Numerische Mathematik

Eine Vorlesung für das

Lehramtsstudium

Franz Hofbauer

I. Rundungsfehler

Beim Rechnen mit reellen Zahlen muss man auf eine gewisse Anzahl von Dezimalstellen runden, da man nicht beliebig viele Dezimalstellen aufschreiben kann. Mit den dabei auftretenden Problemen beschäftigen wir uns in diesem ersten Kapitel.

1. Gleitkommaarithmetik

Ein Computer arbeitet mit ganzen Zahlen (integer), mit denen exakt gerechnet wird, und mit reellen Zahlen (real), mit denen nicht exakt gerechnet werden kann, da man sich auf endlich viele Kommastellen beschränken muss. Reelle Zahlen kann man in Festkommadarstellung aufschreiben, das heißt es stehen n_1 Stellen vor dem Komma und n_2 Stellen nach dem Komma zur Verfügung. Für $n_1 = 4$ und $n_2 = 6$ sind 12.160473 und -3502.896043 Beispiele für Zahlen in Festkommadarstellung.

Computer verwenden üblicherweise die Gleitkommadarstellung von reellen Zahlen. Eine reelle Zahl x wird geschrieben als $x = z \cdot p^n$. Dabei ist p die Basis. Wir werden immer $p = 10$ verwenden, aber Computer rechnen auch mit Zahlen zur Basis $p = 2$. Weiters ist $z = \pm 0.z_1 z_2 \dots z_k$ eine Dezimalzahl mit Vorzeichen und k Dezimalstellen (wir arbeiten mit Basis $p = 10$), für die $z_1 \neq 0$ gilt. Man nennt z die Mantisse der Zahl x. Der Exponent n schließlich ist eine positive oder negative ganze Zahl mit m Dezimalstellen. Durch diese Regeln ist die Gleitkommadarstellung einer reellen Zahl eindeutig bestimmt.

Nehmen wir an, ein Computer arbeitet mit Mantissen, die $k = 6$ Stellen haben, und Exponenten, die $m = 2$ Stellen haben. Die Gleikommazahlen mit 6-stelligen Mantissen und 2-stelligen Exponebten nennt man dann Maschinenzahlen. Die Zahl $x = 28.382$ wird geschrieben als $x = 0.283820 \cdot 10^2$. Die Zahl $y = -4802.6362$ wird zu $y = -0.480264 \cdot 10^4$, wobei gerundet wird, da die Mantissen nur 6 Stellen haben. Die Zahl $a = 0.00065927$ wird zu $a = 0.659270 \cdot 10^{-3}$. Die größte Zahl, die dieser Computer speichern kann, ist $0.999999 \cdot 10^{99}$. Die kleinste positive darstellbare Zahl ist $0.100000 \cdot 10^{-99}$.

Wie rechnet dieser Computer? Will man $a = 0.162905 \cdot 10^2$ und $b = -0.472027 \cdot 10^{-1}$ addieren, dann muss man in einer der Mantissen den Dezimalpunkt so verschieben, dass man gleiche Exponenten erhält, zum Beispiel $a = 162.905 \cdot 10^{-1}$. Dann berechnet man $a+b = 162.432973 \cdot 10^{-1}$ und rundet auf 6 Stellen zu $162.433 \cdot 10^{-1}$, sodass man als Ergebnis $a+b = 0.162433 \cdot 10^2$ erhält. Die Multiplikation ist einfacher. Man multipliziert die beiden Mantissen und addiert die beiden Exponenten und erhält $-0.076895558435 \cdot 10^{-1}$. Durch Runden hat man das Ergebnis $a \cdot b = -0.768956 \cdot 10^{-2}$. Bei Division werden die Mantissen dividiert und die Exponenten subtrahiert. Für b/a erhält man $-2.89755992756 \dots \cdot 10^{-3}$ und durch Runden $b/a = -0.289756 \cdot 10^{-2}$.

Bei Addition und Subtraktion gibt es ein Problem, die sogenannte Auslöschung. Subtrahiert man zwei annähernd gleiche Zahlen (oder addiert man annähernd betragsgleiche Zahlen mit entgegengesetztem Vorzeichen), dann werden die führenden Dezimalstellen der Mantisse null und es bleiben nicht mehr genug Dezimalstellen übrig, um alle Stellen der Mantisse im Ergebnis zu füllen. Es werden Dezimalstellen ausgelöscht. Subtrahiert unser Computer mit 6-stelligen Mantissen die Zahlen $a = 0.375483 \cdot 10^2$ und $b = 0.374802 \cdot 10^2$ voneinander, so ergibt sich $a - b = 0.000681 \cdot 10^2 = 0.681 \cdot 10^{-1}$. Die Mantisse ist nur mehr dreistellig. Man kann drei weitere Dezimalstellen anfügen, aber die sind bedeutungslos. Auch wenn die Mantissen der Zahlen a und b auf sechs Stellen genau angegeben werden, die Mantisse von $a - b$ hat nur mehr drei richtige Dezimalstellen, die anderen sind irrelevant.

Um solche Probleme zu untersuchen, führt man den relativen Fehler einer Zahl ein. Hat man statt eines exakten Wertes x nur einen Näherunswert $\hat{x}$, dann nennt man $\hat{x} - x$ den absoluten und $\varepsilon_x = \frac{\hat{x}-x}{x}$ den relativen Fehler der Zahl x. Diese Definition ist äquivalent zu $\hat{x} = x(1 + \varepsilon_x)$.

Der relative Fehler ε_x einer Zahl x gibt die Anzahl der Stellen an, in denen die Mantisse y von x mit der Mantisse z des Näherungswerts $\hat{x}$ übereinstimmt. Sie stimmen in j Stellen überein, wenn $10^{-j-1} \leq |z - y| \leq 10^{-j}$ gilt, das heißt $\frac{10^{-j-1}}{|y|} \leq |\frac{\hat{x}-x}{x}| \leq \frac{10^{-j}}{|y|}$. Wegen $\frac{1}{10} \leq |y| < 1$ bedeutet das $10^{-j-1} \leq |\varepsilon_x| \leq 10^{-j+1}$. Stimmen die Mantissen von x und $\hat{x}$ in den ersten j Stellen überein, dann gilt $\varepsilon_x \approx 10^{-j}$.

Hier einige Beispiele. Ist $x = 37.54913$ und $\hat{x} = 37.540$, dann gilt $\varepsilon_x = -0.243 \cdot 10^{-4}$. Die Zahlen x und $\hat{x}$ stimmen in den ersten 4 Stellen überein. Ist $x = 0.3999632$ und $\hat{x} = 0.4000370$, dann gilt $\varepsilon_x = 0.185 \cdot 10^{-4}$. Auch hier stimmen die ersten 4 Stellen überein, da bei Dezimalzahlen ja $0.3999\ldots = 0.4000\ldots$ gilt.

2. Fehlerfortpflanzung

Wir kommen zur Addition zurück. Sind ε_x und ε_y die relativen Fehler zweier Zahlen, dann will man wissen, wie sich diese relativen Fehler bei Addition und Subtraktion fortpflanzen.

Satz 1: *Es liegen Näherungswerte $\hat{x}$ und $\hat{y}$ zweier Zahlen mit relativen Fehlern ε_x und ε_y vor. Näherungswerte für Summe und Differenz sind $\hat{x} + \hat{y}$ und $\hat{x} - \hat{y}$. Für die relativen Fehler hat man dann $\varepsilon_{x+y} = \frac{x}{x+y}\varepsilon_x + \frac{y}{x+y}\varepsilon_y$ und $\varepsilon_{x-y} = \frac{x}{x-y}\varepsilon_x - \frac{y}{x-y}\varepsilon_y$.*

Beweis: Nach Definition des relativen Fehlers gilt $\hat{x} = x(1+\varepsilon_x)$ und $\hat{y} = y(1+\varepsilon_y)$. Daraus folgt $\varepsilon_{x+y} = \frac{\hat{x}+\hat{y}-x-y}{x+y} = \frac{x\varepsilon_x+y\varepsilon_y}{x+y} = \frac{x}{x+y}\varepsilon_x + \frac{y}{x+y}\varepsilon_y$.
Ebenso folgt $\varepsilon_{x-y} = \frac{\hat{x}-\hat{y}-(x-y)}{x+y} = \frac{x\varepsilon_x-y\varepsilon_y}{x-y} = \frac{x}{x-y}\varepsilon_x - \frac{y}{x-y}\varepsilon_y$. □

Damit wissen wir, wie sich die relativen Fehler bei Addition und Subtraktion fortpflanzen. Die Faktoren $\frac{x}{x+y}$ und $\frac{y}{x+y}$ geben an, wie sehr sich die relativen Fehler bei Addition verstärken. Sie heißen Verstärkungsfaktoren oder Konditionszahlen. Die Verstärkungsfaktoren bei Subtraktion sind $\frac{x}{x-y}$ und $\frac{y}{x-y}$.

Große relative Fehler sind zu erwarten, wenn die Verstärkungsfaktoren groß sind. Das ist der Fall, wenn zwei annähernd gleich große Zahlen subtrahiert werden. In obigem Beispiel haben wir die Zahlen $a = 0.375483 \cdot 10^2$ und $b = 0.374802 \cdot 10^2$ subtrahiert und festgestellt, dass dabei drei Stellen der Mantisse verloren gehen. In diesem Fall sind die Verstärkungsfaktoren $\frac{a}{a-b} = 0.551 \cdot 10^3$ und $\frac{b}{a-b} = 0.550 \cdot 10^3$, also von der Größenordnung 10^3. Die relativen Fehler werden mit 10^3 multipliziert, woran man erkennt, dass man bei Subtraktion der beiden Zahlen den Verlust von drei Stellen in der Mantisse zu erwarten hat.

Die Addition von zwei Zahlen mit gleichem Vorzeichen und die Subtraktion von zwei Zahlen mit verschiedenen Vorzeichen ist auf jeden Fall unproblematisch, da in diesen Fällen die Verstärkungsfaktoren Betrag < 1 haben und somit die relativen Fehler nicht verstärkt werden.

Jetzt untersuchen wir Multiplikation und Division.

Satz 2: *Es liegen Näherungswerte $\hat{x}$ und $\hat{y}$ zweier Zahlen mit relativen Fehlern ε_x und ε_y vor. Man berechnet Produkt und Quotient durch $\hat{x} \cdot \hat{y}$ und $\hat{x}/\hat{y}$. Für die relativen Fehler hat man dann $\varepsilon_{x \cdot y} = \varepsilon_x + \varepsilon_y + \varepsilon_x \varepsilon_y \approx \varepsilon_x + \varepsilon_y$ und $\varepsilon_{x/y} = \frac{\varepsilon_x - \varepsilon_y}{1+\varepsilon_y} \approx \varepsilon_x - \varepsilon_y$.*

Beweis: Nach Definition des relativen Fehlers gilt $\hat{x} = x(1+\varepsilon_x)$ und $\hat{y} = y(1+\varepsilon_y)$. Daraus folgt $\varepsilon_{x \cdot y} = \frac{\hat{x} \cdot \hat{y} - x \cdot y}{x \cdot y} = \frac{x(1+\varepsilon_x) y(1+\varepsilon_y) - x \cdot y}{x \cdot y} = (1+\varepsilon_x)(1+\varepsilon_y) - 1 = \varepsilon_x + \varepsilon_y + \varepsilon_x \varepsilon_y$.
Ebenso folgt $\varepsilon_{x/y} = \frac{\hat{x}/\hat{y} - x/y}{x/y} = (\frac{x(1+\varepsilon_x)}{y(1+\varepsilon_y)} - \frac{x}{y}) \frac{y}{x} = \frac{1+\varepsilon_x}{1+\varepsilon_y} - 1 = \frac{\varepsilon_x - \varepsilon_y}{1+\varepsilon_y}$.
Da relative Fehler klein sind, kann man $\varepsilon_x \varepsilon_y \approx 0$ und $1 + \varepsilon_y \approx 1$ annehmen. □

Aus diesem Satz sieht man, dass bei Multiplikation und Division keine Verstärkung des relativen Fehlers auftritt.

Genauso kann man Funktionsauswertungen untersuchen, zum Beispiel Wurzelziehen oder Logarithmieren.

Satz 3: *Sei f eine stetig differenzierbare Funktion. Es liegt ein Näherungswert $\hat{x}$ einer Zahl x mit relativen Fehlern ε_x vor. Man berechnet den Funktionswert dieser Zahl durch $f(\hat{x})$. Für den relativen Fehler hat man dann $\varepsilon_{f(x)} \approx \frac{x f'(x)}{f(x)} \varepsilon_x$.*

Beweis: Wir erhalten $f(\hat{x}) - f(x) \approx f'(x)(\hat{x} - x) = x f'(x) \frac{\hat{x} - x}{x} = x f'(x) \varepsilon_x$ mit Hilfe des Mittelwertsatzes. Es folgt $\varepsilon_{f(x)} = \frac{f(\hat{x}) - f(x)}{f(x)} \approx \frac{x f'(x)}{f(x)} \varepsilon_x$. □

Dieser Satz besagt, dass der Verstärkungsfaktor für den relativem Fehler bei Auswertung einer Funktion f im Punkt x gleich $\frac{x f'(x)}{f(x)}$ ist. Ist der Verstärkungsfaktor klein, dann spricht man von einem gut konditionierten Problem. Ansonsten spricht man von schlechter Kondition. Wir untersuchen einige Funktionen.

Sei $\alpha \in \mathbb{R}$ und $f(x) = x^\alpha$. Dann gilt $\frac{x f'(x)}{f(x)} = \frac{x \alpha x^{\alpha-1}}{x^\alpha} = \alpha$. Ist der Exponent α nicht allzu groß, dann wird der relative Fehler kaum verstärkt. Beim Wurzelziehen ($\alpha = \frac{1}{2}$) wird der relative Fehler halbiert.

Sei $a > 1$ und $f(x) = a^x$. Dann gilt $\frac{x f'(x)}{f(x)} = \frac{x a^x \log a}{a^x} = x \log a$. Für $x \in [0,1]$ ist das $\leq \log a$. Der relative Fehler wird kaum verstärkt. Ist x allerdings groß, dann kann der relative Fehler entsprechend größer werden.

Für $f(x) = \sin x$ gilt $\frac{x f'(x)}{f(x)} = \frac{x \cos x}{\sin x}$. Für $x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ hat man $|\frac{x \cos x}{\sin x}| \leq 1$. Der relative Fehler wird nicht größer. Man hat gute Kondition. Liegt x in der Nähe von π, dann ist der Verstärkungsfaktor $\frac{x \cos x}{\sin x}$ jedoch groß.

Für $f(x) = \cos x$ gilt $\frac{x f'(x)}{f(x)} = -\frac{x \sin x}{\cos x}$. Für $x \in [-\frac{\pi}{4}, \frac{\pi}{4}]$ hat man $|\frac{x \sin x}{\cos x}| \leq \frac{\pi}{4} < 1$. Die Kondition ist gut. Liegt x in der Nähe von $\frac{\pi}{2}$, dann ist der Verstärkungsfaktor jedoch groß. Die Kondition ist schlecht.

Für $f(x) = \log(x)$ gilt $\frac{x f'(x)}{f(x)} = \frac{1}{\log(x)}$. Liegt x in der Nähe von 1, dann ist der Verstärkungsfaktor groß. In diesem Fall ist die Funktionsauswertung des Logarithmus schlecht konditioniert.

Diese Überlegungen haben praktische Konsequenzen. Es kann vorkommen, dass man durch einen schlecht gewählten Algorithmus für ein Problem, das eine gute Kondition aufweist, schlechte Ergebnisse erhält. Man sollte darauf achten, dass bei der Ausführung des Algorithmus, wenn möglich, Auslöschung vermieden wird. Wir überlegen uns das an einem Beispiel.

Beispiel: Sei $f(x) = \sqrt{x+1} - \sqrt{x}$. Diese Funktion soll für $x = 75$ bei Rechnung mit vierstelligen Mantissen ausgewertet werden.

Wir berechnen $\frac{xf'(x)}{f(x)} = -\frac{\sqrt{x}}{2\sqrt{x+1}}$. Der Verstärkungsfaktor ist somit $< \frac{1}{2}$. Das Problem ist gut konditioniert.

Ein Algorithmus zur Berechnung von $f(x)$ besteht darin, zuerst $s_1 = \sqrt{x+1}$ und $s_2 = \sqrt{x}$ zu berechnen und dann $s_1 - s_2$. Tun wir das für $x = 75.00$, so erhalten wir $\sqrt{x+1} = 8.71779789$ und $\sqrt{x} = 8.660254038$ und somit $s_1 = 8.718$ und $s_2 = 8.660$, da wir ja nur vierstellige Zahlen abspeichern können. Die Berechnung von $s_1 - s_2$ ergibt 0.058, das heißt $0.58?? \cdot 10^{-1}$ in Gleitkommadarstellung. Wir sehen, dass bei dieser Vorgangsweise zwei Stellen durch Auslöschung verlorengehen.

Nun gilt auch $f(x) = \frac{1}{\sqrt{x+1}+\sqrt{x}}$. Wir können daher die Funktionsauswertung auch so durchführen, dass wir s_1 und s_2 wie oben berechnen, dann aber $s_3 = s_1 + s_2$ und schließlich $\frac{1}{s_3}$. Mit den bereits oben berechneten Werten für s_1 und s_2 erhalten wir $s_1 + s_2 = 17.378$ und somit $s_3 = 17.38$, da wir auf vier Stellen runden müssen. Schließlich erhalten wir $\frac{1}{s_3} = 0.057537399$. Auf vier Stellen gerundet ist das $0.5754 \cdot 10^{-1}$. Das ist das genauere Resultat, da bei dieser Berechnung keine Auslöschung aufgetreten ist. (Der genaue Wert ist 0.0575438492.)

Ähnliches wie in obigem Beispiel geschieht bei der Lösung einer quadratischen Gleichung $ax^2 + bx + c = 0$, wobei $a \neq 0$ und $b^2 - 4ac > 0$ gelte. Die Lösungen berechnet man durch $x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a}$ und $x_2 = \frac{-b-\sqrt{b^2-4ac}}{2a}$. Wenn der Betrag von $4ac$ wesentlich kleiner als b^2 ist, dann liegt $\sqrt{b^2-4ac}$ nahe bei $|b|$. Für $b > 0$ hat man bei der Berechnung von x_1 Auslöschung, für $b < 0$ bei der Berechnung von x_2. Um das zu vermeiden, kann man für $b > 0$ zuerst x_2 durch obige Formel berechnen (keine Auslöschung wegen $b > 0$) und x_1 dann durch $x_1 = \frac{c}{ax_2}$. Für $b < 0$ berechnet man zuerst x_1 durch obige Formel (keine Auslöschung wegen $b < 0$) und x_2 dann durch $x_2 = \frac{c}{ax_1}$.

II. Polynome

Wir behandeln ein einfaches Verfahren für das Rechnen mit Polynomen. Man kann damit ein Polynom und dessen Ableitungen an einer Stelle u auswerten, ein Polynom durch ein anderes dividieren, oder auch ein Polynom in Potenzen von $x - u$ schreiben, wobei u eine feste Zahl ist.

1. Der Horneralgorithmus

Die im folgende Satz angegebene Methode zur Division eines Polynoms $P(x)$ durch ein lineares Polynom $x - u$ heißt Horneralgorithmus.

Satz 4: *Sei $P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_n$ ein Polynom mit $n \geq 1$ und $a_n \neq 0$, sodass n der Grad von $P(x)$ ist. Die Koeffizienten $a_0, a_1, \ldots, a_n$ liegen entweder in $\mathbb{R}$ oder in $\mathbb{C}$. Weiters sei $u \in \mathbb{R}$ oder $u \in \mathbb{C}$. Wir berechnen $b_0 = a_0$ und $b_k = a_k + b_{k-1}u$ für $1 \leq k \leq n$. Wir bilden das Polynom $Q(x) = b_0x^{n-1} + b_1x^{n-2} + \cdots + b_{n-1}$. Dann gilt $P(x) = (x - u)Q(x) + b_n$.*

Beweis: Wir setzen $Q(x) = b_0x^{n-1} + b_1x^{n-2} + \cdots + b_{n-1}$ in $(x - u)Q(x) + b_n$ ein und formen um. Wir erhalten $b_0x^n + (b_1 - b_0u)x^{n-1} + (b_2 - b_1u)x^{n-2} + \cdots + (b_n - b_{n-1}u)$. Weiters gilt $b_0 = a_0$ und $b_k - b_{k-1}u = a_k$ für $1 \leq k \leq n$. Setzt man das ein, so ergibt sich $(x - u)Q(x) + b_n = P(x)$. □

Ist ein Polynom $P(x)$ und eine Zahl u vorgegeben, dann berechnet man die Zahlen $b_0, b_1, \ldots, b_n$ nach der in Satz 4 angegebenen Rekursionsformel und bildet daraus das Polynom $Q(x)$. Dann gilt $P(x) = (x - u)Q(x) + b_n$, das heißt $Q(x)$ ist das Polynom, das man bei Division von $P(x)$ durch $x - u$ erhält und b_n ist der Rest. Setzt man $x = u$ in die Formel $P(x) = (x - u)Q(x) + b_n$ ein, dann sieht man, dass $P(u) = b_n$ gilt. Somit hat man auch P an der Stelle u berechnet.

