vec{x}})^{\mathrm{T}} \end{pmatrix} (\vec{x} - \hat{\vec{x}}) \\
&= \vec{f}(\hat{\vec{x}}) + J_{\vec{f}}(\hat{\vec{x}})(\vec{x} - \hat{\vec{x}}).
\end{aligned}$$

Das lineare Gleichungssystem

$$\vec{\ell}(\vec{x}) = \vec{0}$$

lässt sich nun unter Verwendung der inversen Jacobi Matrix wie folgt auflösen:

$$\begin{aligned}
\vec{f}(\hat{\vec{x}}) + J_{\vec{f}}(\hat{\vec{x}})(\vec{x} - \hat{\vec{x}}) &= \vec{0} \\
J_{\vec{f}}(\hat{\vec{x}})(\vec{x} - \hat{\vec{x}}) &= -\vec{f}(\hat{\vec{x}}) \\
\vec{x} - \hat{\vec{x}} &= -\left(J_{\vec{f}}(\hat{\vec{x}})\right)^{-1} \vec{f}(\hat{\vec{x}}) \\
\vec{x} &= \hat{\vec{x}} - \left(J_{\vec{f}}(\hat{\vec{x}})\right)^{-1} \vec{f}(\hat{\vec{x}}).
\end{aligned}$$

Damit lässt sich die Iterationsvorschrift des mehrstelligen Newton Verfahrens kompakt schreiben durch

$$\vec{x}_{n+1} = \vec{x}_n - \left(J_{\vec{f}}(\vec{x}_n)\right)^{-1} \vec{f}(\vec{x}_n).$$

Vergleicht man diese Vorschrift mit dem eindimensionalen Newton Verfahren

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

so stellt man fest, dass lediglich die Division durch $f'(x_n)$ durch die Multiplikation mit $\left(J_{\vec{f}}(\vec{x}_n)\right)^{-1}$ ersetzt wurde. Die Bedingung, dass $f'(x_n) \neq 0$ ist, entspricht im mehrdimensionalen Fall der Bedingung, dass die Jacobi Matrix regulär ist.

A Anhang

A.1 Rechengesetze für Faltung und Dirac Impuls

- **Kommutativgesetz**

$$f * g = g * f$$

- **Assoziativgesetz**

$$(f * g) * h = f * (g * h)$$

- **Linearität (inkl. Distributivgesetz)**

$$\begin{aligned}(f_1 + f_2) * g &= f_1 * g + f_2 * g \\ (af) * g &= a(f * g)\end{aligned}$$

- **Zeitinvarianz**

$$f_{\hat{t}} * g = (f * g)_{\hat{t}} = f * g_{\hat{t}}$$

- **Vertauschbarkeit von Faltung und Ableitung**

$$f' * g = (f * g)' = f * g'$$

- **Integration durch Faltung mit Sprungfunktion**

$$(f * \sigma)(t) = \int_{-\infty}^{t} f(\tau) d\tau.$$

- **Neutrales Element**

$$\begin{aligned}f * \delta &= f \\ f * \delta_{\hat{t}} &= f_{\hat{t}}\end{aligned}$$

- **Verallgemeinerte Ableitung**

$$\sigma'(t) = \delta(t)$$

- **Ausblendeigenschaft (falls f stetig bei $t = a$)**

$$\begin{aligned}f(t)\delta(t-a) &= f(a)\delta(t-a) \\ \int_{-\infty}^{\infty} f(t)\delta(t-a) &= f(a)\end{aligned}$$

A.2 Die wichtigsten Fourier Transformationspaare

$$\delta(t) \quad \circ\!\!-\!\!\bullet \quad 1$$

$$1 \quad \circ\!\!-\!\!\bullet \quad 2\pi\delta(\omega)$$

$$e^{j\hat{\omega}t} \quad \circ\!\!-\!\!\bullet \quad 2\pi\delta(\omega - \hat{\omega})$$

$$\sigma(t)e^{at} \quad \circ\!\!-\!\!\bullet \quad \frac{1}{j\omega - a} \qquad \text{falls } a < 0$$

$$\cos(\hat{\omega}t) \quad \circ\!\!-\!\!\bullet \quad \pi(\delta(\omega - \hat{\omega}) + \delta(\omega + \hat{\omega}))$$