Die Rechnung führt man in einer Tabelle durch, wie links unten dargestellt. In die erste Zeile schreibt man die Koeffizienten des Polynoms. Die erste Stelle der zweiten Zeile bleibt leer. Man berechnet eine Spalte nach der anderen, indem die Zahlen der ersten beiden Zeilen addiert und in die dritte Zeile einträgt. Dieses Ergebnis multipliziert man mit u und schreibt es in die zweite Zeile der nächsten Spalte. So tut man weiter bis ans Ende der Tabelle. Links ist das allgemeine Rechenschema dargestellt. Man sieht, dass mit dieser Rechenmethode die Rekursionsformeln aus Satz 4 ausgeführt werden. In der Tabelle rechts wird das Verfahren für das Polynom $P(x) = 4x^4 - 5x^3 - 8x^2 + 6$ und $u = 2$ durchgeführt.

	a_0	a_1	a_2	a_3	a_4
u		b_0u	b_1u	b_2u	b_3u
	b_0	b_1	b_2	b_3	b_4

	4	−5	−8	0	6
$u = 2$		8	6	−4	−8
	4	3	−2	−4	−2

Es ergibt sich das Polynom $Q(x) = 4x^3 + 3x^2 - 2x - 4$ und $b_4 = -2$. Somit erhalten wir $P(x) = (x - 2)Q(x) - 2$ und $P(2) = -2$.

Durch wiederholtes Anwenden des Horneralgorithmus können wir auch die Ableitungen eines Polynoms an einer Stelle u berechnen.

Satz 5: *Sei $P(x)$ ein Polynom vom Grad n und u sei vorgegeben. Durch wiederholtes Anwenden des Horneralgorithmus finden wir Polynome $Q_k(x)$ und Zahlen c_k, sodass*

$P(x) = (x-u)Q_1(x) + c_0$ und $Q_{k-1}(x) = (x-u)Q_k(x) + c_{k-1}$ für $2 \le k \le n$ gilt. Dann gilt auch $P^{(j)}(u) = j!c_j$ für $0 \le j \le n$.

Beweis: Es gilt $P(x) = (x-u)Q_1(x) + c_0$. Durch Differenzieren erhalten wir $P'(x) = (x-u)Q_1'(x) + Q_1(x)$. Nochmaliges Differenzieren ergibt $P''(x) = (x-u)Q_1''(x) + 2Q_1'(x)$. Es ist leicht zu sehen, wie es weitergeht. Nach k-maligem Differenzieren hat man $P^{(k)}(x) = (x-u)Q_1^{(k)}(x) + kQ_1^{(k-1)}(x)$. Setzt man $x = u$, so ergibt sich $P^{(k)}(u) = kQ_1^{(k-1)}(u)$.
Wir beweisen $P^{(j)}(u) = j!c_j$ mit Induktion nach j. Durch Einsetzen von $x = u$ in $P(x) = (x-u)Q_1(x) + c_0$ erhalten wir $P(u) = c_0$, sodass die Formel für $j = 0$ gezeigt ist. Nehmen wir also an, dass $P^{(j-1)}(u) = (j-1)!c_{j-1}$ bereits bewiesen ist. Wir können dieses Resultat auf jedes Polynom anwenden, insbesondere auch auf $Q_1(x)$. Anstelle der Zahlen $c_0, c_1, c_2, \ldots$, die für $P(x)$ auftreten, haben wir dann die Zahlen $c_1, c_2, c_3, \ldots$, wenn wir mit $Q_1(x)$ anstelle von $P(x)$ beginnen. Somit gilt $Q_1^{(j-1)}(u) = (j-1)!c_j$. Aus der oben bewiesenen Formel $P^{(j)}(u) = jQ_1^{(j-1)}(u)$ ergibt sich dann $P^{(j)}(u) = j!c_j$. Damit ist der Induktionsbeweis gelungen. □

Um die Ableitungen eines Polynoms mit der Methode aus Satz 5 zu berechnen, schreibt man die Tabellen zur Durchführung des Horneralgorithmus direkt untereinander. Die dritte Zeile jeder dieser Tabellen, abgesehen von der letzten Stelle, enthält ja die Koeffizienten des Polynoms, mit denen man die nächste Tabelle beginnt. Wir führen das für das Beispiel $P(x) = 2x^4 + x^3 - 3x^2 - 2x + 1$ und $u = -1$ durch.

	2	1	−3	−2	1
$u = -1$		−2	1	2	0
	2	−1	−2	0	1
$u = -1$		−2	3	−1	
	2	−3	1	−1	
$u = -1$		−2	5		
	2	−5	6		
$u = -1$		−2			
	2	−7			
$u = -1$					
	2				

Die Zahlen c_0, c_1, c_2, c_3, c_4 treten als die letzten Zahlen in den Zeilen unter den Strichen auf. Wir erhalten somit $P(-1) = 1$, $P'(-1) = -1$, $\frac{1}{2!}P''(-1) = 6$, $\frac{1}{3!}P^{(3)}(-1) = -7$ und $\frac{1}{4!}P^{(4)}(-1) = 2$. Damit sind alle Ableitungen des Polynoms an der Stelle -1 berechnet. Alle weiteren Ableitungen sind ja gleich null, da $P(x)$ Grad 4 hat.

Man kann dieses Resultat auch verwenden, um $P(x)$ in Potenzen von $x-u$ zu schreiben. Es gilt ja $P(x) = d_n(x-u)^n + d_{n-1}(x-u)^{n-1} + \cdots + d_1(x-u) + d_0$ mit $d_j = \frac{P^{(j)}(u)}{j!}$ für ein Polynom $P(x)$ vom Grad n. Das sieht man, indem man diese Gleichung j Mal differenziert und dann $x = u$ setzt (Taylorformel). Die Koeffizienten d_j stimmen also mit den oben berechneten Zahlen c_j überein. Für das Polynom $P(x) = 2x^4 + x^3 - 3x^2 - 2x + 1$ gilt somit $P(x) = 2(x+1)^4 - 7(x+1)^3 + 6(x+1)^2 - (x+1) + 1$.

2. Division durch Polynome höheren Grades

Man kann den Horneralgorithmus erweitern, sodass man damit auch durch Polynome höheren Grades dividieren kann. Wir tun das nur für Polynome zweiten Grades.

Satz 6: *Sei* $P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_n$ *ein Polynom mit Grad* $n \geq 1$ *und* $a_n \neq 0$. *Seien* u *und* v *vorgegebene reelle Zahlen. Wir berechnen der Reihe nach* $b_0 = a_0$, $b_1 = a_1 + b_0u$, $b_k = a_k + b_{k-1}u + b_{k-2}v$ *für* $2 \leq k \leq n-1$ *und schließlich* $b_n = a_n + b_{n-2}v$. *Wir bilden die Polynome* $Q(x) = b_0x^{n-2} + b_1x^{n-3} + \cdots + b_{n-2}$ *und* $R(x) = b_{n-1}x + b_n$. *Dann gilt* $P(x) = (x^2 - ux - v)Q(x) + R(x)$.

Beweis: Wir setzen $Q(x) = b_0x^{n-2} + b_1x^{n-3} + \cdots + b_{n-2}$ und $R(x) = b_{n-1}x + b_n$ in $(x^2 - ux - v)Q(x) + R(x)$ ein, multiplizieren aus und fassen zusammen. Wir erhalten $b_0x^n + (b_1 - b_0u)x^{n-1} + \sum_{k=2}^{n-1}(b_k - b_{k-1}u - b_{k-2}v)x^{n-k} + (b_n - b_{n-2}v)$. Mit Hilfe der im Satz angegebenen Rekursionsformeln ergibt sich $(x^2 - ux - v)Q(x) + R(x) = P(x)$. □

Die Rekursinsformeln aus Satz 6 schreibt man wieder in Tabellenform auf. In die erste Zeile schreibt man die Koeffizienten des Polynoms. Die ersten beiden Stellen der zweiten Zeile und die erste und letzte Stelle der dritten Zeile bleiben leer. Man berechnet eine Spalte nach der anderen, indem die Zahlen der ersten drei Zeilen addiert und in die vierte Zeile einträgt. Dieses Ergebnis multipliziert man mit u und schreibt es in die dritte Zeile der nächsten Spalte; man multiplizert es mit v und schreibt es in die zweite Zeile der übernächsten Spalte. So tut man weiter bis ans Ende der Tabelle. Links ist das allgemeine Rechenschema dargestellt. Man erkennt, dass mit dieser Rechenmethode die Rekursionsformeln aus Satz 6 ausgeführt werden. In der rechten Tabelle wird das Verfahren für das Polynom $P(x) = 2x^4 + x^3 - 3x^2 - 2x + 1$ und $u = -1$, $v = 2$ durchgeführt.

	a_0	a_1	a_2	a_3	a_4
v			b_0v	b_1v	b_2v
u		b_0u	b_1u	b_2u	
	b_0	b_1	b_2	b_3	b_4

	2	1	−3	−2	1
$v = 2$			4	−2	4
$u = -1$		−2	1	−2	
	2	−1	2	−6	5

Es ergeben sich die Polynome $Q(x) = 2x^2 - x + 2$ und $R(x) = -6x + 5$. Somit erhalten wir $P(x) = (x^2 + x - 2)Q(x) + R(x)$.

III. Nullstellen

Sei $I \subseteq \mathbb{R}$ ein Intervall und $f : I \to \mathbb{R}$ eine Abbildung. Ein Punkt $x \in I$ heißt Fixpunkt von f, wenn $f(x) = x$ gilt. Ein Punkt $x \in I$ heißt Nullstelle von f, wenn $f(x) = 0$ gilt. Wir suchen Iterationsverfahren zur Bestimmung von Nullstellen und Fixpunkten. Dazu gibt man eine Startwert $x_0 \in I$ vor und definiert rekursiv eine Folge $x_0, x_1, x_2, \ldots$ in I, die gegen eine Nullstelle oder einen Fixpunkt von f konvergiert.

1. Fixpunkte

Sei $I \subseteq \mathbb{R}$ ein Intervall. Eine Funktion $g : I \to I$ heißt kontrahierend, wenn eine Konstante $q < 1$ existiert mit $|g(x) - g(y)| \leq q|x - y|$ für alle x und y in I. Die Konstante q heißt Kontraktionskonstante. Für kontrahierende Abbildungen ist es einfach, Fixpunkte durch ein Iterationsverfahren zu bestimmen, wie der folgende Satz zeigt.

Satz 7 (Kontraktionssatz) *Sei $I \subseteq \mathbb{R}$ ein abgeschlossenes Intervall und $g : I \to I$ eine kontrahierende Abbildung mit Kontraktionskonstante q. Dann hat g genau einen Fixpunkt u in I. Für alle $x \in I$ gilt $\lim_{n\to\infty} g^n(x) = u$.*

Beweis: Wir zeigen, dass ein Punkt $c \in I$ existiert mit $g(c) \geq c$. Ist I nach unten beschränkt, dann liegt der linke Endpunkt a von I in I. Es gilt dann $g(a) \in I$, also $g(a) \geq a$. Wir können $c = a$ wählen. Ist I nicht nach unten unbeschränkt, dann sei $y \in I$ beliebig. Wir wählen $c < \min(y, \frac{g(y)-qy}{1-q})$. Wegen $c < y$ gilt dann $c \in I$. Wegen $g(y) - g(c) \leq |g(y) - g(c)| \leq q(y - c)$ haben wir auch $g(c) \geq g(y) - qy + qc > c$. Somit hat c die gewünschte Eigenschaft. Ganz analog findet man ein $d \in I$ mit $g(d) \leq d$.

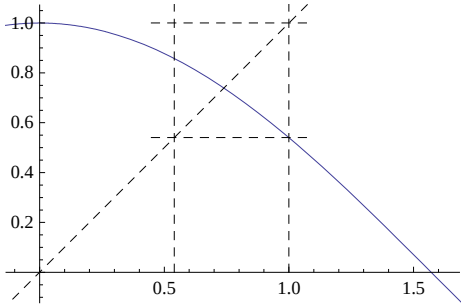
Sei $h(x) = g(x) - x$. Dann gilt $h(c) \geq 0$ und $h(d) \leq 0$. Nach dem Zwischenwertsatz existiert ein u im Intervall mit Endpunkten c und d, also auch $u \in I$, sodass $h(u) = 0$ gilt. Daraus folgt $g(u) = u$. Damit ist die Existenz des Fixpunkts gezeigt.

Sei v ebenfalls ein Fixpunkt von g. Dann gilt $|u - v| = |g(u) - g(v)| \leq q|u - v|$. Weil aber $q < 1$ gilt, folgt $|u - v| = 0$, also $u = v$. Der Fixpunkt ist eindeutig.

Sei $x \in I$ beliebig. Dann gilt $|g^n(x) - u| = |g^n(x) - g(u)| \leq q|g^{n-1}(x) - u|$. Wiederholt man diese Abschätzung, so erhält man schließlich $|g^n(x) - u| \leq q^n|x - u|$. Wegen $q < 1$ ist damit $\lim_{n\to\infty} g^n(x) = u$ gezeigt. □

Bemerkung: Wenn g differenzierbar ist, dann kann man $q = \sup_{t\in I} |g'(t)|$ als Kontraktionskonstante wählen. Sind x und y in I, dann existiert nach dem Mittelwertsatz ein ξ zwischen x und y, sodass $|g(x) - g(y)| = |g'(\xi)(x - y)| \leq q|x - y|$ gilt.

Beispiel: Gibt man eine Zahl in den Taschenrechner ein und drückt dann wiederholt die Cosinustaste, dann erhält man eine Folge von Zahlen, die gegen $0.739085\ldots$ konvergiert. Man hat einen Fixpunkt für die Cosinusfunktion gefunden. Die Konvergenz ergibt sich wegen Satz 7. Sei $g : \mathbb{R} \to \mathbb{R}$ definiert durch $g(x) = \cos x$. Sei $I = [a, b]$ mit $b = 1$ und $a = \cos 1 = 0.5403$. Dann gilt $g(I) \subseteq I$ wegen $g(b) = a$ und $g(a) < b$ und da g auf I monoton fallend ist. Weiters gilt $\sup_{t\in I} |g'(t)| = \sup_{t\in I} \sin t = \sin 1$. Somit ist $q = \sin 1 = 0.8415 < 1$



eine Kontraktionskonstante für die Funktion $g : I \to I$. Wir können Satz 7 anwenden. Die Abbildung $g : I \to I$ hat einen eindeutig bestimmten Fixpunkt $u \in I$. Für einen beliebigen Startwert $x_0 \in I$ konvergiert die Folge $(x_n)_{n\geq 0}$ definiert durch $x_n = g^n(x_0)$ gegen den eindeutigen Fixpunkt u. Wir berechnen diese Folge iterativ. Wir können zum Beispiel mit $x_0 = 1$ beginnen und berechnen $x_1 = g(x_0) = 0.5403$, $x_2 = g(x_1) = 0.8576$, $x_3 = g(x_2) = 0.6543$, $x_4 = g(x_3) = 0.7935$, $x_5 = g(x_4) = 0.7014$, und so weiter. Nach ungefähr 25 Iterationen erreicht man den Wert 0.7391, der sich bei vierstelliger Rechnung nicht mehr ändert. Damit hat man den Fixpunkt gefunden.

Beispiel: Wir suchen die größte Nullstelle der Funktion $f(x) = x^3 - ax - b$, wobei $a > 0$ und $b > 0$ vorgegebene Konstanten sind. Wegen $f(0) = -b$ liegt diese Nullstelle in $\mathbb{R}^+$. Es gibt viele Möglichkeiten, die Gleichung $f(x) = 0$ so umzuformen, dass die Nullstelle zu einem Fixpunkt einer Funktion wird, zum Beispiel $x = \frac{x^3-b}{a}$, $x = \frac{ax+b}{x^2}$, $x = \sqrt{a + b/x}$ oder $x = \sqrt[3]{ax + b}$. Die Nullstelle ist ein Fixpunkt dieser vier rechts vom Gleichheitszeichen stehenden Funktionen. Um Satz 7 anwenden zu können, müssen wir ein Intervall I finden, das den Fixpunkt enthält und auf dem die Funktion kontrahierend ist. Mit manchen dieser Funktionen funktioniert es gar nicht, mit anderen schlecht. Am besten ist die Funktion $g(x) = \sqrt[3]{ax + b}$ geeignet. Es gilt $g'(x) = \frac{a}{3}(ax + b)^{-\frac{2}{3}}$. Wir wählen $I = [\sqrt{a}, \infty)$. Für $x \in I$ gilt $g(x) \geq \sqrt[3]{a\sqrt{a} + b} \geq \sqrt{a}$, das heißt $g(x) \in I$, womit $g(I) \subseteq I$ gezeigt ist. Für $x \in I$ gilt auch $|g'(x)| \leq \frac{a}{3}(a\sqrt{a} + b)^{-\frac{2}{3}} \leq \frac{1}{3}$, sodass g auf I kontrahierend ist mit Kontrktionskonstante $\frac{1}{3}$. Wählt man $x_0 \in I$ beliebig und berechnet $x_n = g(x_{n-1})$ für $n \geq 1$, dann wird diese Folge gegen den eindeutigen Fixpunkt von g in I konvergieren, das ist die größte Nullstelle der Funktion $f(x) = x^3 - ax - b$.
Sei $a = 1$ und $b = 4$. Als Startwert wählen wir $x_0 = 5$. Eine Berechnung der Folge ergibt $x_1 = 2.08008$, $x_2 = 1.82517$, $x_3 = 1.79930$, $x_4 = 1.79663$, $x_5 = 1.79635$, $x_6 = 1.79633$, $x_7 = 1.79632$, $x_8 = 1.79632$. Nach acht Iterationsschritten ändert sich der Wert nicht mehr. Damit ist die größte Nullstelle von $x^3 - x - 4$ auf 5 Dezimalstellen genau gefunden.

Bemerkung: Je kleiner die Kontraktionskonstante ist, um so schneller ist die Konvergenz. Im verletzten Beispiel ist $q = \sin 1 = 0.8415$ eine Kontraktionskonstante. Sie liegt nahe bei 1. Es waren auch 25 Iterationen erforderlich, um den Fixpunkt auf 4 Dezimalstellen genau zu berechnen. Im letzten Beispiel ist $q = \frac{1}{3}$ eine Kontraktionskonstante. Sie ist wesentlich kleiner. Es waren auch nur 8 Iterationen erforderlich, um den Fixpunkt auf 5 Dezimalstellen genau zu berechnen.

2. Das Newtonverfahren zur Bestimmung von Nullstellen

Sei $f : \mathbb{R} \to \mathbb{R}$ eine differenzierbare Funktion. Wir suchen eine Nullstelle dieser Funktion. Angenommen, es liegt ein Näherungswert t dieser Nullstelle vor. Um einen besseren Näherungswert zu erhalten, legen wir die Tangente im Punkt $(t, f(t))$ an die Funktion und berechnen die Nullstelle dieser Tangente. Wir können hoffen, dadurch einen besseren Näherungswert zu erhalten, da die Tangente die Funktion approximiert. Die Gleichung der Tangente ist $x \mapsto f(t) + f'(t)(x - t)$. Ihre Nullstelle ist $t - \frac{f(t)}{f'(t)}$. Aus dem Näherungswert t haben wir einen neuen Näherungswert $t - \frac{f(t)}{f'(t)}$ berechnet. Wir definieren $g(t) = t - \frac{f(t)}{f'(t)}$. Liegt ein Startwert x_0 vor, so kann man die Folge $x_n = g(x_{n-1})$ für $n \geq 1$ definieren. Bei geeignetem Startwert kann man so immer bessere Approximationen der Nullstelle gewinnen.

Satz 8 (Newtonverfahren) *Sei $f : \mathbb{R} \to \mathbb{R}$ zweimal stetig differenzierbar und u eine einfache Nullstelle von f, das heißt $f(u) = 0$ und $f'(u) \neq 0$. Sei $g(x) = x - \frac{f(x)}{f'(x)}$. Dann gilt $g(u) = u$ und es existiert eine Umgebung I von u mit $g(I) \subseteq I$, sodass g auf I kontrahierend ist.*

Beweis: Wegen $f(u) = 0$ erhalten wir $g(u) = u$. Somit ist u ein Fixpunkt von f. Weiters gilt $g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}$. Wegen $f(u) = 0$ und $f'(u) \neq 0$ erhalten wir $g'(u) = 0$. Da f zweimal stetig differenzierbar ist, ist auch g' stetig. Es existiert ein $\varepsilon > 0$, sodass $|g'(t)| \leq \frac{1}{2}$ für alle $t \in I := [u - \varepsilon, u + \varepsilon]$ gilt. Sind x und y in I, dann gilt $g(x) - g(y) = g'(\xi)(x - y)$ für ein ξ zwischen x und y, also $\xi \in I$, nach dem Mittelwertsatz. Es folgt $|g(x) - g(y)| \leq \frac{1}{2}|x - y|$. Somit ist g auf I kontrahierend. Insbesondere gilt für jedes $x \in I$ auch $|g(x) - u| = |g(x) - g(u)| \leq \frac{1}{2}|x - u| < |x - u| \leq \varepsilon$, woraus $g(x) \in I$ folgt. Damit ist auch $g(I) \subseteq I$ gezeigt. □

Um eine Nullstelle u einer zweimal stetig differenzierbaren Funktion f zu finden, berechnen wir die Funktion $g(x) = x - \frac{f(x)}{f'(x)}$. Nach Satz 8 ist u ein Fixpunkt von g und es existiert eine Umgebung I von u, auf der g kontrahierend ist. Wählt man $x_0 \in I$ und berechnet $x_n = g(x_{n-1})$ für $n \geq 1$, dann konvergiert die Folge $(x_n)_{n \geq 0}$ nach Satz 7 gegen die Nullstelle u. Das Problem besteht darin, dass man I nicht kennt. Deshalb ist es oft schwer, einen Startwert für das Iterationsverfahren zu finden.

Beispiel: Wir berechnen den Fixpunkt für die Cosinusfunktion, also die Lösung der Gleichung $\cos x = x$ mit Hilfe des Newtonverfahrens. Dieser Fixpunkt ist eine Nullstelle der Funktion $f(x) = x - \cos x$. Es folgt $f'(x) = 1 - \sin x$ und $g(x) = x - \frac{x - \cos x}{1 - \sin x}$. Als Startwert versuchen wir $x_0 = 1$ und probieren, was passiert. Wir berechnen $x_1 = g(x_0) = 0.750364$, $x_2 = g(x_1) = 0.739113$, $x_3 = g(x_2) = 0.739085$ und $x_4 = g(x_3) = 0.739085$. Bei 6-stelliger Rechnung stellt sich bereits nach vier Iterationen ein Wert ein, der sich nicht mehr ändert. Das Newtonverfahren konvergiert sehr schnell.

Bemerkung: Sei $f : \mathbb{R} \to \mathbb{R}$ zweimal stetig differenzierbar und u eine einfache Nullstelle von f. Sei $g(x) = x - \frac{f(x)}{f'(x)}$. Wir berechnen $\lim_{x \to u} \frac{g(x) - u}{(x - u)^2}$. Wegen $g(z) = z$ können wir die Regel von de l Hospital anwenden. Im Beweis von Satz 8 wurde $g'(x) = \frac{f(x)f''(x)}{f'(x)^2}$ berechnet. Wir erhalten $\lim_{x \to u} \frac{g(x) - u}{(x - u)^2} = \lim_{x \to u} \frac{g'(x)}{2(x - u)} = \lim_{x \to u} \frac{f''(x)}{2f'(x)^2} \lim_{x \to u} \frac{f(x)}{x - u}$. Wegen $f'(u) \neq 0$ und $\lim_{x \to u} \frac{f(x)}{x - u} = \lim_{x \to u} \frac{f(x) - f(u)}{x - u} = f'(u)$ ergibt sich

$$\lim_{x \to u} \frac{g(x) - u}{(x - u)^2} = \frac{f''(u)}{2f'(u)}.$$

Sei c eine Konstante mit $|\frac{f''(u)}{2f'(u)}| < c$. Dann existiert ein $\varepsilon > 0$, sodass $|\frac{g(x) - u}{(x - u)^2}| < c$, das heißt $|g(x) - u| < c(x - u)^2$ für alle $x \in (u - \varepsilon, u + \varepsilon)$ gilt. Ist der Abstand von x zur Nullstelle u klein genug, dann verkleinert sich dieser Abstand bei Anwenden der Iterationsfunktion sehr schnell (er wird quadriert). Das erklärt die schnelle Konvergenz des Newtonverfahrens.

Bemerkung: Wir haben das Newtonverfahren nur für einfache Nullstellen untersucht. Es funktioniert auch für mehrfache Nullstellen, die Konvergenz ist allerdings langsamer. Sei u eine m-fache Nullstelle von f, das heißt $f(x) = (x - u)^m h(x)$ mit $h(u) \neq 0$ und h sei stetig differenzierbar. Sei wieder $g(x) = x - \frac{f(x)}{f'(x)}$. Es folgt $\frac{g(x) - u}{x - u} = 1 - \frac{f(x)}{(x - u)f'(x)} = 1 - \frac{(x - u)^m h(x)}{(x - u)^{m+1} h'(x) + m(x - u)^m h(x)}$ und somit $\lim_{x \to u} \frac{g(x) - u}{x - u} = 1 - \frac{1}{m}$. Sei c eine Konstante mit $1 - \frac{1}{m} < c < 1$. Dann gilt $|g(x) - u| \leq c|x - u|$, wenn x nahe genug bei u liegt. Startet

man mit einem x_0, das nahe genug bei u liegt, und berechnet $x_n = g(x_{n-1})$ für $n \geq 1$, dann gilt $|x_n - u| \leq c|x_{n-1} - u|$. Wegen $c < 1$ konvergiert die Folge gegen u. Aber die Konvergenz kann sehr langsam sein.

Beispiel: Das Polynom $P(x) = x^3 + 2x^2 - x - 2$ hat eine einfache Nullstelle bei 1. Wendet man das Newtonverfahren mit Startwert $x_0 = 2$ an, so erhält man $x_1 = 1.36842$, $x_2 = 1.07716$, $x_3 = 1.00452$, $x_4 = 1.0000169$ und $x_5 = 1.0000000002$. Man sieht, dass die Folge sehr schnell gegen die Nullstelle 1 konvergiert. Man kann hier auch beobachten, dass die Abstände des Näherungswertes x_n zur Nullstelle 1 von einem Schritt zum nächsten quadriert werden, sobald x_n nahe genug bei 1 liegt.
Das Polynom $P(x) = x^3 - x^2 - x + 1$ hat eine zweifache Nullstelle bei 1. Wendet man das Newtonverfahren mit Startwert $x_0 = 2$ an, so erhält man $x_5 = 1.04422$, $x_{10} = 1.00141$ und $x_{15} = 1.00004$. Man sieht, dass die Folge gegen die Nullstelle 1 konvergiert, aber langsamer als bei der einfachen Nullstelle.
Je höher die Ordnung der Nullstelle ist, umso langsamer ist die Konvergenz. Das Polynom $P(x) = x^4 - 2x^3 + 2x - 1$ hat eine dreifache Nullstelle bei 1. Wendet man das Newtonverfahren mit Startwert $x_0 = 2$ an, so erhält man $x_{10} = 1.02093$, $x_{15} = 1.00277$, $x_{20} = 1.00036$ und $x_{25} = 1.00005$. Die Konvergenz ist also noch langsamer als bei der zweifachen Nullstelle.

3. Nullstellen von Polynomen

Wir beschäftigen uns noch mit dem Problem einen Startwert für das Newtonverfahren zu finden. Man kann die Funktion f, deren Nullstellen man sucht, in einigen Punkten auswerten und sich so eine Vorstellung verschaffen, wo die Nullstellen ungefähr liegen. So kann man Näherungswerte für die Nullstellen erraten und diese als Startwerte verwenden.

In gewissen Situationen, wie der im folgenden Satz, kann man die Konvergenz des Newtonverfahrens auch beweisen.

Satz 9: *Sei $u \in \mathbb{R}$ und $f : \mathbb{R} \to \mathbb{R}$ zweinmal stetig differenzierbar mit $f(u) = 0$. Sei wieder $g(x) = x - \frac{f(x)}{f'(x)}$ die Iterationsfunktion des Newtonverfahrens. Weiters gelte $f(x) > 0$, $f'(x) > 0$ und $f''(x) \geq 0$ für alle $x > u$. Sei $x_0 > u$ beliebig gewählt und $x_n = g(x_{n-1})$ für $n \geq 1$. Dann konvergiert die Folge $(x_n)_{n\geq 0}$ monoton fallend gegen u.*

Beweis: Sei $x > u$. Dann gilt $f(x) > 0$ und $f'(x) > 0$ nach Voraussetzung, woraus $g(x) = x - \frac{f(x)}{f'(x)} < x$ folgt. Weiters ist die Funktion f' auf dem Intervall $[u, x]$ monoton wachsend ist, da dort ja $f'' \geq 0$ gilt. Es gilt somit $f'(t) \leq f'(x)$ für alle $t \in [u, x]$. Integriert man über t von u bis x, so erhält man $f(x) - f(u) \leq f'(x)(x - u)$, woraus wegen $f(u) = 0$ dann $u \leq x - \frac{f(x)}{f'(x)}$ folgt. Damit ist $u < g(x)$ gezeigt. Wir haben somit gezeigt, dass

$$u < x \quad \Rightarrow \quad u < g(x) < x \tag{1}$$

gilt. Es wird $x_0 > u$ vorausgesetzt. Aus (1) folgt dann $u < g(x_0) = x_1 < x_0$. Wenden wir nochmals (1) an, so erhalten wir $u < g(x_1) = x_2 < x_1$. Tun wir so weiter, dann ergibt sich $u < \cdots < x_5 < x_4 < x_3 < x_2 < x_1 < x_0$.
Die Folge $(x_n)_{n\geq 0}$ ist monoton fallend und nach unten durch u beschränkt. Somit hat sie einen Grenzwert v und es gilt $v \geq u$. Wäre $v > u$, dann würde $v = g(v)$ folgen, da $x_n = g(x_{n-1})$ für $n \geq 1$ gilt und g auf dem Intervall (u, ∞) stetig ist. Daraus würde dann $f(v) = 0$ folgen im Widerspruch zu $f(x) > 0$ für $x > u$. Wir haben somit $v = u$ gezeigt, das heißt $\lim_{n\to\infty} x_n = u$. □

Diesen Satz kann man auf Polynome anwenden. Sei $P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_n$ ein Polynom vom Grad $n \geq 2$ mit $a_0 > 0$, das nur reelle Nullstellen hat (zum Beispiel das charakteristische Polynom einer symmetrischen Matrix). Dann sind die Voraussetzungen von Satz 9 für die größte Nullstelle u von $P(x)$ erfüllt. Die Nullstellen von $P'(x)$ liegen nach dem Mittelwertsatz zwischen den Nullstellen von $P(x)$ und sind somit alle $\leq u$. Wegen $a_0 > 0$ folgt dann $P'(x) > 0$ für $x > u$. Dasselbe gilt für $P''(x)$. Die Nullstellen von $P''(x)$ liegen nach dem Mittelwertsatz zwischen den Nullstellen von $P'(x)$ und sind somit alle $\leq u$. Wegen $a_0 > 0$ folgt dann auch $P''(x) > 0$ für $x > u$. Wegen Satz 9 konvergiert das Newtonverfahren dann gegen u, wenn man den Startwert x_0 größer als u wählt. Man kann zeigen, dass $u < \max\{|\frac{a_n}{a_0}|, 1 + |\frac{a_{n-1}}{a_0}|, \ldots, 1 + |\frac{a_1}{a_0}|\}$ gilt. Dieses Maximum ist somit ein geeigneter Startwert.

Beispiel: Das Polynom $P(x) = x^4 + x^3 - 8x^2 - x + 4$ hat nur reelle Nullstellen. Es gilt $P'(x) = 4x^3 + 3x^2 - 16x - 1$ und $g(x) = x - \frac{P(x)}{P'(x)} = \frac{3x^4+2x^3-8x^2+4}{4x^3+3x^2-16x-1}$. Berechnet man aus den Koeffizienten des Polynoms das oben angegebene Maximum, dann erhält man 9. Startet man mit $x_0 = 9$ und iteriert mit der Funktion g so erhält man $x_5 = 2.689433$, $x_6 = 2.415169$, $x_7 = 2.326514$, $x_8 = 2.317292$, $x_9 = 2.317196$, $x_{10} = 2.317196$. Nach 10 Iterationen ändert sich der Wert (bei 7-stelliger Rechnung) nicht mehr. Damit haben wir die größte Nullstelle des Polynoms gefunden. Man kann weitere Nullstellen berechnen, indem man $P(x)$ durch $x - 2.317196$ dividiert und von dem so erhaltenen Polynom wieder die größte Nullstelle berechnet. Und so weiter.

Wählt man einen ungünstigen Startwert oder hat die Funktion gar keine Nullstelle, dann liefert das Newtonverfahren eine Folge, die nicht konvergiert.

Das Newtonverfahren funktioniert auch für komplexe Nullstellen. Wählt man eine komplexe Zahl als Startwert, die nahe genug bei einer komplexen Nullstelle liegt, dann konvergiert das Newtonverfahren dorthin.

Beispiel: Sei $f(x) = x^2 + 1$. Dann gilt $g(x) = x - \frac{f(x)}{f'(x)} = \frac{x}{2} - \frac{1}{2x}$. Ist der Startwert x_0 reell, dann wird $x_n \in \mathbb{R}$ für alle $n \geq 1$ gelten und man wird keine Nullstelle finden. In $\mathbb{R}$ existiert ja keine. Wir versuchen es in $\mathbb{C}$. Wir wählen $x_0 = 1 + i$. Es folgt dann $x_1 = \frac{1+i}{2} - \frac{1-i}{4} = \frac{1+3i}{4}$. Wir erhalten $x_2 = \frac{1+3i}{8} - \frac{1-3i}{5} = \frac{-3+39i}{40}$ und im nächsten Schritt $x_3 = \frac{-3+39i}{80} - \frac{-2-26i}{51} = \frac{7+4069i}{4080}$. Man erkennt bereits die Konvergenz gegen die Nullstelle i der Funktion f.

4. Sekantenverfahren und Regula falsi

Das Sekantenverfahren funktioniert genauso wie das Newtonverfahren, nur dass man statt der Tangente die Sekante verwendet. Sei $f : \mathbb{R} \to \mathbb{R}$ eine Funktion und r und s zwei Näherungswerte einer Nullstelle u. Die Sekante durch die Punkte $(r, f(r))$ und $(s, f(s))$ ist dann $x \mapsto f(r) + (x - r)\frac{f(s)-f(r)}{s-r}$ und $\ell(r, s) = r - \frac{(s-r)f(r)}{f(s)-f(r)} = \frac{rf(s)-sf(r)}{f(s)-f(r)}$ ist die Nullstelle der Sekante. Man benötigt zwei Startwerte x_0 und x_1. Dann kann man $x_n = \ell(x_{n-1}, x_{n-2})$ für $n \geq 2$ berechnen. Für das Konvergenzverhalten der dadurch berechneten Folge $(x_n)_{n\geq 0}$ gilt Ähnliches wie beim Newtonverfahren. Es hat den Vorteil, dass man sich das Berechnen der Ableitung erspart.

Beispiel: Wir suchen eine Nullstelle des Polynoms $P(x) = x^4 + x^3 - 8x^2 - x + 4$ mit Hilfe des Sekantenverfahrens, das heißt $x_n = \frac{x_{n-1}f(x_{n-2})-x_{n-2}f(x_{n-1})}{f(x_{n-2})-f(x_{n-1})}$ für $n \geq 2$.

Starten wir mit $x_0 = 4$ und $x_1 = 2$, dann erhalten wir $x_2 = 2.060606$, $x_3 = 2.485177$, $x_4 = 2.266073$, $x_5 = 2.308414$, $x_6 = 2.317724$, $x_7 = 2.317191$, $x_8 = 2.317196$, $x_9 = 2.317196$ und $x_{10} = 2.317196$, womit wir einen Wert erreicht haben, der sich nicht mehr ändert. Wir brechen die Iteration ab. Eine Nullstelle ist auf 7 Dezimalstellen genau berechnet.

Starten wir hingegen mit $x_0 = 0$ und $x_1 = 2$, dann erhalten wir $x_2 = 0.8$, $x_3 = 0.560461$, $x_4 = 0.691279$, $x_5 = 0.696309$, $x_6 = 0.696051$, $x_7 = 0.696052$, $x_8 = 0.696052$ und $x_9 = 0.696052$, womit wir einen Wert erreicht haben, der sich nicht mehr ändert. Wir brechen die Iteration ab. Wir haben wieder eine Nullstelle auf 6 Dezimalstellen genau berechnet, allerdings eine andere als vorher. Das liegt an der Wahl der Startwerte.

Das Sekantenverfahren hat genauso wie das Newtonverfahren den Nachteil, dass Konvergenz nicht garantiert ist. Die Regula falsi ist eine Version des Sekantenverfahrens, bei der eine Nullstelle der Funktion f zwischen zwei Werten eingeschlossen wird. Man sucht zwei Startwerte x_0 und y_0, sodass $f(x_0) < 0$ und $f(y_0) > 0$ gilt. Sind x_{n-1} und y_{n-1} mit $f(x_{n-1}) < 0$ und $f(y_{n-1}) > 0$ bereits bestimmt, dann berechnet man $c_n = \ell(x_{n-1}, y_{n-1})$, das zwischen x_{n-1} und y_{n-1} liegt. Gilt $f(c_n) > 0$, dann setzt $x_n = c_n$ und $y_n = y_{n-1}$. Gilt $f(c_n) < 0$, dann setzt $x_n = x_{n-1}$ und $y_n = c_n$. (Gilt $f(c_n) = 0$, dann bricht man ab, da eine Nullstelle gefunden ist.) Man hat dann wieder $f(x_n) < 0$ und $f(y_n) > 0$. Ist I_k das Intervall mit den Endpunkten x_k und y_k, dann gilt $I_0 \supseteq I_1 \supseteq I_2 \supseteq I_3 \supseteq \dots$ und alle diese Intervalle enthalten eine Nullstelle von f. Man hat eine Intervallschachtelung für die Nullstelle konstruiert. Allerdings kann es passieren, dass die Längen der Intervalle nicht gegen null gehen.

Eine vereinfachte Version der Regula falsi ist das Bisektionsverfahren. Es funktioniert genauso, nur wird $c_n = \frac{x_{n-1}+y_{n-1}}{2}$ genommen. Da beim Bisektionsverfahren die Länge von I_{k+1} halb so groß ist wie die von I_k, ziehen sich die Intervalle zu einem Punkt zusammen, der dann eine Nullstelle der Funktion ist.

Beispiel: Wir suchen eine Nullstelle des Polynoms $P(x) = x^4+x^3-8x^2-x+4$ mit Hilfe der Regula Falsi und des Bisektionsverfahrens. Wir starten mit $I_0 = [2, 4]$. Es gilt $f(2) < 0$ und $f(4) > 0$. Bei der Regula Falsi erhalten wir $I_5 = [2.21862, 4]$, $I_{10} = [2.29110, 4]$, $I_{15} = [2.31062, 4]$, $I_{20} = [2.31556, 4]$, $I_{25} = [2.31679, 4]$, $I_{30} = [2.31709, 4]$. Der rechte Endpunkt ist immer 4. Der linke Endpunkt konvergiert gegen eine Nullstelle, allerdings langsam. Man weiß nicht wie weit man noch von der Nullstelle entfernt ist.

Für das Bisektionsverfahren hingegen erhält man $I_1 = [2, 3]$, $I_2 = [2, 2.5]$, $I_3 = [2.25, 2.5]$, $I_4 = [2.25, 2.375]$, $I_5 = [2.3125, 2.375]$, $I_6 = [2.3125, 2.34375]$, $I_7 = [2.3125, 2.328125]$, $I_8 = [2.3125, 2.3203125]$, $I_9 = [2.31640625, 2.3203125]$, $I_{10} = [2.31640625, 2.318359375]$, $I_{11} = [2.31640625, 2.3173828125]$. Die Intervalle enthalten eine Nullstelle und ihre Längen gehen gegen null.

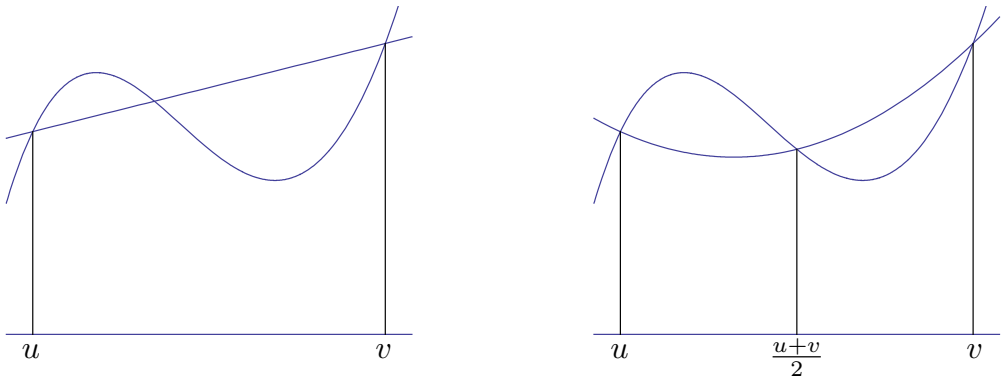
Bemerkung: Beim Bisektionsverfahren kann man das Intervall I_k auch in zehn gleich lange Intervalle teilen und dasjanige suchen, in dessen Endpunkten die Funktion verschiedenes Vorzeichen hat. Dieses Intervall ist dann I_{k+1}. Diese Vorgangsweise ist dem Dezimalsystem besser angepasst.

IV. Integration

Ein Integral $\int_a^b f(x)\,\mathrm{d}x$ soll näherungsweise berechnet werden. Wir tun das, indem wir die Funktion durch ein Polynom (üblicherweise linear oder quadratisch) approximieren, und das Integral des Polynoms berechnen.

1. Trapez– und Simpsonregel

Sei I ein Intervall und $f : I \to \mathbb{R}$ eine integrierbare Funktion. Seien u und v in I mit $u < v$. Wir erhalten eine Näherungsformel für das Integral $\int_u^v f(x)\,\mathrm{d}x$, indem wir die lineare Funktion, die in den Punkten u und v dieselben Funktionswerte wie f hat (siehe linke Zeichnung), von u bis v integrieren. Das ist auch die Fläche eines Trapezes, daher nennt man diese Näherungsformel die Trapezregel. Anstatt einer linearen Funktion kann man auch



eine quadratische Funktion bilden, die in den drei Punkten u, $\frac{u+v}{2}$ und v dieselben Funktionswerte wie f hat. Das Integral dieser quadratischen Funktion in den Grenzen von u bis v ist dann ebenfalls eine Näherungsformel für das Integral $\int_u^v f(x)\,\mathrm{d}x$, die man Simpsonregel nennt.