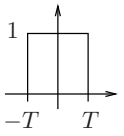
$$\sin(\hat{\omega}t) \quad \circ\!\!-\!\!\bullet \quad -j\pi(\delta(\omega - \hat{\omega}) - \delta(\omega + \hat{\omega}))$$

$$\text{sign}(t) \quad \circ\!\!-\!\!\bullet \quad \frac{2}{j\omega}$$

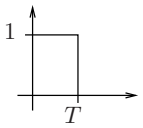
$$\frac{1}{t} \quad \circ\!\!-\!\!\bullet \quad -j\pi\text{sign}(\omega)$$

$$\sigma(t) \quad \circ\!\!-\!\!\bullet \quad \frac{1}{j\omega} + \pi\delta(\omega)$$

$$\frac{j}{\pi t} + \delta(t) \quad \circ\!\!-\!\!\bullet \quad 2\sigma(\omega)$$

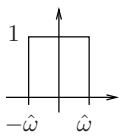


$$\circ\!\!-\!\!\bullet \quad 2T\text{si}(\omega T)$$



$$\circ\!\!-\!\!\bullet \quad \begin{cases} \frac{1}{j\omega}(1 - e^{-j\omega T}) & \text{falls } \omega \neq 0 \\ T & \text{falls } \omega = 0 \end{cases}$$

$$\frac{\hat{\omega}}{\pi}\text{si}(\hat{\omega}t) \quad \circ\!\!-\!\!\bullet$$



$$T_s \sum_{n=-\infty}^{\infty} \delta(t - nT_s) = \sum_{k=-\infty}^{\infty} e^{jk\omega_s t} \quad \circ\!\!-\!\!\bullet \quad 2\pi \sum_{k=-\infty}^{\infty} \delta(\omega - k\omega_s), \quad \omega_s = \frac{2\pi}{T_s}$$

A.3 Rechengesetze für die Fourier Transformation

Symmetrie

$$\begin{aligned} f(t) \text{ reell} &\quad \circ\!\!-\!\!\bullet \quad F(-\omega) = \overline{F(\omega)} \\ f(t) \text{ reell, gerade} &\quad \circ\!\!-\!\!\bullet \quad F(\omega) \text{ reell, gerade} \\ f(t) \text{ reell, ungerade} &\quad \circ\!\!-\!\!\bullet \quad F(\omega) \text{ imaginär, ungerade} \end{aligned}$$

Linearität

$$\begin{aligned} f(t) + g(t) &\quad \circ\!\!-\!\!\bullet \quad F(\omega) + G(\omega) \\ af(t) &\quad \circ\!\!-\!\!\bullet \quad aF(\omega) \end{aligned}$$

Zeitverschiebung

$$f(t - \hat{t}) \quad \circ\!\!-\!\!\bullet \quad e^{-j\omega\hat{t}} F(\omega)$$

Frequenzverschiebung

$$f(t)e^{j\hat{\omega}t} \quad \circ\!\!-\!\!\bullet \quad F(\omega - \hat{\omega})$$

Modulation

$$f(t)\cos(\hat{\omega}t) \quad \circ\!\!-\!\!\bullet \quad \frac{1}{2}(F(\omega - \hat{\omega}) + F(\omega + \hat{\omega}))$$

Zeitumkehr

$$f(-t) \quad \circ\!\!-\!\!\bullet \quad F(-\omega)$$

Zeitdehnung

$$f(at) \quad \circ\!\!-\!\!\bullet \quad \frac{1}{|a|} F\left(\frac{\omega}{a}\right)$$

Ableitung im Zeitbereich

$$\begin{aligned} f'(t) &\quad \circ\!\!-\!\!\bullet \quad (j\omega)F(\omega) \\ f''(t) &\quad \circ\!\!-\!\!\bullet \quad (j\omega)^2 F(\omega) \end{aligned}$$

Integration im Zeitbereich

$$\int_{-\infty}^{t} f(u)du \quad \circ\!\!-\!\!\bullet \quad \left(\frac{1}{j\omega} + \pi\delta(\omega)\right) F(\omega)$$

Ableitung im Frequenzbereich

$$\begin{aligned} (-jt)f(t) &\quad \circ\!\!-\!\!\bullet \quad F'(\omega) \\ (-jt)^2 f(t) &\quad \circ\!\!-\!\!\bullet \quad F''(\omega) \end{aligned}$$