Im folgenden Satz wird die Trapezregel angegeben und eine Formel für den Approximationsfehler, den man bei Verwendung der Trapezregel begeht.

Satz 10 (Trapezregel) *Sei f eine zweimal stetig differenzierbare Funktion und $u < v$. Dann gilt $\int_u^v f(x)\,\mathrm{d}x = (v-u)\big(\frac{1}{2}f(u) + \frac{1}{2}f(v)\big) - \frac{1}{12}(v-u)^3 f''(\xi)$ für ein $\xi \in (u,v)$.*

Beweis: Wir beweisen den Satz zuerst für $u = -1$ und $v = 1$, das heißt für $\int_{-1}^1 g(y)\,\mathrm{d}y$, wobei g eine zweimal stetig differenzierbare Funktion ist. Sei $\ell_1(y) = py + q$ die lineare Funktion, für die $\ell_1(1) = g(1)$ und $\ell_1(-1) = g(-1)$ gilt. Wir approximieren $\int_{-1}^1 g(y)\,\mathrm{d}y$ durch $\int_{-1}^1 \ell_1(y)\,\mathrm{d}y = p\frac{y^2}{2} + qy\big|_{-1}^1 = 2q = \ell_1(-1) + \ell_1(1) = g(-1) + g(1)$. Wir setzen $h(y) = g(y) - \ell_1(y)$. Dann ist $\int_{-1}^1 h(y)\,\mathrm{d}y = \int_{-1}^1 g(y)\,\mathrm{d}y - g(-1) - g(1)$ der Fehler, den wir bei dieser Approximation machen. Weiters gilt $h(1) = h(-1) = 0$ und $h'' = g''$.

Sei $w(y) = \frac{y^2-1}{2}$. Es gilt $w''(y) = 1$ für alle y. Zweimalige partiellle Integration ergibt $\int_{-1}^1 h(y)\,\mathrm{d}y = \int_{-1}^1 w''(y)h(y)\,\mathrm{d}y = w'(y)h(y)\big|_{-1}^1 - w(y)h'(y)\big|_{-1}^1 + \int_{-1}^1 w(y)h''(y)\,\mathrm{d}y$. Nun gilt $h(1) = h(-1) = 0$ und $w(1) = w(-1) = 0$, sodass wir $\int_{-1}^1 h(y)\,\mathrm{d}y = \int_{-1}^1 w(y)h''(y)\,\mathrm{d}y$ erhalten. Wegen $w(y) \le 0$ für $y \in [-1,1]$ können wir den Mittelwertsatz der Integralrechnung anwenden und erhalten $\int_{-1}^1 h(y)\,\mathrm{d}y = h''(\eta)\int_{-1}^1 w(y)\,\mathrm{d}y$ für ein $\eta \in (-1,1)$. Wegen

$\int_{-1}^1 w(y)\,\mathrm{d}y = \frac{y^3}{6} - \frac{y}{2}\big|_{-1}^1 = -\frac{2}{3}$ und $h'' = g''$ ergibt sich

$$\int_{-1}^1 h(y)\,\mathrm{d}y = -\tfrac{2}{3}g''(\eta) \quad \text{für ein} \quad \eta \in (-1,1) \tag{1}$$

Sei jetzt $u < v$ und f erfülle die Voraussetzungen des Satzes. Sei $\psi(y) = \frac{v-u}{2}y + \frac{v+u}{2}$. Das ist eine lineare Abbildung mit $\psi(-1) = u$ und $\psi(1) = v$, sodass $\psi([-1,1]) = [u,v]$ gilt. Sei $g(y) = \frac{v-u}{2}f(\psi(y))$. Wir können dann die oben bewiesen Resultate auf diese Funktion g anwenden. Oben wurde $\int_{-1}^1 h(y)\,\mathrm{d}y = \int_{-1}^1 g(y)\,\mathrm{d}y - g(-1) - g(1)$ gezeigt. Wir berechnen $\int_{-1}^1 g(y)\,\mathrm{d}y = \int_{-1}^1 \frac{v-u}{2}f(\psi(y))\,\mathrm{d}y = \int_u^v f(x)\,\mathrm{d}x$ durch Einführen der neuen Integrationsvariable $x = \psi(y) = \frac{v-u}{2}y + \frac{v+u}{2}$. Da auch $g(-1) - g(1) = \frac{v-u}{2}(f(u) + f(v))$ gilt, erhalten wir dann $\int_{-1}^1 h(y)\,\mathrm{d}y = \int_u^v f(x)\,\mathrm{d}x - \frac{v-u}{2}(f(u) + f(v))$. Weiters haben wir $g''(\eta) = \frac{v-u}{2}f''(\psi(\eta))\psi'(\eta)^2 = (\frac{v-u}{2})^3 f''(\psi(\eta))$. Wir setzen $\xi = \psi(\eta)$. Wegen $\eta \in (-1,1)$ gilt dann $\xi \in (u,v)$ und $g''(\eta) = (\frac{v-u}{2})^3 f''(\xi)$. Setzt man das alles in (1) ein, so ergibt sich, dass $\int_u^v f(x)\,\mathrm{d}x = \frac{v-u}{2}\big(f(u) + f(v)\big) - \frac{2}{3}(\frac{v-u}{2})^3 f''(\xi)$ für ein $\xi \in (u,v)$ gilt. □

Die Trapezregel zur Berechnung von $\int_u^v f(x)\,\mathrm{d}x$ lautet somit $(v-u)(\frac{1}{2}f(u) + \frac{1}{2}f(v))$. Der Approximationsfehler lässt sich abschätzen durch $\frac{1}{12}(v-u)^3 \sup_{t\in[u,v]} |f''(t)|$, wie man in Satz 10 sieht. Der nächste Satz behandelt die Simpsonregel.

Satz 11 (Simpsonregel) *Sei f eine viermal stetig differenzierbare Funktion und $u < v$. Dann gilt $\int_u^v f(x)\,\mathrm{d}x = \frac{v-u}{2}\big(\frac{1}{3}f(u) + \frac{4}{3}f(\frac{v+u}{2}) + \frac{1}{3}f(v)\big) - \frac{1}{90}(\frac{v-u}{2})^5 f^{(4)}(\xi)$ für ein $\xi \in (u,v)$.*

Beweis: Wir beweisen den Satz zuerst für $u = -1$ und $v = 1$, das heißt für $\int_{-1}^1 g(y)\,\mathrm{d}y$, wobei g eine viermal stetig differenzierbare Funktion ist. Sei $\ell_2(y) = py^2 + qy + r$ die quadratische Funktion, für die $\ell_2(1) = g(1)$, $\ell_2(0) = g(0)$ und $\ell_2(-1) = g(-1)$ gilt. Es folgt $r = \ell_2(0) = g(0)$ und $2p + 2r = \ell_2(-1) + \ell_2(1) = g(-1) + g(1)$. Damit ergibt sich

$$\int_{-1}^1 \ell_2(y)\,\mathrm{d}y = p\tfrac{y^3}{3} + q\tfrac{y^2}{2} + ry\big|_{-1}^1 = \tfrac{2}{3}p + 2r = \tfrac{1}{3}(2p+2r) + \tfrac{4}{3}r = \tfrac{1}{3}g(-1) + \tfrac{4}{3}g(0) + \tfrac{1}{3}g(1)$$

Wir setzen $h(y) = g(y) - \ell_2(y)$. Approximiert man $\int_{-1}^1 g(y)\,\mathrm{d}y$ durch $\int_{-1}^1 \ell_2(y)\,\mathrm{d}y$, dann ist $\int_{-1}^1 h(y)\,\mathrm{d}y = \int_{-1}^1 g(y)\,\mathrm{d}y - \frac{1}{3}g(-1) - \frac{4}{3}g(0) - \frac{1}{3}g(1)$ der Fehler, den man bei dieser Approximation begeht. Weiters gilt $h(-1) = h(0) = h(1) = 0$ und $h^{(4)} = g^{(4)}$.

Sei $w(y) = \frac{y^4}{24} - \frac{y^3}{9} + \frac{y^2}{12} - \frac{1}{72}$ für $y \in [0,1]$ und $w(y) = \frac{y^4}{24} + \frac{y^3}{9} + \frac{y^2}{12} - \frac{1}{72}$ für $y \in [-1,0]$. Wir integrieren zuerst nur über das Intervall $[0,1]$. Es gilt $w^{(4)}(y) = 1$ für alle $y \in [0,1]$, also $\int_0^1 h(y)\,\mathrm{d}y = \int_0^1 w^{(4)}(y)h(y)\,\mathrm{d}y$. Durch viermalige partielle Integration erhält man

$$\int_0^1 h(y)\,\mathrm{d}y = w^{(3)}(y)h(y)\big|_0^1 - w''(y)h'(y)\big|_0^1 + w'(y)h''(y)\big|_0^1 - w(y)h^{(3)}(y)\big|_0^1 + \int_0^1 w(y)h^{(4)}(y)\,\mathrm{d}y$$

Nun gilt $h(1) = h(0) = 0$. Man berechnet $w''(0) = \frac{1}{6}$, $w'(0) = 0$, $w(0) = -\frac{1}{72}$ und $w''(1) = w'(1) = w(1) = 0$. Es folgt $\int_0^1 h(y)\,\mathrm{d}y = \frac{1}{6}h'(0) - \frac{1}{72}h^{(3)}(0) + \int_0^1 w(y)h^{(4)}(y)\,\mathrm{d}y$. Ganz analog ergibt sich $\int_{-1}^0 h(y)\,\mathrm{d}y = -\frac{1}{6}h'(0) + \frac{1}{72}h^{(3)}(0) + \int_{-1}^0 w(y)h^{(4)}(y)\,\mathrm{d}y$, wobei man $h(-1) = h(0) = 0$ und $w''(-1) = w'(-1) = w(-1) = 0$ verwendet. Addiert man diese beiden Gleichungen, so hat man $\int_{-1}^1 h(y)\,\mathrm{d}y = \int_{-1}^1 w(y)h^{(4)}(y)\,\mathrm{d}y$. Wegen $w(y) \le 0$ für $y \in [-1,1]$ können wir den Mittelwertsatz der Integralrechnung anwenden und erhalten $\int_{-1}^1 h(y)\,\mathrm{d}y = h^{(4)}(\eta)\int_{-1}^1 w(y)\,\mathrm{d}y$ für ein $\eta \in (-1,1)$. Da w eine gerade Funktion ist, ergibt sich $\int_{-1}^1 w(y)\,\mathrm{d}y = 2\int_0^1 w(y)\,\mathrm{d}y = 2(\frac{y^5}{120} - \frac{y^4}{36} + \frac{y^3}{36} - \frac{y}{72})\big|_0^1 = -\frac{1}{90}$. Da auch $h^{(4)} = g^{(4)}$ gilt, erhalten wir schließlich

$$\int_{-1}^1 h(y)\,\mathrm{d}y = -\tfrac{1}{90}g^{(4)}(\eta) \quad \text{für ein} \quad \eta \in (-1,1) \tag{1}$$

Wie im letzten Beweis sei $\psi(y) = \frac{v-u}{2}y + \frac{v+u}{2}$ und $g(y) = \frac{v-u}{2} f(\psi(y))$. Für oben definierte Funktion h wurde $\int_{-1}^{1} h(y)\,\mathrm{d}y = \int_{-1}^{1} g(y)\,\mathrm{d}y - \frac{1}{3}g(-1) - \frac{4}{3}g(0) - \frac{1}{3}g(1)$ gezeigt. Wegen $\int_{-1}^{1} g(y)\,\mathrm{d}y = \int_u^v f(x)\,\mathrm{d}x$ folgt $\int_{-1}^{1} h(y)\,\mathrm{d}y = \int_u^v f(x)\,\mathrm{d}x - \frac{v-u}{2}(\frac{1}{3}f(u) + \frac{4}{3}f(\frac{v+u}{2}) + \frac{1}{3}f(v))$. Weiters gilt $g^{(4)}(\eta) = \frac{v-u}{2} f^{(4)}(\psi(\eta))\psi'(\eta)^4 = (\frac{v-u}{2})^5 f^{(4)}(\psi(\eta))$. Wir setzen $\xi = \psi(\eta)$. Wegen $\eta \in (-1,1)$ gilt dann $\xi \in (u,v)$ und $g^{(4)}(\eta) = (\frac{v-u}{2})^5 f^{(4)}(\xi)$. Setzt man das in (1) ein, so ergibt sich, dass $\int_u^v f(x)\,\mathrm{d}x = \frac{v-u}{2}\big(\frac{1}{3}f(u) + \frac{4}{3}f(\frac{v+u}{2}) + \frac{1}{3}f(v)\big) - \frac{1}{90}(\frac{v-u}{2})^5 f^{(4)}(\xi)$ für ein $\xi \in (u,v)$ gilt. □

Die Simpsonregel zur Berechnung von $\int_u^v f(x)\,\mathrm{d}x$ lautet $\frac{v-u}{2}\big(\frac{1}{3}f(u) + \frac{4}{3}f(\frac{v+u}{2}) + \frac{1}{3}f(v)\big)$. Der Approximationsfehler lässt sich abschätzen durch $\frac{1}{90}(\frac{v-u}{2})^5 \sup_{t\in[u,v]} |f^{(4)}(t)|$, wie man in Satz 11 sieht.

Man kann weitere Formeln dieser Art herleiten, indem man Polynome höheren Grades zur Approximation heranzieht. Wir gehen noch kurz auf die $\frac{3}{8}$–Regel ein, bei der durch Polynome dritten Grades approximiert wird.

Satz 12 ($\frac{3}{8}$-Regel) *Sei f eine viermal stetig differenzierbare Funktion und $u < v$. Dann gilt* $\int_u^v f(x)\,\mathrm{d}x = \frac{v-u}{6}\big(\frac{3}{4}f(u) + \frac{9}{4}f(\frac{2u+v}{3}) + \frac{9}{4}(\frac{u+2v}{3}) + \frac{3}{4}g(v)\big) - \frac{6}{5}(\frac{v-u}{6})^5 f^{(4)}(\xi)$ *für ein $\xi \in (u,v)$.*

Beweis: Wir beweisen den Satz zuerst für $u = -3$ und $v = 3$, das heißt für $\int_{-3}^{3} g(y)\,\mathrm{d}y$, wobei g eine viermal stetig differenzierbare Funktion ist. Sei $\ell_3(y) = py^3 + qy^2 + ry + s$ jetzt ein Polynom dritten Grades, sodass $\ell_3(3) = g(3)$, $\ell_3(1) = g(1)$, $\ell_3(-1) = g(-1)$ und $\ell_3(-3) = g(-3)$ gilt. Es folgt $18q + 2s = \ell_3(-3) + \ell_3(3) = g(-3) + g(3)$ und $2q + 2s = \ell_3(-1) + \ell_3(1) = g(-1) + g(1)$. Damit ergibt sich $\int_{-3}^{3} \ell_3(y)\,\mathrm{d}y = p\frac{y^4}{4} + q\frac{y^3}{3} + r\frac{y^2}{2} + sy\big|_{-1}^{1} = 18q + 6s = \frac{3}{4}(18q + 2s) + \frac{9}{4}(2q + 2s) = \frac{3}{4}g(-3) + \frac{9}{4}g(-1) + \frac{9}{4}g(1) + \frac{3}{4}g(3)$. Wir setzen $h(y) = g(y) - \ell_3(y)$. Dann gilt $\int_{-1}^{1} h(y)\,\mathrm{d}y = \int_{-1}^{1} g(y)\,\mathrm{d}y - \frac{3}{4}g(-3) - \frac{9}{4}g(-1) - \frac{9}{4}g(1) - \frac{3}{4}g(3)$ und $h(-3) = h(-1) = h(1) = h(3) = 0$.
Wir wählen $w(y) = \frac{1}{24}y(y+3)^3$ für $y \in [-3,-1]$, $w(y) = \frac{1}{24}(y^4 - 9)$ für $y \in [-1,1]$ und $w(y) = \frac{1}{24}y(y-3)^3$ für $y \in [1,3]$. Dann zeigt man $\int_{-3}^{3} h(y)\,\mathrm{d}y = h^{(4)}(\eta)\int_{-3}^{3} w(y)\,\mathrm{d}y$ für ein $\eta \in (-3,3)$ ähnlich wie im letzten Beweis. Wegen $\int_{-3}^{3} w(y)\,\mathrm{d}y = -\frac{6}{5}$ ergibt sich

$$\int_{-3}^{3} h(y)\,\mathrm{d}y = -\tfrac{6}{5}g^{(4)}(\eta) \quad \text{für ein} \quad \eta \in (-3,3)$$

Da wir jetzt das Intervall $[-3,3]$ gewählt haben, verwenden wir $\psi(y) = \frac{v-u}{6}y + \frac{v+u}{2}$, wodurch $[-3,3]$ auf $[u,v]$ abgebildet wird. Wir setzen $g(y) = \frac{v-u}{6} f(\psi(y))$ und wenden die oben bewiesenen Resultate auf die Funktion g an. Genauso wie im letzten Beweis ergibt sich dann $\int_u^v f(x)\,\mathrm{d}x = \frac{v-u}{6}\big(\frac{3}{4}f(u) + \frac{9}{4}f(\frac{2u+v}{3}) + \frac{9}{4}(\frac{u+2v}{3}) + \frac{3}{4}g(v)\big) - \frac{6}{5}(\frac{v-u}{6})^5 f^{(4)}(\xi)$ für ein $\xi \in (u,v)$. □

Wir fassen die Approximationsformeln zur Berechnung des Integrals $\int_u^v f(x)\,\mathrm{d}x$ und die Approximationsfehler, die man dabei begeht, in folgender Tabelle zusammen.

Name	Approximationsformel	Fehler
Trapezregel	$(v-u)\big(\frac{1}{2}f(u) + \frac{1}{2}f(v)\big)$	$-\frac{1}{12}(v-u)^3 f''(\xi)$
Simpsonregel	$\frac{v-u}{2}\big(\frac{1}{3}f(u) + \frac{4}{3}f(\frac{v+u}{2}) + \frac{1}{3}f(v)\big)$	$-\frac{1}{90}(\frac{v-u}{2})^5 f^{(4)}(\xi)$
$\frac{3}{8}$–Regel	$\frac{v-u}{3}\big(\frac{3}{8}f(u) + \frac{9}{8}f(\frac{2u+v}{3}) + \frac{9}{8}(\frac{u+2v}{3}) + \frac{3}{8}g(v)\big)$	$-\frac{6}{5}(\frac{v-u}{6})^5 f^{(4)}(\xi)$

2. Zusammengesetzte Integrationsformeln

Will man ein Integral $\int_a^b f(x)\,\mathrm{d}x$ mit einem numerischen Verfahren berechnen, dann wendet man die Regeln aus dem letzten Kapitel nicht auf das gesamte Integrationsintervall $[a,b]$ an, sondern zerlegt es in n gleich lange Teilintervalle und wendet die Integrationsregel auf jedes der Teilintervalle an.

Satz 13 (Zusammengesetzte Trapezregel) *Sei f zweimal stetig differenzierbar und $a<b$. Sei $n\geq 1$. Setzt man $h=\frac{b-a}{n}$ und approximiert das Integral $\int_a^b f(x)\,\mathrm{d}x$ durch*

$$h\big(\tfrac{1}{2}f(a)+f(a+h)+f(a+2h)+f(a+3h)+\cdots+f(a+(n-2)h)+f(a+(n-1)h)+\tfrac{1}{2}f(a+nh)\big)$$

dann ist der Fehler, den man dabei begeht, durch $\frac{(b-a)^3}{12n^2}\sup_{t\in[a,b]}|f''(t)|$ beschränkt.

Beweis: Wir unterteilen das Intervall $[a,b]$ in die Teilintervalle $[a,a+h]$, $[a+h,a+2h]$, $[a+2h,a+3h]$, $\ldots$, $[a+(n-1)h,a+nh]$, wobei $a+nh=b$ gilt, da $h=\frac{b-a}{n}$ gewählt wurde. Nach der Trapezregel werden die Integrale über diese Intervalle approximiert durch $h(\frac{1}{2}f(a)+\frac{1}{2}f(a+h))$, $h(\frac{1}{2}f(a+h)+\frac{1}{2}f(a+2h))$, $h(\frac{1}{2}f(a+2h)+\frac{1}{2}f(a+3h))$, und so weiter bis $h(\frac{1}{2}f(a+(n-1)h)+\frac{1}{2}f(a+nh))$. Addiert man diese Ausdrücke, so erhält man $h(\frac{1}{2}f(a)+f(a+h)+f(a+2h)+f(a+3h)+\cdots+f(a+(n-2)h)+f(a+(n-1)h)+\frac{1}{2}f(a+nh))$ als Approximation für das Integral $\int_a^b f(x)\,\mathrm{d}x$. Der Fehler, den man für jedes dieser Teilintervalle begeht, lässt sich wegen Satz 10 abschätzen durch $\frac{1}{12}h^3\sup_{t\in[a,b]}|f''(t)|$. Da diese Abschätzung für alle Teilintervalle die gleiche ist, ist der Gesamtfehler das n-fache dieser Abschätzung, das heißt $n\frac{1}{12}h^3\sup_{t\in[a,b]}|f''(t)|=\frac{(b-a)^3}{12n^2}\sup_{t\in[a,b]}|f''(t)|$. □

Satz 14 (Zusammengesetzte Simpsonregel) *Sei f zweimal stetig differenzierbar und $a<b$. Sei $n\geq 1$. Setzt man $h=\frac{b-a}{2n}$ und approximiert das Integral $\int_a^b f(x)\,\mathrm{d}x$ durch*

$$\begin{aligned}&h\big(\tfrac{1}{3}f(a)+\tfrac{4}{3}f(a+h)+\tfrac{2}{3}f(a+2h)+\tfrac{4}{3}f(a+3h)+\tfrac{2}{3}f(a+4h)+\tfrac{4}{3}f(a+5h)+\ldots\\&\quad\cdots+\tfrac{4}{3}f(a+(2n-3)h)+\tfrac{2}{3}f(a+(2n-2)h)+\tfrac{4}{3}f(a+(2n-1)h)+\tfrac{1}{3}f(a+2nh)\big)\end{aligned}$$

dann ist der Fehler, den man dabei begeht, durch $\frac{(b-a)^5}{2880n^4}\sup_{t\in[a,b]}|f^{(4)}(t)|$ beschränkt.

Beweis: Wir unterteilen das Intervall $[a,b]$ in die Teilintervalle $[a,a+2h]$, $[a+2h,a+4h]$, $[a+4h,a+6h]$, $\ldots$, $[a+(2n-2)h,a+2nh]$, wobei $a+2nh=b$ gilt, da $h=\frac{b-a}{2n}$ gewählt wurde. Nach der Simpsonregel werden die Integrale über diese Intervalle approximiert durch $h(\frac{1}{3}f(a)+\frac{4}{3}f(a+h)+\frac{1}{3}f(a+2h))$, $h(\frac{1}{3}f(a+2h)+\frac{4}{3}f(a+3h)+\frac{1}{3}f(a+4h))$, $h(\frac{1}{3}f(a+4h)+\frac{4}{3}f(a+5h)+\frac{1}{3}f(a+6h))$, und so weiter bis zum letzten Intervall, für das man $h(\frac{1}{3}f(a+(n-1)h)+\frac{4}{3}f(a+(n-1)h)+\frac{1}{3}f(a+nh))$ erhält. Addiert man diese Ausdrücke, so ergibt sich die im Satz angegebene Approximation für das Integral $\int_a^b f(x)\,\mathrm{d}x$. Der Fehler, den man für jedes dieser Teilintervalle begeht, lässt sich wegen Satz 11 abschätzen durch $\frac{1}{90}h^5\sup_{t\in[a,b]}|f^{(4)}(t)|$. Da diese Abschätzung für alle Teilintervalle die gleiche ist, ist der Gesamtfehler das n-fache dieser Abschätzung, das heißt $n\frac{1}{90}h^5\sup_{t\in[a,b]}|f^{(4)}(t)|=\frac{(b-a)^5}{2880n^4}\sup_{t\in[a,b]}|f^{(4)}(t)|$. □

Beispiel: Das Integral $\int_0^1 f(x)\,\mathrm{d}x$ mit $f(x)=x^6\cos x$ soll auf drei Kommastellen genau berechnet werden. Wie groß muss man n in der zusammengesetzten Trapez– und Simpsonregel wählen.

Die Fehlerabschätzung in der zusammengesetzten Trapezregel ist $\frac{(b-a)^3}{12n^2}\sup_{t\in[a,b]}|f''(t)|$. Wir haben $b=1$ und $a=0$. Wir berechnen $f''(x)=30x^4\cos x-18x^5\sin x-x^6\cos x$,

sodass auf jeden Fall $|f''(x)| \leq 30 + 18 + 1 = 49$ auf dem Intervall $[0, 1]$ gilt. Somit ist der Fehler $\leq \frac{49}{12n^2}$. Wir bestimmen n so, dass $\frac{49}{12n^2} < 10^{-3}$ gilt, und erhalten $63.9 < n$. Es reicht also $n = 64$ zu wählen. Wir erhalten 0.091033 als Näherungswert für das Integral.

Die Fehlerabschätzung in der zusammengesetzten Simpsonregel ist $\frac{(b-a)^5}{2880n^4} \sup_{t\in[a,b]} |f^{(4)}(t)|$. Wir berechnen $f^{(4)}(x) = 360x^2 \cos x - 480x^3 \sin x - 180x^4 \cos x + 24x^5 \sin x + x^6 \cos x$, sodass auf jeden Fall $|f^{(4)}(x)| \leq 360+480+180+24+1 = 1045$ auf dem Intervall $[a, b] = [0, 1]$ gilt. Somit ist der Fehler $\leq \frac{1045}{2880n^4}$. Wir bestimmen n so, dass $\frac{1045}{2880n^4} < 10^{-3}$ gilt, und erhalten $4.37 < n$. Es reicht also $n = 5$ zu wählen. Wir erhalten 0.090974 als Näherungswert für das Integral.

Man sieht an diesem Beispiel, dass man bei der Simpsonregel mit einem kleineren n auskommt. Das Berechnen und Abschätzen der zweiten oder vierten Ableitung kann jedoch schwierig sein. Wir bezeichnen die Näherungsformel der zusammengesetzten Trapezregel für n Intervalle mit T_n und die der zusammengesetzten Simpsonregel für n Intervalle mit S_n. Man berechnet der Reihe nach T_2, T_4, T_8, T_{16}, T_{32}, ... (oder S_2, S_4, S_8, S_{16}, S_{32}, ...) und tut so lange weiter, bis sich die Dezimalstellen, die man genau haben will, nicht mehr ändern. Man kann dann annehmen, dass man die gewünschte Genauigkeit erreicht hat. Der Grund für die Verdopplung der Intervallanzahl n ist der, dass man dann die vorher berechneten Funktionswerte im nächsten Schritt wiederverwenden kann.

Beispiel: Das Integral $\int_0^1 f(x)\,\mathrm{d}x$ mit $f(x) = x^6 \cos x$ soll auf vier Kommastellen genau berechnet werden.

Wir berechnen $T_8 = 0.094116$, $T_{16} = 0.091766$, $T_{32} = 0.091180$, $T_{64} = 0.091033$ und $T_{128} = 0.090996$. Wir haben den Näherungswert 0.0910 für das Integral gefunden.

Ebenso berechnen wir $S_8 = 0.0909826$, $S_{16} = 0.0909842$ und $S_{32} = 0.0909843$. Wir haben den Näherungswert 0.090984 für das Integral gefunden. Dieser Wert ist bereits auf 6 Dezimalstellen genau. Man sieht wieder, dass man mit der Simpsonregel mit weniger Aufwand einen genaueren Wert erhält.

Bemerkung: Für $n \geq 1$ gilt $S_n = \frac{4}{3}T_{2n} - \frac{1}{3}T_n$.

Um diese Formel zu beweisen, setzen wir $u = \frac{b-a}{2n}$. Dann gilt
$S_n = u\big(\frac{1}{3}f(a) + \frac{4}{3}f(a+u) + \frac{2}{3}f(a+2u) + \frac{4}{3}f(a+3u) + \frac{2}{3}f(a+4u) + \frac{4}{3}f(a+5u) + \dots\big)$,
$T_n = 2u\big(\frac{1}{2}f(a) + f(a+2u) + f(a+4u) + f(a+6u) + \dots\big)$ und
$T_{2n} = u\big(\frac{1}{2}f(a) + f(a+u) + f(a+2u) + f(a+3u) + f(a+4u) + f(a+5u) + \dots\big)$.
Es folgt $\frac{4}{3}T_{2n} = u\big(\frac{2}{3}f(a)+\frac{4}{3}f(a+u)+\frac{4}{3}f(a+2u)+\frac{4}{3}f(a+3u)+\frac{4}{3}f(a+4u)+\frac{4}{3}f(a+5u)+\dots\big)$ und $\frac{1}{3}T_n = u\big(\frac{1}{3}f(a) + \frac{2}{3}f(a+2u) + \frac{2}{3}f(a+4u) + \frac{2}{3}f(a+6u) + \dots\big)$. Damit kann man dann leicht nachprüfen, dass $\frac{4}{3}T_{2n} - \frac{1}{3}T_n = S_n$ gilt.

Hat man T_4, T_8, T_{16}, T_{32}, T_{64}, ... berechnet, so kann man mit dieser Formel daraus leicht auch S_4, S_8, S_{16}, S_{32}, ... berechnen.

V. Lineare Gleichungssysteme

In diesem Kapitel geht es um das Lösen von linearen Gleichungssystemen und damit verwandten Verfahren. Diese behandeln überbestimmte Gleichungssysteme, die zum Beispiel beim Berechnen einer Regressionsgeraden auftreten, und lineare Optimierung.

1. Das Gaußsche Eliminationsverfahren

Das Gaußsche Eliminationsverfahren ist eine effiziente Methode zum Lösen von linearen Gleichungssystemen. Durch Äquivalenzumformungen wird das lineare Gleichungssystem auf Dreiecksgestalt gebracht. Damit meint man, dass alle Koeffizienten unterhalb der Diagonale gleich null sind. Dann ist es einfach, das Gleichungssystem zu lösen. Das Verfahren soll anhand des folgenden Beispiels erklärt werden.

$$\begin{aligned} 2x_1 + \ x_2 - \ 4x_3 &= -7 \\ -4x_1 + \ x_2 + 10x_3 &= \ 15 \\ 4x_1 + 8x_2 \qquad\quad &= -4 \end{aligned}$$

Um die Variablen nicht immer anschreiben zu müssen, arbeiten wir mit einer rechteckigen Koeffiziententabelle. Diese Tabelle sieht man links unten. Die rechts vom Gleichheitszeichen stehenden Zahlen sind durch einen Strich abgetrennt.

Wir ändern an der Lösung des Gleichungssystems nichts, wenn wir ein Vielfaches einer Gleichung zu einer anderen Gleichung addieren oder von ihr subtrahieren. In der ersten Spalte stehen die Zahlen 2, -4 und 4. Die in der Diagonale stehende Zahl 2 nennt man Pivotelement (in der Tabelle eingerahmt) und die erste Zeile die Pivotzeile. Um unterhalb der Diagonale Nullen zu erhalten, addieren wir das 2-fache der ersten Zeile (Pivotzeile) zur zweiten Zeile und subtrahieren das 2-fache der ersten Zeile (Pivotzeile) von der dritten Zeile (in Kurzschreibweise: $\mathrm{II} + 2\mathrm{I} \to \mathrm{II}$ und $\mathrm{III} - 2\mathrm{I} \to \mathrm{III}$). Wir erhalten die in der Mitte stehende Tabelle. Jetzt gehen wir zur zweiten Spalte. Da steht 3 in der Diagonale, das ist jetzt das Pivotelement, und unterhalb der Diagonale steht noch 6. Um unterhalb der Diagonale 0 zu erhalten, subtrahieren wir das 2-fache der zweiten Zeile, das ist jetzt die Pivotzeile, von der dritten Zeile (in Kurzschreibweise: $\mathrm{III} - 2\mathrm{II} \to \mathrm{III}$). So entsteht die rechts stehende Tabelle.

x_1	x_2	x_3	
$\boxed{2}$	1	-4	-7
-4	1	10	15
4	8	0	-4

x_1	x_2	x_3	
2	1	-4	-7
0	$\boxed{3}$	2	1
0	6	8	10

x_1	x_2	x_3	
2	1	-4	-7
0	3	2	1
0	0	4	8

Unterhalb der Diagonale stehen jetzt Nullen. Wir können das Gleichungssystem direkt lösen. Die letzte Zeile besagt, dass $4x_3 = 8$ gilt, woraus $x_3 = 2$ folgt. Die vorletzte Zeile besagt, dass $3x_2 + 2x_3 = 1$ gilt, woraus $x_2 = \frac{1}{3} - \frac{2}{3}x_3 = -1$ folgt. Die erste Zeile besagt, dass $2x_1 + x_2 - 4x_3 = -7$ gilt, woraus $x_1 = -\frac{7}{2} - \frac{1}{2}x_2 + \frac{4}{2}x_3 = 1$ folgt. Damit ist die Lösung gefunden.

Allerdings lässt sich diese Vorgangsweise so nicht durchführen, wenn das Pivotelement null ist. Liegt zum Beispiel folgendes Gleichungssystem vor

$$\begin{aligned} -\ x_2 + \ x_3 &= \ 2 \\ x_1 + 3x_2 + 2x_3 &= \ 2 \\ 2x_1 + 4x_2 - 6x_3 &= -4 \end{aligned}$$

dann wäre bei obiger Vorgangsweise das Pivotelement null und man könnte durch Addition

oder Subtraktion eines Vielfachen der ersten Zeile die Koeffizienten in der ersten Spalte nicht gleich null machen. Links unten steht die Koeffiziententabelle des Gleichungssystems. Man sucht in der ersten Spalte einen Koeffizienten, der nicht null ist, und wählt ihn als Pivotelement (in der Tabelle eingerahmt). Die Zeile, in der er steht, heißt Pivotzeile. Man nennt diese Vorgangsweise Spaltenpivotsuche. Im Beispiel wurde die dritte Zeile als Pivotzeile gewählt (es könnte auch die zweite sein). Wir vertauschen die Pivotzeile und die erste Zeile, in Kurzschreibweise $\mathrm{I} \leftrightarrow \mathrm{III}$, und führen den Eliminationsschritt $\mathrm{II} - \frac{1}{2}\mathrm{III} \to \mathrm{II}$ durch (es werden immer Vielfache der Pivotzeile addiert oder subtrahiert). So erhält man die in der Mitte stehende Tabelle. Damit haben wir Nullen in der ersten Spalte unterhalb der Diagonale. Jetzt kommen wir zur zweiten Spalte. In der Diagonale steht die Zahl 1, die wir als Pivotelement wählen (wir könnten auch die daruntersthehende Zahl -1 wählen). Die zweite Zeile ist die Pivotzeile. Der Eliminationsschritt $\mathrm{III} + \mathrm{II} \to \mathrm{III}$ führt zur rechtsstehenden Tabelle. Damit haben wir Dreiecksgestalt erreicht.

x_1	x_2	x_3	
0	-1	1	2
1	3	2	2
$\boxed{2}$	4	-6	-4

x_1	x_2	x_3	
2	4	-6	-4
0	$\boxed{1}$	5	4
0	-1	1	2

x_1	x_2	x_3	
2	4	-6	-4
0	1	5	4
0	0	6	6

Die Lösung ist jetzt wieder einfach zu berechnen. Aus $6x_3 = 6$ erhalten wir $x_3 = 1$. Aus $x_2 + 5x_3 = 4$ folgt dann $x_2 = -1$ und aus $2x_1 + 4x_2 - 6x_3 = -4$ schließlich $x_1 = 3$.

Bemerkung: Oft führt man die Spaltenpivotsuche auch dann durch, wenn sie nicht durch eine Null in der Koeffiziententabelle erzwungen wird, wie das in obigem Beispiel der Fall war. Das tut man, da die Wahl des betragsgrößten Koeffizienten in der Spalte als Pivotelement die Fortpflanzung der Rundungsfehler vermindert. Da wir jedoch nur Beispiele haben, wo exakt gerechnet wird, ist das für uns bedeutungslos.

Zur Übung lösen wir folgendes lineare Gleichungssystem mit 4 Variablen.

$$\begin{aligned} 4x_1 - 3x_2 + 4x_3 + x_4 &= 4 \\ -2x_1 + \tfrac{3}{2}x_2 - x_3 + \tfrac{3}{2}x_4 &= 5 \\ -4x_1 + 7x_2 + \tfrac{4}{3}x_3 - 5x_4 &= 12 \\ 2x_1 + \tfrac{3}{2}x_2 + \tfrac{7}{2}x_4 &= 8 \end{aligned}$$

Links unten steht die Koeffiziententabelle. Wir wählen die erste Zeile als Pivotzeile und 4 als Pivotelement. Die Eliminationsschritte $\mathrm{II} + \frac{1}{2}\mathrm{I} \to \mathrm{II}$, $\mathrm{III} + \mathrm{I} \to \mathrm{III}$ und $\mathrm{IV} - \frac{1}{2}\mathrm{I} \to \mathrm{IV}$ führen zur rechts stehenden Tabelle und sorgen dafür, dass die Koeffizienten in der ersten Spalte unterhalb der Diagonale null werden.

x_1	x_2	x_3	x_4	
$\boxed{4}$	-3	4	1	4
-2	$\frac{3}{2}$	-1	$\frac{3}{2}$	5
-4	7	$\frac{4}{3}$	-5	12
2	$\frac{3}{2}$	0	$\frac{7}{2}$	8

x_1	x_2	x_3	x_4	
4	-3	4	1	4
0	0	1	2	7
0	1	$\frac{4}{3}$	-1	4
0	$\boxed{3}$	-2	3	6

Wir kommen zur zweiten Spalte. Wir können die zweite Zeile nicht als Pivotzeile wählen, da in der Diagonale Null steht. Wir durchsuchen die zweite Spalte unterhalb der Diagonale und wählen 3 als Pivotelement und die vierte Zeile als Pivotzeile. Die Eliminationsschritte $\mathrm{IV} \leftrightarrow \mathrm{II}$ und $\mathrm{III} - \frac{1}{3}\mathrm{IV} \to \mathrm{III}$ führen zur links unten stehenden Tabelle und sorgen dafür, dass die Koeffizienten in der zweiten Spalte unterhalb der Diagonale null werden. Schließlich kommen wir zur dritten Spalte. In der Diagonale steht 2, das wir als

Pivotelement wählen. Die Pivotzeile ist somit die dritte Zeile. Der Eliminationsschritt $\mathrm{IV} - \frac{1}{2}\mathrm{III} \to \mathrm{IV}$ führt zur rechts stehenden Tabelle und sorgt dafür, dass der Koeffizient in der dritten Spalte unterhalb der Diagonale null wird.

x_1	x_2	x_3	x_4	
4	−3	4	1	4
0	3	−2	3	6
0	0	$\boxed{2}$	−2	2
0	0	1	2	7

x_1	x_2	x_3	x_4	
4	−3	4	1	4
0	3	−2	3	6
0	0	2	−2	2
0	0	0	3	6

Die Lösung ist jetzt wieder einfach zu berechnen. Aus $3x_4 = 6$ erhalten wir $x_4 = 2$ und aus $2x_3 - 2x_4 = 2$ folgt dann $x_3 = 3$. Aus $3x_2 - 2x_3 + 3x_4 = 6$ ergibt sich $x_2 = 2$ und aus $4x_1 - 3x_2 + 4x_3 + x_4 = 4$ schließlich $x_1 = -1$.

Es gibt auch lineare Gleichungssysteme, deren Lösung nicht eindeutig ist. Wir untersuchen diesen Fall wieder anhand eines Beispiels. Gesucht ist die Lösung von

$$\begin{aligned} -2x_2 - 3x_3 + x_4 &= 6 \\ 4x_1 - 2x_2 - 8x_3 &= 8 \\ 2x_1 + x_2 - x_3 + 3x_4 &= 6 \\ 2x_1 + 3x_2 + 2x_3 + 4x_4 &= 4 \end{aligned}$$

Wir führem das Eliminationsverfahren durch. Links unten steht die Koeffiziententabelle. Da in der ersten Spalte das Diagonalelement gleich Null ist, wählen wir die dritte Zeile als Pivotzeile mit 2 als Pivotelement. Die Eliminationsschritte $\mathrm{I} \leftrightarrow \mathrm{III}$, $\mathrm{II} - 2\mathrm{III} \to \mathrm{II}$ und $\mathrm{IV} - \mathrm{III} \to \mathrm{IV}$ führen zur rechts stehenden Tabelle.

x_1	x_2	x_3	x_4	
0	−2	−3	1	6
4	−2	−8	0	8
$\boxed{2}$	1	−1	3	6
2	3	2	4	4

x_1	x_2	x_3	x_4	
2	1	−1	3	6
0	$\boxed{-4}$	−6	−6	−4
0	−2	−3	1	6
0	2	3	1	−2

Jetzt ist -4 das Diagonalelement der zweiten Spalte. Wir wählen es als Pivotelement und die zweite Zeile als Pivotzeile. Die Eliminationsschritte $\mathrm{III} - \frac{1}{2}\mathrm{II} \to \mathrm{III}$ und $\mathrm{IV} + \frac{1}{2}\mathrm{II} \to \mathrm{IV}$ führen zur links unten stehenden Tabelle. Jetzt kommen wir zur dritten Spalte. Dort steht sowohl in als auch unterhalb der Diagonale eine Null. Daher gehen wir gleich weiter zur vierten Spalte. Wir wählen die dritte Zeile als Pivotzeile mit 4 als Pivotelement. Der Eliminationsschritt $\mathrm{IV} + \frac{1}{2}\mathrm{III} \to \mathrm{IV}$ führt zur rechts stehenden Tabelle.

x_1	x_2	x_3	x_4	
2	1	−1	3	6
0	−4	−6	−6	−4
0	0	0	$\boxed{4}$	8
0	0	0	−2	−4

x_1	x_2	x_3	x_4	
2	1	−1	3	6
0	−4	−6	−6	−4
0	0	0	4	8
0	0	0	0	0

Die vierte Gleichung ist $0 = 0$. Die dritte Gleichung ist $4x_4 = 8$, woraus $x_4 = 2$ folgt. Die zweite Gleichung ist $-4x_2 - 6x_3 - 6x_4 = -4$, woraus $x_2 = -2 - \frac{3}{2}x_3$ folgt. Wir können x_3 beliebig wählen. Wir wählen einen Parameter $t \in \mathbb{R}$ und setzen $x_3 = t$. Dann erhalten wir $x_2 = -2 - \frac{3}{2}t$. Aus der ersten Gleichung $2x_1 + x_2 - x_3 + 3x_4 = 6$ folgt dann $x_1 = 1 + \frac{5}{4}t$. Wir haben eine einparametrige Famile von Lösungen erhalten.