Faltung im Zeitbereich

$$(f * g)(t) \quad \circ\!\!-\!\!\bullet \quad F(\omega)G(\omega)$$

Faltung im Frequenzbereich

$$f(t)g(t) \quad \circ\!\!-\!\!\bullet \quad \frac{1}{2\pi}(F * G)(\omega)$$

Abtastung

$$\underbrace{f(t) \sum_{n=-\infty}^{\infty} \delta(t - nT_s)}_{\text{Abtastung}} = \sum_{n=-\infty}^{\infty} f_n \delta(t - nT_s) \quad \circ\!\!-\!\!\bullet \quad \frac{1}{T_s} \underbrace{\sum_{k=-\infty}^{\infty} F(\omega - k\omega_s)}_{\text{periodische Fortsetzung}}$$

Theorem von Parseval

$$\int_{-\infty}^{\infty} |f(t)|^2 \quad = \quad \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(\omega)|^2 d\omega$$

A.4 Die wichtigsten Laplace Transformationspaare

$$\sigma(t) \quad \circ\!\!-\!\!\bullet \quad \frac{1}{s}$$

$$\delta(t) \quad \circ\!\!-\!\!\bullet \quad 1$$

$$t\,\sigma(t) \quad \circ\!\!-\!\!\bullet \quad \frac{1}{s^2}$$

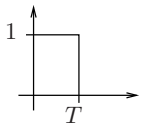
$$t^n\,\sigma(t) \quad \circ\!\!-\!\!\bullet \quad \frac{n!}{s^{n+1}}, \qquad n \in \mathbb{N}$$

$$e^{at}\,\sigma(t) \quad \circ\!\!-\!\!\bullet \quad \frac{1}{s-a}$$

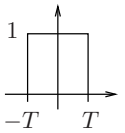
$$t^n e^{at}\sigma(t) \quad \circ\!\!-\!\!\bullet \quad \frac{n!}{(s-a)^{n+1}}$$

$$\sin(\omega t)\,\sigma(t) \quad \circ\!\!-\!\!\bullet \quad \frac{\omega}{s^2+\omega^2}$$

$$\cos(\omega t)\,\sigma(t) \quad \circ\!\!-\!\!\bullet \quad \frac{s}{s^2+\omega^2}$$



$$\circ\!\!-\!\!\bullet \quad \frac{1-e^{-sT}}{s}$$



$$\circ\!\!-\!\!\bullet \quad \frac{e^{sT}-e^{-sT}}{s}$$

$$\delta(t-\hat{t}) \quad \circ\!\!-\!\!\bullet \quad e^{-s\hat{t}}$$

$$\lfloor t \rfloor\,\sigma(t) \quad \circ\!\!-\!\!\bullet \quad \frac{1}{s(e^s-1)}$$

$$p(t) = \sum_{n=0}^{\infty} \delta(t-nT) \quad \circ\!\!-\!\!\bullet \quad \frac{1}{1-e^{-sT}}$$

A.5 Rechengesetze für die Laplace Transformation

- Falls

$$f(t) \quad \circ\!\!-\!\!\bullet \quad F(s)$$

Linearität

$$\begin{aligned} f(t) + g(t) \quad &\circ\!\!-\!\!\bullet \quad F(s) + G(s) \\ af(t) \quad &\circ\!\!-\!\!\bullet \quad aF(s) \end{aligned}$$

Verschiebung

$$f(t - \hat{t}) \quad \circ\!\!-\!\!\bullet \quad e^{-s\hat{t}} F(s)$$

Ähnlichkeit

$$f(at) \quad \circ\!\!-\!\!\bullet \quad \frac{1}{a} F\left(\frac{s}{a}\right), \quad a > 0$$

Dämpfung

$$e^{-at} f(t) \quad \circ\!\!-\!\!\bullet \quad F(s + a)$$

Ableitung im Zeitbereich

$$f'(t) \quad \circ\!\!-\!\!\bullet \quad sF(s)$$

Ableitung im Frequenzbereich

$$tf(t) \quad \circ\!\!-\!\!\bullet \quad -F'(s)$$

Integration im Zeitbereich

$$\int_{-\infty}^{t} f(u)du \quad \circ\!\!-\!\!\bullet \quad \frac{1}{s} F(s)$$

Faltung

$$(f * g)(t) \quad \circ\!\!-\!\!\bullet \quad F(s)G(s)$$

- Falls

$$\sigma(t) f(t) \quad \circ\!\!-\!\!\bullet \quad F(s)$$

Verschiebung

$$\sigma(t - \hat{t}) f(t - \hat{t}) \quad \circ\!\!-\!\!\bullet \quad e^{-s\hat{t}} F(s)$$