Was passiert, wenn ein lineares Gleichungssystem keine Lösung hat? Ändert man in obigem Beispiel die letzte Gleichung um in $2x_1 + 3x_2 + 2x_3 + 4x_4 = 6$ und führt das

Eliminationsverfahren durch genauso wie wir es oben getan haben, dann endet man mit folgender Tabelle (es ändert sich nur die Zahl rechts unten in der Tabelle)

x_1	x_2	x_3	x_4	
2	1	−1	3	6
0	−4	−6	−6	−4
0	0	0	4	8
0	0	0	0	2

Die vierte Gleichung lautet $0 = 2$, ein Widerspruch. Somit gibt es keine Lösung.

2. Determinante und inverse Matrix.

Mit Hilfe des Eliminationsverfahrens kann man auch die Determinante einer Matrix A berechnen. Addiert man ein Vielfaches einer Zeile von A zu einer anderen Zeile oder subtrahiert es von ihr, dann ändert sich die Determinante nicht. Transformiert man die Matrix A durch Eliminationsschritte in eine Dreiecksmatrix B, dann gilt $\det A = \det B$. Die Determinante der Dreiecksmatrix B ist das Produkt ihrer Diagonalelemente. Das ist dann auch die Determinante von A. Allerdings gilt das nur, wenn man keine Zeilenvertauschungen vornimmt. Jede Zeilenvertauschung ändert das Vorzeichen der Determinante. Es gilt dann $\det A = (-1)^p \det B$, wobei p die Anzahl der Zeilenvertauschungen ist.

Diese Methode zur Berechnung der Determinante soll an einem Beispiel demonstriert werden. Gesucht ist die Determinante der Matrix

$$A = \begin{pmatrix} 0 & -2 & -5 & -4 \\ -2 & -4 & 2 & 6 \\ 1 & 3 & 0 & -1 \\ 2 & 5 & 2 & -3 \end{pmatrix}$$

Unten wird die Transformation der Matrix A in eine Dreiecksmatrix mit Hilfe von Eliminationsschritten durchgeführt. Ganz links ist nochmals die Matrix A aufgeschrieben. Wir beginnen mit der ersten Spalte. Wir müssen eine Pivotsuche durchführen. Um die Rechnung zu vereinfachen, wählen wir die dritte Zeile als Pivotzeile mit 1 als Pivotelement. Die Eliminationsschritte $\mathrm{I} \leftrightarrow \mathrm{III}$, $\mathrm{II} + 2\mathrm{III} \to \mathrm{II}$ und $\mathrm{IV} - 2\mathrm{III} \to \mathrm{IV}$ ergeben die zweite Tabelle und sorgen dafür, dass in der ersten Spalte unterhalb der Diagonale Nullen stehen. Jetzt kommen wir zur zweiten Spalte. Wir wählen die Diagonaleintragung 2 in der zweiten Spalte als Pivotelement und somit die zweite Zeile als Pivotzeile. Die Eliminationsschritte $\mathrm{III} + \mathrm{II} \to \mathrm{III}$ und $\mathrm{IV} + \frac{1}{2}\mathrm{II} \to \mathrm{IV}$ ergeben die dritte Tabelle. Jetzt stehen auch in der zweiten Spalte unterhalb der Diagonale Nullen. Schließlich kommen wir zur dritten Spalte. Wir wählen die dritte Zeile als Pivotzeile mit der Diagonaleintragung -3 als Pivotelement. Der Eliminationsschritt $\mathrm{IV} + \mathrm{III} \to \mathrm{IV}$ ergibt die rechts stehende Tabelle.

$$\begin{pmatrix} 0 & -2 & -5 & -4 \\ -2 & -4 & 2 & 6 \\ \boxed{1} & 3 & 0 & -1 \\ 2 & 5 & 2 & -3 \end{pmatrix} \quad \begin{pmatrix} 1 & 3 & 0 & -1 \\ 0 & \boxed{2} & 2 & 4 \\ 0 & -2 & -5 & -4 \\ 0 & -1 & 2 & -1 \end{pmatrix} \quad \begin{pmatrix} 1 & 3 & 0 & -1 \\ 0 & 2 & 2 & 4 \\ 0 & 0 & \boxed{-3} & 0 \\ 0 & 0 & 3 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 3 & 0 & -1 \\ 0 & 2 & 2 & 4 \\ 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Damit ist die Matrix in eine Dreiecksmatrix transformiert. Die Determinante der Dreiecksmatrix ist das Produkt der Diagonalelemente, also -6. Da wir eine Zeilenvertauschung durchgeführt haben, erhalten wir $\det A = 6$, die Determinante mit geändertem Vorzeichen.

Man kann das Gaußsche Eliminationsverfahren auch zum Berechnen der inversen Matrix verwenden. Die k-te Spalte der Inversen einer Matrix A ist die Lösung $\mathbf{x}$ des linearen

Gleichungssystems $A\mathbf{x} = \mathbf{e}_k$, wobei $\mathbf{e}_k$ der k-te Einheitsvektor ist. Man kann diese Gleichungssysteme alle auf einmal lösen, indem man in der Koeffiziententabelle rechts vom senkrechten Strich alle Vektoren $\mathbf{e}_k$ hinschreibt, sodass dort die Einheitsmatrix steht, und dann das Eliminationsverfahren durchführt. Wir wenden jedoch ein erweitertes Eliminationsverfahren an, das die Lösung gleich in einem Durchgang berechnet. Wir erlauben nicht nur Eliminationsschritte wie oben, bei denen ein Vielfaches der Pivotzeile zu anderen Zeilen addiert oder von ihnen subtrahiert wird, sondern auch die Multiplikation der Pivotzeile mit einer Zahl ungleich Null. Wenn wir beides zulassen, sprechen wir von Zeilenoperationen. Mit Hilfe von Zeilenoperationen muss die links vom senkrechten Strich stehende Matrix A in die Einheitsmatrix übergeführt werden. Steht rechts vom senkrechten Strich ein Vektor $\mathbf{v}$, dann geht er dabei in die Lösung des Gleichungssystems $A\mathbf{x} = \mathbf{v}$ über. Steht aber rechts vom senkrechten Strich nicht nur ein Vektor, sondern die Einheitsmatrix, deren Spalten die Vektoren $\mathbf{e}_k$ sind, dann entstehen durch diese Zeilenoperationen rechts vom Strich die Lösungen der Gleichungssysteme $A\mathbf{x} = \mathbf{e}_k$ und somit die inverse Matrix A^{-1}.

Dieses Verfahren zur Berechnung der inversen Matrix soll an links unten stehender Matrix A demonstriert werden. Rechts sieht man die Matrix A mit angefügter Einheitsmatrix.

$$A = \begin{pmatrix} 0 & 2 & 1 & 0 \\ -2 & 6 & -4 & -2 \\ 0 & -4 & -2 & 1 \\ 2 & -3 & 5 & 1 \end{pmatrix} \qquad \left(\begin{array}{cccc|cccc} 0 & 2 & 1 & 0 & 1 & 0 & 0 & 0 \\ \boxed{-2} & 6 & -4 & -2 & 0 & 1 & 0 & 0 \\ 0 & -4 & -2 & 1 & 0 & 0 & 1 & 0 \\ 2 & -3 & 5 & 1 & 0 & 0 & 0 & 1 \end{array}\right)$$

Im ersten Durchgang soll aus der ersten Spalte der Einheitsvektor $\mathbf{e}_1$ werden. Da die Diagonaleintragung der ersten Spalte null ist, wählen wir -2 als Pivotelement und die zweite Zeile als Pivotzeile. Die Zeilenoperationen $\mathrm{I} \to \mathrm{II}$, $-\frac{1}{2}\mathrm{II} \to \mathrm{I}$ und $\mathrm{IV} + \mathrm{II} \to \mathrm{IV}$ machen aus der ersten Spalte den ersten Einheitsvektor $\mathbf{e}_1$ und ergeben die links unten stehende Tabelle. Wir kommen zur zweiten Spalte. Dort ist die Diagonaleintragung jetzt 2. Wir wählen sie als Pivotelement und die zweite Zeile als Pivotzeile. Die Zeilenoperationen $\mathrm{I} + \frac{3}{2}\mathrm{II} \to \mathrm{I}$, $\frac{1}{2}\mathrm{II} \to \mathrm{II}$, $\mathrm{III} + 2\mathrm{II} \to \mathrm{III}$ und $\mathrm{IV} - \frac{3}{2}\mathrm{II} \to \mathrm{IV}$ machen aus der zweiten Spalte den zweiten Einheitsvektor $\mathbf{e}_2$ und ergeben die rechtsstehende Tabelle.

$$\left(\begin{array}{cccc|cccc} 1 & -3 & 2 & 1 & 0 & -\frac{1}{2} & 0 & 0 \\ 0 & \boxed{2} & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -4 & -2 & 1 & 0 & 0 & 1 & 0 \\ 0 & 3 & 1 & -1 & 0 & 1 & 0 & 1 \end{array}\right) \qquad \left(\begin{array}{cccc|cccc} 1 & 0 & \frac{7}{2} & 1 & \frac{3}{2} & -\frac{1}{2} & 0 & 0 \\ 0 & 1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & \boxed{-\frac{1}{2}} & -1 & -\frac{3}{2} & 1 & 0 & 1 \end{array}\right)$$

Wir kommen zur dritten Spalte. Die Diagonaleintragung ist null. Wir wählen daher die vierte Zeile als Pivotzeile mit $-\frac{1}{2}$ als Pivotelement. Die Zeilenoperationen $\mathrm{I} + 7\mathrm{IV} \to \mathrm{I}$, $\mathrm{II} + \mathrm{IV} \to \mathrm{II}$, $-2\mathrm{IV} \to \mathrm{III}$ und $\mathrm{III} \to \mathrm{IV}$ machen aus der dritten Spalte den dritten Einheitsvektor $\mathbf{e}_3$ und ergeben die links unten stehende Tabelle. Es bleibt schließlich noch die vierte Spalte. Wir wählen die vierte Zeile als Pivotzeile mit 1 als Pivotelement. Die Zeilenoperationen $\mathrm{I} + 6\mathrm{IV} \to \mathrm{I}$, $\mathrm{II} + \mathrm{IV} \to \mathrm{II}$ und $\mathrm{III} - 2\mathrm{IV} \to \mathrm{III}$ machen die vierte Spalte zum vierten Einheitsvektor $\mathbf{e}_4$ und ergeben die rechtsstehende Tabelle

$$\left(\begin{array}{cccc|cccc} 1 & 0 & 0 & -6 & -9 & \frac{13}{2} & 0 & 7 \\ 0 & 1 & 0 & -1 & -1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 & 3 & -2 & 0 & -2 \\ 0 & 0 & 0 & \boxed{1} & 2 & 0 & 1 & 0 \end{array}\right) \qquad \left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 3 & \frac{13}{2} & 6 & 7 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & -1 & -2 & -2 & -2 \\ 0 & 0 & 0 & 1 & 2 & 0 & 1 & 0 \end{array}\right)$$

Rechts vom senkrechten Strich ist die Inverse der Matrix A entstanden. Indem man diese Matrix mit A multipliziert, kann man nachprüfen, ob man sich nicht verrechnet hat.

VI. Lineare Optimierung

In der linearen Optimierung geht es darum, das Maximum oder Minimum einer linearen Funktion zu finden unter Nebenbedingungen, die ebenfalls linear sind.

1. Graphische Lösung

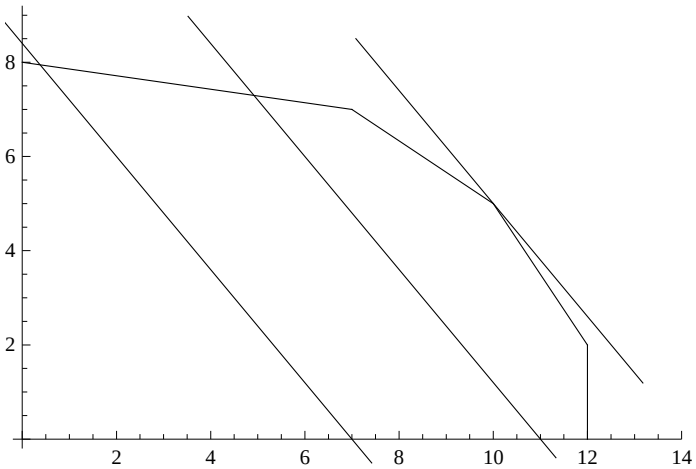
Wir beginnen mit einem Beispiel: Eine Firma stellt zwei Produkte P_1 und P_2 her, wobei drei verschiedene Maschinen M_1, M_2 und M_3 benutzt werden. Für die Herstellung einer Mengeneinheit des Produkts P_1 werden 1 Maschinenstunde auf M_1, 2 Maschinenstunden auf M_2 und 3 Maschinenstunden auf M_3 benötigt. Für die Herstellung einer Mengeneinheit des Produkts P_2 werden 7 Maschinenstunden auf M_1, 3 Maschinenstunden auf M_2 und 2 Maschinenstunden auf M_3 benötigt. Pro Woche steht Maschine M_1 für 56 Stunden zur Verfügung, Maschine M_2 für 35 Stunden und Maschine M_3 für 40 Stunden. Außerdem können pro Woche höchstens 12 Mengeneinheiten von Produkt P_1 abgesetzt werden. Für eine Mengeneinheit von P_1 erzielt man einen Preis von von 6 Euro und für eine Mengeneinheit von P_2 einen Preis von von 5 Euro. Wie viele Mengeneinheiten von P_1 und P_2 muss man pro Woche produzieren, um den Gewinn zu maximieren?

Seien x_1 und x_2 die Mengeneinheiten, die von P_1 und P_2 produziert werden. Aus obigen Angaben lassen sich folgende Ungleichungen bilden.

$$\begin{array}{rl} x_1 + 7x_2 \leq 56 & \\ 2x_1 + 3x_2 \leq 35 & \qquad x_1 \geq 0 \\ 3x_1 + 2x_2 \leq 40 & \qquad x_2 \geq 0 \\ x_1 \leq 12 & \end{array} \qquad 6x_1 + 5x_2 \to \text{Max}$$

Die Maschinenkapazitäten und die beschränkte Absatzmöglichkeit für P_1 ergeben die links stehenden Ungleichungen. Die Mengeneinheiten sind natürlich größer oder gleich null. Das ergibt $x_1 \geq 0$ und $x_2 \geq 0$. Der Gewinn für die produzierten Mengeneinheiten ist $6x_1 + 5x_2$ und soll ein Maximum annehmen. Man nennt $z = 6x_1 + 5x_2$ die Zielfunktion.

Die Menge G aller $(x_1, x_2) \in \mathbb{R}^2$, die obige Ungleichungen erfüllen, kann man zeichnen. Zum Beispiel bilden die Punkte $(x_1, x_2) \in \mathbb{R}^2$, für die $x_1 + 7x_2 \leq 56$ gilt, die Halbebene, die durch die Gerade $x_1 + 7x_2 = 56$ begrenzt wird und den Punkt $(0, 0)$ enthält. Analoges gilt für die anderen Ungleichungen. Nebenstehend ist die Menge G gezeichnet, die obige Ungleichungen erfüllt. Außerdem sind die Geraden $6x_1 + 5x_2 = c$ für $c = 42$, $c = 66$ und $c = 85$ eingezeichnet. Auf diesen Geraden nimmt die Zielfunktion den Wert c an. Gesucht ist das maximale c, sodass die Gerade $6x_1 + 5x_2 = c$ noch einen Punkt der durch die Ungleichungen bestimmten Menge G enthält. Das ist für $c = 85$ der Fall. Die Gerade $6x_1 + 5x_2 = 85$ verläuft durch den Punkt $(10, 5)$, der noch in der Menge G liegt. Für $c > 85$ hat die Gerade jedoch leeren Durchschnitt mit G.



Somit ist $x_1 = 10$ und $x_2 = 5$ die Lösung des Optimierungsproblems. Für diese Werte sind alle Ungleichungen erfüllt und die Zielfunktion nimmt ihren maximalen Wert 85 an.

2. Das Simplexverfahren

Diese graphische Lösungsmethode funktioniert jedoch nur für zwei Variable. Hätte man nicht zwei, sondern drei oder vier Produkte, dann hätte man auch drei oder vier Variable und man könnte das Optimierungsproblem nicht mehr graphisch lösen. Deshalb gibt es ein anderes Lösungsverfahren, die sogenannte Simplexmethode. Wir werden diese Simplexmethode anhand des obigen Beispiels behandeln.

Im ersten Schritt werden sogenannte Schlupfvariable eingeführt, und zwar für jede Ungleichung eine. Wir bezeichnen sie mit x_3, x_4, x_5 und x_6. So werden aus den Ungleichungen Gleichungen. Das Optimierungsproblem lässt sich dann so schreiben

$$\begin{array}{rrrrl|ll|l} x_1 & +7x_2 & +x_3 & & = 56 & x_3 \geq 0 & & \\ 2x_1 & +3x_2 & & +x_4 & = 35 & x_4 \geq 0 & x_1 \geq 0 & 6x_1+5x_2 \to \text{ Max} \\ 3x_1 & +2x_2 & & +x_5 & = 40 & x_5 \geq 0 & x_2 \geq 0 & \\ x_1 & & & +x_6 & = 12 & x_6 \geq 0 & & \end{array}$$

Wir können umformen. Wir können ein Vielfaches einer Gleichung zu einer anderen addieren. Wir können eine Gleichung mit einer Zahl $\neq 0$ multiplizieren. Dadurch verändern wir nichts. Die Zielfunktion $z = 6x_1 + 5x_2$, die wir als $z = 0 - (-6x_1 - 5x_2)$ schreiben, können wir ebenfalls umformen. Zum Beispiel können wir 72 addieren und $6x_1 + 6x_6$ subtrahieren, da wegen der vierten Gleichung ja $6x_1 + 6x_6 = 72$ gilt. Wir erhalten dann $z = 72+5x_2-6x_6$, das wir als $z = 72-(-5x_2+6x_6)$ schreiben. Wir fassen die Gleichungen und die Zielfunktion in einer Koeffiziententabelle zusammen

x_1	x_2	x_3	x_4	x_5	x_6	
1	7	1	0	0	0	56
2	3	0	1	0	0	35
3	2	0	0	1	0	40
$\boxed{1}$	0	0	0	0	1	12
-6	-5	0	0	0	0	0

Die ersten vier Zeilen stellen die Gleichungen dar, die letzte Zeile die Zielfunktion, wobei ganz rechts die Konstante steht und davor die Koeffizienten der Variablen wie sie in der Darstellung $z = 0 - (-6x_1 - 5x_2)$ auftreten. Wie bereits oben erklärt, können wir dann ein Vielfaches einer Gleichung (einer der ersten vier Zeilen) zur letzten Zeile addieren, ohne an der Zielfunktion etwas zu ändern. Das Gleichungssystem, das aus 4 Gleichungen mit 6 Variablen besteht, steht bereits in Diagonalform da, sodass man Lösungen leicht angeben kann, indem man x_1 und x_2 irgendwie wählt, und die anderen Variablen aus den Gleichungen berechnet. Die Variablen x_3, x_4, x_5 und x_6 heißen Basisvariable, da die Spalten unter diesen Variablen die Standardbasis eines Vektorraums bilden. Die anderen Variablen x_1 und x_2 nennen wir Nullvariable. Wir setzen nämlich $x_1 = 0$ und $x_2 = 0$. Dann kann man aus der Tabelle sofort $x_3 = 56$, $x_4 = 35$, $x_5 = 40$ und $x_6 = 12$ ablesen. Der Wert der Zielfunktion ist 0. Alle Variablen sind ≥ 0, daher haben wir eine zulässige Lösung (die Lösung kann rechts vom senkrechten Strich abgelesen werden).

Um eine zulässige Lösung zu finden, für die die Zielfunktion einen größeren Wert ergibt, suchen wir die Nullvariable, die in der Zielfunktion den größten Koeffizienten hat, das ist die Nullvariable, wo in der letzten Zeile der Tabelle die negative Zahl mit dem größten Betrag steht. Es ist die Variable x_1. In obiger Lösung ist $x_1 = 0$. Vergrößern wir x_1, dann

wird auch die Zielfunktion größer, da x_1 in der Zielfunktion einen Koeffizienten > 0 hat.

Da alle Variablen ≥ 0 sein müssen, erhalten wir aus den Gleichungen in obiger Tabelle die Ungleichungen $x_1 \leq 56$, $2x_1 \leq 35$, $3x_1 \leq 40$ und $x_1 \leq 12$. Der maximale Wert, den x_1 annehmen kann, ist 12. In diesem Fall wird die vierte Ungleichung zur Gleichung. Wir wählen daher die vierte Zeile als Pivotzeile mit 1 in der x_1-Spalte als Pivotelement. Durch Zeilenoperationen soll in der ersten Spalte der Einheitsvektor $\mathbf{e}_4$ entstehen, das heißt alle Eintragungen dieser Spalte werden null, nur das Pivotelement wird zu 1. Dazu sind die Zeilenoperationen $\mathrm{I} - \mathrm{IV} \to \mathrm{I}$, $\mathrm{II} - 2\mathrm{IV} \to \mathrm{II}$, $\mathrm{III} - 3\mathrm{IV} \to \mathrm{III}$ und $\mathrm{V} + 6\mathrm{IV} \to \mathrm{V}$ notwendig (die letzte Zeilenoperation führt die oben besprochene Änderung der Zielfunktion durch), die folgende Tabelle ergeben.

x_1	x_2	x_3	x_4	x_5	x_6	
0	7	1	0	0	-1	44
0	3	0	1	0	-2	11
0	$\boxed{2}$	0	0	1	-3	4
1	0	0	0	0	1	12
0	-5	0	0	0	6	72

Jetzt sind x_1, x_3, x_4 und x_5 die Basisvariablen (in diesen Spalten stehen Standardbasisvektoren) und x_2 und x_6 die Nullvariablen. Wir setzten $x_2 = 0$ und $x_6 = 0$ und erhalten $x_1 = 12$, $x_3 = 44$, $x_4 = 11$, $x_5 = 4$ und $z = 72 - (-5x_2 + 6x_6) = 72$ aus der Tabelle. Alle Variablen sind ≥ 0, daher haben wir eine zulässige Lösung.

Wir haben eine zulässige Lösung mit einem größeren Wert für die Zielfunktion gefunden. Um die Zielfunktion weiter zu vergrößeren, suchen wir die Nullvariable, die in der Zielfunktion den größten Koeffizienten hat (negative Eintragung in der letzten Zeile). Es ist die Variable x_2. In obiger Lösung ist $x_2 = 0$. Wenn wir x_2 größer machen, wird auch die Zielfunktion größer.

Wir ändern an $x_6 = 0$ nichts, nur an den anderen Variablen. Da diese ≥ 0 sein müssen, erhalten wir aus den Gleichungen in obiger Tabelle die Ungleichungen $7x_2 \leq 44$, $3x_2 \leq 11$, $2x_2 \leq 4$ und $0x_2 \leq 12$. Der maximale Wert, den x_2 annehmen kann, ist somit 2. In diesem Fall wird die dritte Ungleichung zur Gleichung. Wir wählen daher die dritte Zeile als Pivotzeile mit 2 in der x_2-Spalte als Pivotelement. Die Zeilenoperationen $\mathrm{I} - \frac{7}{2}\mathrm{III} \to \mathrm{I}$, $\mathrm{II} - \frac{3}{2}\mathrm{III} \to \mathrm{II}$, $\frac{1}{2}\mathrm{III} \to \mathrm{III}$ und $\mathrm{V} + \frac{5}{2}\mathrm{III} \to \mathrm{V}$ erzeugen den Einheitsvektor $\mathbf{e}_3$ in der der x_2-Spalte, das heißt alle Eintragungen dieser Spalte werden null, nur das Pivotelement wird zu 1. Wir erhalten folgende Tabelle

x_1	x_2	x_3	x_4	x_5	x_6	
0	0	1	0	-3.5	9.5	30
0	0	0	1	-1.5	$\boxed{2.5}$	5
0	1	0	0	0.5	-1.5	2
1	0	0	0	0	1	12
0	0	0	0	2.5	-1.5	82

Jetzt sind x_1, x_2, x_3 und x_4 die Basisvariablen (in diesen Spalten stehen Standardbasisvektoren) und x_5 und x_6 die Nullvariablen. Wir setzten $x_5 = 0$ und $x_6 = 0$ und erhalten $x_1 = 12$, $x_2 = 2$, $x_3 = 30$, $x_4 = 5$ und $z = 82 - (2.5x_5 - 1.5x_6) = 82$ aus der Tabelle. Alle Variablen sind ≥ 0, daher haben wir eine zulässige Lösung.

Es gibt noch immer eine Nullvariable, die in der Zielfunktion einen Koeffizienten > 0 hat (negative Eintragung in der letzten Zeile), nämlich x_6. Wir können daher die Zielfunktion weiter vergrößern, indem wir x_6 größer machen.

Wir ändern an $x_5 = 0$ nichts, nur an den anderen Variablen. Da diese ≥ 0 sein müssen, erhalten wir die Ungleichungen $9.5x_6 \leq 30$, $2.5x_6 \leq 5$, $-1.5x_6 \leq 2$ und $x_6 \leq 12$. Genau dann sind alle diese Ungleichungen erfüllt, wenn $x_6 \leq 2$ gilt. Der maximale Wert, den x_6 annehmen kann, ist somit 2. In diesem Fall wird die zweite Ungleichung zur Gleichung. Wir wählen daher die zweite Zeile als Pivotzeile mit 2.5 in der x_6-Spalte als Pivotelement. Die Zeilenoperationen $\mathrm{I} - \frac{19}{5}\,\mathrm{II} \to \mathrm{I}$, $\mathrm{III} + \frac{3}{5}\,\mathrm{II} \to \mathrm{III}$, $\mathrm{IV} - \frac{2}{5}\,\mathrm{II} \to \mathrm{IV}$, $\mathrm{V} + \frac{3}{5}\,\mathrm{II} \to \mathrm{V}$ und $\frac{2}{5}\,\mathrm{II} \to \mathrm{II}$ machen die x_6-Spalte zum Einheitsvektor $\mathbf{e}_2$ und führen zu

x_1	x_2	x_3	x_4	x_5	x_6	
0	0	1	−3.8	2.2	0	11
0	0	0	0.4	−0.6	1	2
0	1	0	0.6	−0.4	0	5
1	0	0	−0.4	0.6	0	10
0	0	0	0.6	1.6	0	85

Die Nullvariablen sind x_4 und x_5. Die Zielfunktion ist $z = 85 - (0.6x_4 + 1.6x_5)$. Die Koeffizienten der Nullvariablen in der Zielfunktion sind negativ. Durch Vergrößern der Nullvariablen wird die Zielfunktion nicht größer, sondern kleiner. Wir haben den maximalen Wert gefunden, den die Zielfunktion annehmen kann. Die Werte für die Variablen, in denen die Zielfunktion das Maximum annimmt, sind $x_1 = 10$, $x_2 = 5$, $x_3 = 11$, $x_4 = 0$, $x_5 = 0$ und $x_6 = 2$. Man muss also 10 Mengeneinheiten des Produkts P_1 und 5 Mengeneinheiten des Produkts P_2 produzieren, um den maximalen Gewinn von 85 Euro zu erzielen. Die Variablen x_3, x_4 und x_5 kann man als freie Maschinenkapazitäten interpretieren. Auf Maschine M_1 bleiben 11 Maschinenstunden pro Woche ungenützt, während die Maschinen M_2 und M_3 ausgelastet sind. Weiters besagt $x_6 = 2$, dass der maximal mögliche Absatz von Produkt P_1 um 2 Mengeneinheiten pro Woche unterschritten wird.

Wir behandeln noch ein Beispiel mit drei Variablen. Hat man in obigem Beispiel drei Produkte und vier Maschinen, dann könnte das folgendes Optimierungsproblem ergeben.

$$\begin{aligned} 2x_1 + x_2 &\leq 12 \\ x_1 + 2x_2 + x_3 &\leq 12 \\ 3x_1 + 4x_2 + 5x_3 &\leq 38 \\ 2x_2 + x_3 &\leq 8 \end{aligned} \qquad \begin{aligned} x_1 &\geq 0 \\ x_2 &\geq 0 \\ x_3 &\geq 0 \end{aligned} \qquad 8x_1 + 7x_2 + 5x_3 \to \text{Max}$$

Wir führen Schlupfvariablen x_4, x_5, x_6 und x_7 ein und schreiben die dadurch entstehenden Gleichungen und die Zielfunktion wie oben in eine Koeffiziententabelle

x_1	x_2	x_3	x_4	x_5	x_6	x_7	
$\boxed{2}$	1	0	1	0	0	0	12
1	2	1	0	1	0	0	12
3	4	5	0	0	1	0	38
0	2	1	0	0	0	1	8
−8	−7	−5	0	0	0	0	0

Die Schlupfvariablen x_4, x_5, x_6 und x_7 sind die Basisvariablen. In den Spalten unter diesen Variablen stehen Einheitsvektoren. Die anderen drei Variablen x_1, x_2 und x_3 sind die Nullvariablen. Rechts vom senkrechten Strich kann man die Werte der Basisvariablen und der Zielfunktion ablesen. Um eine zulässige Lösung zu finden, für die die Zielfunktion einen größeren Wert ergibt, suchen wir die Nullvariable, unter der in der letzten Zeile der Tabelle die negative Zahl mit dem größten Betrag steht. Es ist die Variable x_1. Da alle Variablen ≥ 0 sein müssen, erhalten wir aus den Gleichungen in obiger Tabelle die

Ungleichungen $2x_1 \leq 12$, $x_1 \leq 12$, $3x_1 \leq 38$ und $0x_1 \leq 8$. Der maximale Wert, den x_1 annehmen kann, ist 6. In diesem Fall wird die erste Ungleichung zur Gleichung. Wir wählen daher die erste Zeile als Pivotzeile mit 2 in der x_1-Spalte als Pivotelement. Die Zeilenoperationen $\mathrm{II} - \frac{1}{2}\mathrm{I} \to \mathrm{II}$, $\mathrm{III} - \frac{3}{2}\mathrm{I} \to \mathrm{III}$, $\mathrm{V} + 4\mathrm{II} \to \mathrm{V}$ und $\frac{1}{2}\mathrm{I} \to \mathrm{I}$ machen die x_1-Spalte zum Einheitsvektor $\mathbf{e}_1$ und ergeben folgende Tabelle

x_1	x_2	x_3	x_4	x_5	x_6	x_7	
1	0.5	0	0.5	0	0	0	6
0	1.5	1	−0.5	1	0	0	6
0	2.5	$\boxed{5}$	−1.5	0	1	0	20
0	2	1	0	0	0	1	8
0	−3	−5	4	0	0	0	48

Jetzt sind x_1, x_5, x_6 und x_7 die Basisvariablen und x_2, x_3 und x_4 die Nullvariablen. In der letzten Zeile gibt es negative Eintragungen. Die Lösung ist noch nicht optimal. Die betragsgrößte negative Zahl in der letzten Zeile steht unter der Variable x_3. Wir ändern an $x_2 = 0$ und $x_4 = 0$ nichts, nur an den anderen Variablen. Da diese ≥ 0 sein müssen, erhalten wir aus den Gleichungen in obiger Tabelle die Ungleichungen $0x_3 \leq 6$, $x_3 \leq 6$, $5x_3 \leq 20$ und $x_3 \leq 8$. Der maximale Wert, den x_3 annehmen kann, ist somit 4. In diesem Fall wird die dritte Ungleichung zur Gleichung. Wir wählen daher die dritte Zeile als Pivotzeile mit 5 in der x_3-Spalte als Pivotelement. Die Zeilenoperationen $\mathrm{II} - \frac{1}{5}\mathrm{III} \to \mathrm{II}$, $\mathrm{IV} - \frac{1}{5}\mathrm{III} \to \mathrm{IV}$, $\mathrm{V} + \mathrm{III} \to \mathrm{V}$ und $\frac{1}{5}\mathrm{III} \to \mathrm{III}$ machen die x_3-Spalte zum Einheitsvektor $\mathbf{e}_3$ und ergeben folgende Tabelle

x_1	x_2	x_3	x_4	x_5	x_6	x_7	
1	0.5	0	0.5	0	0	0	6
0	$\boxed{1}$	0	−0.2	1	−0.2	0	2
0	0.5	1	−0.3	0	0.2	0	4
0	1.5	0	0.3	0	−0.2	1	4
0	−0.5	0	2.5	0	1	0	68

Jetzt sind x_1, x_3, x_5 und x_7 die Basisvariablen und x_2, x_4 und x_6 die Nullvariablen. Es gibt noch eine Nullvariable mit negativer Eintragung in der letzten Zeile, nämlich x_2. Wir ändern an $x_4 = 0$ und $x_6 = 0$ nichts, nur an den anderen Variablen. Da diese ≥ 0 sein müssen, erhalten wir die Ungleichungen $0.5x_2 \leq 6$, $x_2 \leq 2$, $0.5x_2 \leq 4$ und $1.5x_2 \leq 4$. Der maximale Wert, den x_2 annehmen kann, ist 2. In diesem Fall wird die zweite Ungleichung zur Gleichung. Wir wählen daher die zweite Zeile als Pivotzeile mit 1 in der x_2-Spalte als Pivotelement. Die Zeilenoperationen $\mathrm{I} - \frac{1}{2}\mathrm{II} \to \mathrm{I}$, $\mathrm{III} - \frac{1}{2}\mathrm{II} \to \mathrm{III}$, $\mathrm{IV} - \frac{3}{2}\mathrm{II} \to \mathrm{IV}$ und $\mathrm{V} + \frac{1}{2}\mathrm{II} \to \mathrm{V}$ machen die x_2-Spalte zum Einheitsvektor $\mathbf{e}_2$ und ergeben folgende Tabelle

x_1	x_2	x_3	x_4	x_5	x_6	x_7	
1	0	0	0.6	−0.5	0.1	0	5
0	1	0	−0.2	1	−0.2	0	2
0	0	1	−0.2	−0.5	0.3	0	3
0	0	0	0.6	−1.5	0.1	1	1
0	0	0	2.4	0.5	0.9	0	69

Jetzt gibt es in der letzten Zeile keine negativen Eintragungen mehr. Die optimale Lösung ist gefunden. Die Werte für die Variablen, in denen die Zielfunktion das Maximum annimmt, sind $x_1 = 5$, $x_2 = 2$, $x_3 = 3$, $x_4 = 0$, $x_5 = 0$, $x_6 = 0$ und $x_7 = 1$. Der Wert der Zielfunktion ist 69. Das ist der maximal mögliche Wert, den die Zielfunktion unter den vorgegeben Nebenbedingungen annehmen kann.

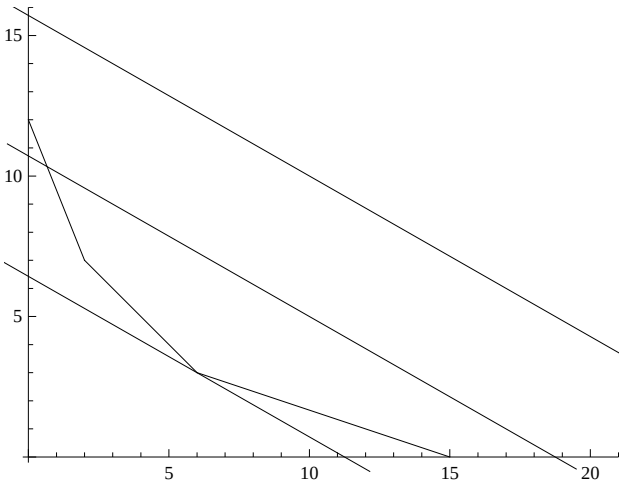
3. Minimumprobleme

Wir behandeln ein Beispiel: Ein Landwirt verfügt über zwei Sorten von Düngemitteln. Ein Sack der Sorte 1 enthält 5 kg Kalzium, 3 kg Stickstoff und 1 kg Phosphor. Ein Sack der Sorte 2 enthält 2 kg Kalzium, 3 kg Stickstoff und 3 kg Phosphor. Der Preis der Sorte 1 verhält sich zum Preis der Sorte 2 wie 4:7. Es soll eine möglichst billige Mischung hergestellt werden, die mindestens 24 kg Kalzium, 27 kg Stickstoff und 15 kg Phosphor enthält. Ist x_1 die Anzahl der Säcke von Sorte 1 und x_2 die Anzahl der Säcke von Sorte 2, die man mischt, dann ergibt sich folgendes Optimierungsproblem

$$\begin{array}{r} 5x_1+\ 2x_2 \geq 24 \\ 3x_1+\ 3x_2 \geq 27 \\ x_1+\ 3x_2 \geq 15 \end{array} \qquad \begin{array}{l} x_1 \geq 0 \\ x_2 \geq 0 \end{array} \qquad 4x_1+7x_2 \to \text{ Min}$$

Die Ungleichungen sind jetzt umgedreht und die Zielfunktion soll minimal werden.

Die Menge G aller $(x_1, x_2) \in \mathbb{R}^2$, die obige Ungleichungen erfüllen, ist das Gebiet oberhalb des Polygonzuges in nebenstehender Zeichnung. Weiters sind die Geraden $4x_1 + 7x_2 = c$ für $c = 45$, $c = 75$ und $c = 110$ eingezeichnet. Auf diesen Geraden nimmt die Zielfunktion jeweils den Wert c an. Gesucht ist das minimale c, sodass die Gerade $4x_1 + 7x_2 = c$ noch einen Punkt der Menge G enthält. Das ist für $c = 45$ der Fall. Die Gerade $4x_1 + 7x_2 = 45$ verläuft durch den Punkt $(6, 3)$, der noch in der Menge G liegt. Für $c < 45$ hat die Gerade leeren Durchschnitt mit G. Somit ist $x_1 = 6$ und $x_2 = 3$ die Lösung des Optimierungsproblems. Für diese Werte von x_1 und x_2 sind alle Ungleichungen erfüllt und die Zielfunktion nimmt ihren minimalen Wert 45 an.



Man kann auch ein Minimumproblem mit dem Simplexverfahren lösen. Wir führen Schlupfvariable x_3, x_4 und x_5 ein und formulieren das gestellte Problem folgendermaßen um (man beachte, dass die Schlupfvariablen in den Gleichungen negatives Vorzeichen haben)

$$\begin{array}{rrrr} 5x_1+2x_2-x_3 & & = 24 \\ 3x_1+3x_2 & -x_4 & = 27 \\ x_1+3x_2 & & -x_5 = 15 \end{array} \qquad \begin{array}{l} x_1 \geq 0 \\ x_2 \geq 0 \end{array} \qquad \begin{array}{l} x_3 \geq 0 \\ x_4 \geq 0 \\ x_5 \geq 0 \end{array} \qquad -4x_1-7x_2 \to \text{ Max}$$

Jetzt haben wir ein Maximumproblem, da die Zielfunktion mit -1 multipliziert wurde. Wir stellen genauso wie früher die Koeffiziententabelle auf:

x_1	x_2	x_3	x_4	x_5	
5	2	-1	0	0	24
3	3	0	-1	0	27
$\boxed{1}$	3	0	0	-1	15
4	7	0	0	0	0

Jetzt läuft das Simplexverfahren genauso ab wie vorher. Allerdings haben wir das Problem, nicht mit der Lösung $x_1 = x_2 = 0$ beginnen zu können, da diese Lösung nicht zulässig ist. Dann wäre $x_3 = -24$, $x_4 = -27$ und $x_3 = -15$ und negative Werte sind nicht zulässig.

Wir müssen eine zulässige Lösung suchen. Es gibt dafür ein Verfahren (zweistufiges Simplexverfahren), aber darauf gehen wir nicht ein. Wir versuchen es mit $x_2 = 0$ und machen x_1 gerade so groß, dass alle Variablen ≥ 0 werden. Aus der dritten Gleichung sieht man, dass x_1 mindestens 15 sein muss. Das genügt auch. Es ergibt sich die Lösung $x_1 = 15$, $x_2 = 0$, $x_3 = 51$, $x_4 = 18$ und $x_5 = 0$. Das ist eine zulässige Lösung, da alle Variablen ≥ 0 sind. Die Nullvariablen sind x_2 und x_5, die anderen Variablen müssen demnach die Basisvariablen sein. Bevor wir mit dem Simplexverfahren beginnen können, müssen wir dafür sorgen, dass in der Tabelle unter den Basisvariablen Einheitsvektoren stehen. Für die Spalten unter x_3 und x_4 ist das so gut wie erfüllt, wir müssen uns jedoch um die x_1-Spalte kümmern. Wir wählen die dritte Zeile als Pivotzeile, das ist die, in der die Schlupfvariable x_5 gleich null wird, mit 1 in der x_1-Spalte als Pivotelement. Die Zeilenoperationen $-\mathrm{I} + 5\mathrm{III} \to \mathrm{I}$, $-\mathrm{II} + 3\mathrm{III} \to \mathrm{II}$ und $\mathrm{IV} - 4\mathrm{III} \to \mathrm{IV}$ machen die x_1-Spalte zum Einheitsvektor $\mathbf{e}_3$ und ergeben folgende Tabelle

x_1	x_2	x_3	x_4	x_5	
0	13	1	0	−5	51
0	$\boxed{6}$	0	1	−3	18
1	3	0	0	−1	15
0	−5	0	0	4	−60

Jetzt können wir starten. Unter den Basisvariablen stehen die Einheitsvektoren. Der Wert der Zielfunktion ist negativ, da wir sie mit -1 multipliziert haben. Es gibt nur mehr eine Variable, unter der in der letzten Zeile eine negative Eintragung steht, nämlich x_2. Da x_5 null ist und auch null bleibt und die anderen Variablen ≥ 0 sind, müssen die Ungleichungen $13x_2 \leq 51$, $6x_2 \leq 18$ und $3x_2 \leq 15$ gelten. Der größte Wert, den x_2 annehmen kann, ist somit 3. Gilt $x_2 = 3$, dann wird die zweite Ungleichung zur Gleichung. Wir wählen daher die zweite Zeile als Pivotzeile mit 6 in der x_2-Spalte als Pivotelement. Die Zeilenoperationen $\mathrm{I} - \frac{13}{6}\mathrm{II} \to \mathrm{I}$, $\mathrm{III} - \frac{1}{2}\mathrm{II} \to \mathrm{III}$, $\mathrm{IV} + \frac{5}{6}\mathrm{II} \to \mathrm{IV}$ und $\frac{1}{6}\mathrm{II} \to \mathrm{II}$ machen die x_2-Spalte zum Einheitsvektor $\mathbf{e}_2$ und ergeben folgende Tabelle

x_1	x_2	x_3	x_4	x_5	
0	0	1	$-\frac{13}{6}$	−5	12
0	1	0	$\frac{1}{6}$	0	3
1	0	0	$-\frac{1}{2}$	−1	6
0	0	0	$\frac{5}{6}$	4	−45

Es gibt keine Variable mehr, unter der in der letzten Zeile eine negative Eintragung steht. Damit haben wir die Lösung gefunden, in der die mit -1 multiplizierte Zielfunktion das Maximum -45 und die ursprüngliche Zielfunktion das Minimum 45 annimmt. Diese Lösung ist $x_1 = 6$, $x_2 = 3$, $x_3 = 12$, $x_4 = 0$ und $x_5 = 0$. Der Punkt, in dem die Zielfunktion ihr Minimum 45 annimmt, ist somit $x_1 = 6$ und $x_2 = 3$.