Ableitung im Zeitbereich

$$\begin{aligned} \sigma(t) f'(t) \quad &\circ\!\!-\!\!\bullet \quad sF(s) - f(0^-) \\ \sigma(t) f''(t) \quad &\circ\!\!-\!\!\bullet \quad s^2 F(s) - sf(0^-) - f'(0^-) \end{aligned}$$

A.6 Die wichtigsten z-Transformationspaare

$$
\begin{aligned}
\delta_k &\circ\!\!-\!\!\bullet\; 1 \\
\sigma_k &\circ\!\!-\!\!\bullet\; \frac{z}{z-1} \\
\sigma_k\, k &\circ\!\!-\!\!\bullet\; \frac{z}{(z-1)^2} \\
\sigma_k\, k^2 &\circ\!\!-\!\!\bullet\; \frac{z(z+1)}{(z-1)^3} \\
\sigma_{k-1} &\circ\!\!-\!\!\bullet\; \frac{1}{z-1} \\
\sigma_k\, a^k &\circ\!\!-\!\!\bullet\; \frac{z}{z-a} \\
\sigma_k\, e^{ak} &\circ\!\!-\!\!\bullet\; \frac{z}{z-e^a} \\
a^{k-n}\binom{k-1}{n-1} &\circ\!\!-\!\!\bullet\; \frac{1}{(z-a)^n}
\end{aligned}
$$

A.7 Rechengesetze für die z-Transformation

- Falls

$$f_k \quad \circ\!\!-\!\!\bullet \quad F(z)$$

Linearität

$$\begin{aligned} f_k + g_k &\quad \circ\!\!-\!\!\bullet \quad F(z) + G(z) \\ a f_k &\quad \circ\!\!-\!\!\bullet \quad a F(z) \end{aligned}$$

Dämpfung

$$a^k f_k \quad \circ\!\!-\!\!\bullet \quad F\left(\frac{z}{a}\right)$$

Zeitverschiebung

$$f_{k-m} \quad \circ\!\!-\!\!\bullet \quad z^{-m} F(z), \qquad m \in \mathbb{Z}$$

Ableitung im Frequenzbereich

$$k f_k \quad \circ\!\!-\!\!\bullet \quad -z F'(z)$$

Faltung

$$(f * g)_k \quad \circ\!\!-\!\!\bullet \quad F(z) G(z)$$

- Falls

$$\sigma_k f_k \quad \circ\!\!-\!\!\bullet \quad F(z)$$

Zeitverschiebung

$$\begin{aligned} \sigma_{k-m} f_{k-m} &\quad \circ\!\!-\!\!\bullet \quad z^{-m} F(z), \qquad m \in \mathbb{Z} \\ \sigma_k f_{k-m} &\quad \circ\!\!-\!\!\bullet \quad z^{-m} \left(F(z) + \sum_{k=-m}^{-1} f_k z^{-k} \right), \qquad m \geq 0 \\ \sigma_k f_{k+m} &\quad \circ\!\!-\!\!\bullet \quad z^{m} \left(F(z) - \sum_{k=0}^{m-1} f_k z^{-k} \right), \qquad m \geq 0 \end{aligned}$$

Index

Literatur

[1] Brigola: *Fourier-Analysis und Distributionen.* editon swk, 2012

[2] Burg ; Haf ; Wille: *Höhere Mathematik für Ingenieure, Band 3.* Teubner, 1993

[3] Burg ; Haf ; Wille: *Höhere Mathematik für Ingenieure, Band 1.* Teubner, 1997

[4] Girod ; Rabenstein ; Stenger: *Einführung in die Systemtheorie.* Teubner, 2007

[5] Glatz ; Grieb ; Hohloch ; Kümmerer ; Mohr: *Brücken zur Mathematik Band 7, Fourier Analysis.* Cornelsen, 1996

[6] Richards ; Youn: *Theory of Distributions.* Cambridge University Press, 1990