VII. Ausgleichsrechnung

In der Ebene sind k Punkte gegeben. Gesucht ist eine Gerade, sodass die Punkte möglichst geringen Abstand von der Geraden haben. Diese Gerade heißt Regressionsgerade. Statt einer Geraden kann man auch ein Polynom wählen. Wir behandeln auch eine kontinuierliche Version dieser Fragestellung.

1. Überbestimmte Gleichungssyteme

Sei M eine $k \times n$-Matrix mit $k > n$ und $\mathbf{c} \in \mathbb{R}^k$. Da das lineare Gleichungssystem $M\mathbf{x} = \mathbf{c}$ aus k Gleichungen besteht und n Variable enthält, also mehr Gleichungen als Variable, darf man nicht erwarten, dass eine Lösung existiert. Daher sucht man einen Vektor $\mathbf{x} \in \mathbb{R}^n$, sodass $\|M\mathbf{x} - \mathbf{c}\|$ minimal wird.

Satz 15: *Sei M eine $k \times n$-Matrix. Wir nehmen an, dass ein $\mathbf{u} \in \mathbb{R}^n$ existiert mit $M^tM\mathbf{u} = M^t\mathbf{c}$ (man kann zeigen, dass das immer der Fall ist). Für einen Vektor $\mathbf{x} \in \mathbb{R}^n$ ist $\|M\mathbf{x} - \mathbf{c}\|$ minimal, genau dann wenn $M^tM\mathbf{x} = M^t\mathbf{c}$ gilt.*

Beweis: Mit Hilfe von $M^tM\mathbf{u} - M^t\mathbf{c} = \mathbf{0}$ erhalten wir

$$\langle M\mathbf{y}, M\mathbf{u} - \mathbf{c}\rangle = \langle \mathbf{y}, M^tM\mathbf{u} - M^t\mathbf{c}\rangle = 0 \quad \text{für alle} \quad \mathbf{y} \in \mathbb{R}^n$$

Daraus ergibt sich dann

$$\begin{aligned}\|M\mathbf{x} - \mathbf{c}\|^2 &= \|M\mathbf{x} - M\mathbf{u} + M\mathbf{u} - \mathbf{c}\|^2 = \langle M\mathbf{x} - M\mathbf{u} + M\mathbf{u} - \mathbf{c}, M\mathbf{x} - M\mathbf{u} + M\mathbf{u} - \mathbf{c}\rangle \\ &= \langle M\mathbf{x} - M\mathbf{u}, M\mathbf{x} - M\mathbf{u}\rangle + 2\langle M\mathbf{x} - M\mathbf{u}, M\mathbf{u} - \mathbf{c}\rangle + \langle M\mathbf{u} - \mathbf{c}, M\mathbf{u} - \mathbf{c}\rangle \\ &= \|M\mathbf{x} - M\mathbf{u}\|^2 + \|M\mathbf{u} - \mathbf{c}\|^2\end{aligned}$$

Da $\|M\mathbf{u} - \mathbf{c}\|^2$ nicht von $\mathbf{x}$ abhängt, ist $\|M\mathbf{x} - \mathbf{c}\|$ genau dann minimal, wenn $M\mathbf{x} = M\mathbf{u}$ gilt. Wir zeigen, dass das äquivalent zu $M^tM\mathbf{x} = M^t\mathbf{c}$ ist.
Es gelte $M\mathbf{x} = M\mathbf{u}$. Durch Multiplikation mit M^t erhalten wir $M^tM\mathbf{x} = M^tM\mathbf{u}$. Wegen $M^tM\mathbf{u} = M^t\mathbf{c}$ folgt $M^tM\mathbf{x} = M^t\mathbf{c}$.
Es gelte $M^tM\mathbf{x} = M^t\mathbf{c}$. Wegen $M^tM\mathbf{u} = M^t\mathbf{c}$ folgt $M^tM\mathbf{x} = M^tM\mathbf{u}$. Damit erhalten wir $\|M(\mathbf{x} - \mathbf{u})\|^2 = \langle M(\mathbf{x} - \mathbf{u}), M(\mathbf{x} - \mathbf{u})\rangle = \langle M^tM(\mathbf{x} - \mathbf{u}), \mathbf{x} - \mathbf{u}\rangle = 0$ und somit $M(\mathbf{x} - \mathbf{u}) = \mathbf{0}$, das heißt $M\mathbf{x} = M\mathbf{u}$. □

Ist ein Vektor $\mathbf{x}$ gesucht, für den $\|M\mathbf{x} - \mathbf{c}\|$ minimal wird, dann lösen wir das lineare Gleichungssystem $M^tM\mathbf{x} = M^t\mathbf{c}$. Nach Satz 15 nimmt $\|M\mathbf{x} - \mathbf{c}\|$ sein Minimum genau auf der Lösungsmenge an.

Wir rechnen ein einfaches Beispiel. Für die Matrix M und den Vektor $\mathbf{c}$ links unten suchen wir $\mathbf{x} = \binom{x_1}{x_2} \in \mathbb{R}^2$, sodass $\|M\mathbf{x} - \mathbf{c}\|$ minimal wird. Es ist das Gleichungssystem $M^tM\mathbf{x} = M^t\mathbf{c}$ zu lösen. Daher wurde auch gleich M^tM und $M^t\mathbf{c}$ berechnet.

$$M = \begin{pmatrix} 2 & 1 \\ 2 & 2 \\ 1 & 1 \end{pmatrix} \quad \mathbf{c} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad M^tM = \begin{pmatrix} 2 & 2 & 1 \\ 1 & 2 & 1 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 2 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 9 & 7 \\ 7 & 6 \end{pmatrix} \quad M^t\mathbf{c} = \begin{pmatrix} 2 & 2 & 1 \\ 1 & 2 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

Wir lösen das Gleichungssystem $\binom{9\ \ 7}{7\ \ 6}\binom{x_1}{x_2} = \binom{2}{1}$. Die Lösung ist $x_1 = 1$ und $x_2 = -1$. Somit wird $\|M\mathbf{x} - \mathbf{c}\|$ für $\mathbf{x} = \binom{1}{-1}$ minimal.

2. Die Methode der kleinsten Quadrate

Die wichtigste Anwendung von Satz 15 ist die Berechnung einer Regressionsgerade. In der Ebene sind k Punkte (x_1, y_1), (x_2, y_2), ... , (x_k, y_k) gegeben. Gesucht ist eine Gerade $y = a_1x + a_0$, sodass $\sum_{i=1}^{k}(a_1x_i + a_0 - y_i)^2$ minimal wird. Die Summe der Quadrate der vertikalen Abstände der Punkte von der Geraden soll möglichst klein sein.

Die Unbekannten sind a_1 und a_0. Sei M die $k \times 2$-Matrix mit den Spalten $(1, 1, \ldots, 1)$ und $(x_1, x_2, \ldots, x_k)$. Sei $\mathbf{c}$ der Spaltenvektor $(y_1, y_2, \ldots, y_k)$. Dann sind a_1 und a_0 so zu bestimmen, dass $\|M\binom{a_0}{a_1} - \mathbf{c}\|$ minimal wird. Wir setzen

$$S_0 = k \qquad S_1 = \textstyle\sum_{i=1}^{k} x_i \qquad S_2 = \sum_{i=1}^{k} x_i^2 \qquad T_0 = \sum_{i=1}^{k} y_i \qquad T_1 = \sum_{i=1}^{k} x_iy_i$$

Dann gilt $M^tM = \binom{S_0\ \ S_1}{S_1\ \ S_2}$ und $M^t\mathbf{c} = \binom{T_0}{T_1}$. Nach Satz 15 ist das Gleichungssystem bestehend aus $S_0a_0 + S_1a_1 = T_0$ und $S_1a_0 + S_2a_1 = T_1$ zu lösen. Die Lösung ist

$$a_0 = \tfrac{S_2T_0 - S_1T_1}{S_0S_2 - S_1^2} \quad \text{und} \quad a_1 = \tfrac{S_0T_1 - S_1T_0}{S_0S_2 - S_1^2}$$

Für die k Punkte (x_1, y_1), (x_2, y_2), ... , (x_k, y_k) haben wir damit die Regressionsgerade $y = a_1x + a_0$ berechnet.

Statt der Geraden $y = a_1x + a_0$ kann man auch ein Polynom $y = \sum_{j=0}^{n-1} a_jx^j$ wählen. Wir setzen $\varphi_j(x) = x^j$ für $j \geq 0$, damit wir später für φ_j auch andere Funktionen einsetzen können, und schreiben $P(x) = \sum_{j=0}^{n-1} a_j\varphi_j(x)$. Wir suchen so ein verallgemeinertes Polynom $P(x)$, sodass die k Punkte (x_1, y_1), (x_2, y_2), ... , (x_k, y_k) möglichst nahe an der Kurve $y = P(x)$ liegen. Die Koeffizienten $a_0, a_1, \ldots, a_{n-1}$ in $P(x)$ sind so zu bestimmen, dass $\sum_{i=1}^{k}(P(x_i) - y_i)^2$ minimal wird. Die Summe der Quadrate der vertikalen Abstände der Punkte von der Kurve soll möglichst klein sein.

Sei M die $k \times n$-Matrix mit den Spalten $(\varphi_j(x_1), \varphi_j(x_2), \ldots, \varphi_j(x_k))$ für $0 \leq j \leq n-1$ und $\mathbf{c}$ der Spaltenvektor $(y_1, y_2, \ldots, y_k)$. Sei $\mathbf{a}$ der Spaltenvektor $(a_0, a_1, \ldots, a_{n-1})$. Dann ist $M\mathbf{a}$ der Vektor $(P(x_1), P(x_2), \ldots, P(x_k))$ und $\mathbf{a}$ ist so zu bestimmen, dass $\|M\mathbf{a} - \mathbf{c}\|$ minimal wird. Wir nehmen an, dass $k > n$ gilt und die Spalten von M linear unabhängig sind. Die Voraussetzung von Satz 15 sind erfüllt. Es gilt $M^tM = R = (R_{uv})_{0 \leq u,v \leq n-1}$ mit $R_{uv} = \sum_{i=1}^{k} \varphi_u(x_i)\varphi_v(x_i)$ und $M^t\mathbf{c} = \mathbf{q} = (q_v)_{0 \leq v \leq n-1}$ mit $q_v = \sum_{i=1}^{k} \varphi_v(x_i)y_i$. Nach Satz 15 ist das Gleichungssystem $R\mathbf{a} = \mathbf{q}$ zu lösen. Die Lösung $\mathbf{a}$ gibt die Koeffizienten in $P(x)$ an, für die $\sum_{i=1}^{k}(P(x_i) - y_i)^2$ minimal wird.

Wir behandeln zwei Spezialfälle, zuerst ein Regressionspolynom, eine Verallgemeinerung der Regressionsgerade. Man wählt $P(x) = \sum_{j=0}^{n-1} a_jx^j$, ein Polynom vom Grad $\leq n-1$. Das ist der Spezialfall $\varphi_j(x) = x^j$. Wir setzen $S_r = \sum_{i=1}^{k} x_i^r$ und $T_r = \sum_{i=1}^{k} x_i^r y_i$ für $r \geq 0$. Dann gilt $R_{uv} = S_{u+v}$ und $q_v = T_v$. Wir erhaltem also

$$R = M^tM = \begin{pmatrix} S_0 & S_1 & \cdots & S_{n-1} \\ S_1 & S_2 & \cdots & S_n \\ \vdots & \vdots & \ddots & \vdots \\ S_{n-1} & S_n & \cdots & S_{2n-2} \end{pmatrix} \qquad \mathbf{q} = M^t\mathbf{c} = \begin{pmatrix} T_0 \\ T_1 \\ \vdots \\ T_{n-1} \end{pmatrix}$$

Zu lösen ist das Gleichungssystem $R\mathbf{a} = \mathbf{q}$. Die Lösung $\mathbf{a}$ gibt die Koeffizienten im Polynom $P(x)$ an, für die $\sum_{i=1}^{k}(P(x_i) - y_i)^2$ minimal wird. Der Spezialfall $n = 2$ ergibt die oben behandelte Regressionsgerade.

Beispiel: Wir suchen die Regressionspolynome vom Grad 1 und Grad 2 für die Punkte $(-3,1)$, $(-2,0)$, $(-1,1)$, $(0,6)$, $(1,5)$, $(2,6)$ und $(3,5)$. Da die x-Werte symmetrisch zum Nullpunkt liegen, erhalten wir $S_1 = S_3 = 0$. Weiters ist $S_0 = 7$, die Anzahl der Punkte. Wir berechnen $S_2 = 2(3^2+2^2+1^2) = 28$ und $S_4 = 2(3^4+2^4+1^4) = 196$. Weiters ist $T_0 = 24$, die Summe der y-Werte. Wir berechnen $T_1 = -3\cdot 1-2\cdot 0-1\cdot 1+0\cdot 6+1\cdot 5+2\cdot 6+3\cdot 5 = 28$ und $T_2 = 3^2\cdot 1+2^2\cdot 0+1^2\cdot 1+0^2\cdot 6+1^2\cdot 5+2^2\cdot 6+3^2\cdot 5 = 84$.
Um die Regressionsgerade zu berechnen, ist $R\mathbf{a} = \mathbf{q}$ zu lösen mit $R = \left(\begin{smallmatrix} 7 & 0 \\ 0 & 28 \end{smallmatrix}\right)$ und $q = \binom{24}{28}$. Da R bereits eine Diagonalmatrix ist, kann man $a_0 = \frac{24}{7}$ und $a_1 = 1$ leicht berechnen. Die Regressionsgerade ist $y = x + \frac{24}{7}$.
Um die Regressionspolynom zweiten Grades zu berechnen, ist $R\mathbf{a} = \mathbf{q}$ zu lösen mit

$$R = \begin{pmatrix} 7 & 0 & 28 \\ 0 & 28 & 0 \\ 28 & 0 & 196 \end{pmatrix} \qquad \mathbf{q} = \begin{pmatrix} 24 \\ 28 \\ 84 \end{pmatrix}$$

Links unten steht die Koeffiziententabelle zu diesem Gleichungssystem. Mit einem Eliminationsschritt erhält man die rechtsstehende Tabelle, die bereits Dreiecksgestalt hat.

a_0	a_1	a_2	
7	0	28	24
0	28	0	28
28	0	196	84

a_0	a_1	a_2	
7	0	28	24
0	28	0	28
0	0	84	−12

Man berechnet daraus $a_2 = -\frac{1}{7}$, $a_1 = 1$ und $a_0 = \frac{24}{7} + \frac{4}{7} = 4$. Damit ist das Regressionspolynom $y = -\frac{1}{7}x^2 + x + 4$ gefunden.

Der zweite Spezialfall, den wir behandeln, sind trigonometrische Polynome. Wir setzen $n = 2m+1$ für ein $m \geq 1$. Wir wählen $\varphi_0(x) = 1$ und $\varphi_j(x) = \cos jx$ und $\varphi_{m+j}(x) = \sin jx$ für $1 \leq j \leq m$. Wir erhalten dann $P(x) = a_0 + \sum_{j=1}^{m} a_j \cos jx + \sum_{j=1}^{m} a_{m+j} \sin jx$. So eine Funktion nennt man ein trigonometrisches Polynom. Durch Einsetzen in obige Formeln kann man R_{uv} und q_u berechnen und das Gleichungssystem $R\mathbf{a} = \mathbf{q}$ lösen. Die Lösung $\mathbf{a}$ gibt dann wieder die Koeffizienten im trigonometrischen Polynom $P(x)$ an, für die $\sum_{i=1}^{k}(P(x_i) - y_i)^2$ minimal wird.

3. Fourierapproximation

Bisher waren n Punkte in der Ebene vorgegeben und gesucht war ein Polynom, das möglichst gut an diese Punkte herankommt. Jetzt geben wir ein beschränktes Intervall I vor und eine integrierbare Funktion $g : I \to \mathbb{R}$ und suchen ein Polynom, das die Funktion g auf dem Intervall I möglichst gut approximiert. Wir tun das wieder für ein verallgemeinertes Polynom $P(x) = \sum_{j=0}^{n-1} a_j \varphi_j(x)$, wobei die Funktionen φ_j stetig und vorgegeben sind, zum Beispiel $\varphi_j(x) = x^j$ für $j \geq 0$. Zum Approximieren verwenden wir eine kontinuierliche Version der Methode der kleinsten Quadrate: Die Koeffizienten $a_0, a_1, \ldots, a_{n-1}$ im verallgemeinerten Polynom $P(x)$ sind so zu bestimmen, dass $\int_I (P(x) - g(x))^2 \,\mathrm{d}x$ minimal wird. Die Lösung findet man auf dieselbe Art wie im letzten Kapitel.

Satz 16: *Sei $R = (R_{uv})_{0\leq u,v\leq n-1}$ mit $R_{uv} = \int_I \varphi_u(x)\varphi_v(x)\,\mathrm{d}x$ und sei $\mathbf{q} = (q_v)_{0\leq v\leq n-1}$ mit $q_v = \int_I \varphi_v(x)g(x)\,\mathrm{d}x$. Wir nehmen an, dass das Gleichungssystem $R\mathbf{a} = \mathbf{q}$ eine eindeutige Lösung $\mathbf{a} = (a_0, a_1, \ldots, a_{n-1})$ besitzt. Diese Lösung gibt dann die Koeffizienten in $P(x) = \sum_{j=0}^{n-1} a_j\varphi_j(x)$ an, für die $\int_I (P(x) - g(x))^2\,\mathrm{d}x$ minimal wird.*

Beweis: Wir setzen $Q(x) = \sum_{j=0}^{n-1} a_j \varphi_j(x)$, wobei die Koeffizienten die Komponenten des Lösungsvektors des Gleichungssystems $R\mathbf{a} = \mathbf{q}$ sind. Dann gilt für $0 \le i \le n-1$

$$\begin{aligned}\int_I \varphi_i(x)(Q(x) - g(x))\,\mathrm{d}x &= \int_I \varphi_i(x)(\textstyle\sum_{j=0}^{n-1} a_j\varphi_j(x) - g(x))\,\mathrm{d}x \\ &= \textstyle\sum_{j=0}^{n-1} a_j \int_I \varphi_i(x)\varphi_j(x)\,\mathrm{d}x - \int_I \varphi_i(x)g(x)\,\mathrm{d}x = \sum_{j=0}^{n-1} a_j R_{ij} - q_i = 0\end{aligned}$$

Es folgt, dass $\int_I (\sum_{j=0}^{n-1} c_j\varphi_i(x))(Q(x) - f(x))\,\mathrm{d}x = 0$ gilt für jedes verallgemeinerte Polynom $\sum_{j=0}^{n-1} c_j\varphi_i(x)$. Damit erhalten wir

$$\begin{aligned}&\int_I (P(x) - g(x))^2\,\mathrm{d}x = \int_I (P(x) - Q(x) + Q(x) - g(x))^2\,\mathrm{d}x \\ &\quad = \int_I (P(x) - Q(x))^2\,\mathrm{d}x + 2\int_I (P(x) - Q(x))(Q(x) - g(x))\,\mathrm{d}x + \int_I (Q(x) - g(x))^2\,\mathrm{d}x \\ &\quad = \int_I (P(x) - Q(x))^2\,\mathrm{d}x + \int_I (Q(x) - g(x))^2\,\mathrm{d}x\end{aligned}$$

Da $\int_I (Q(x) - g(x))^2\,\mathrm{d}x$ nicht von $P(x)$ abhängt, ist $\int_I (P(x) - g(x))^2\,\mathrm{d}x$ minimal genau dann, wenn $\int_I (P(x) - Q(x))^2\,\mathrm{d}x = 0$ gilt, das heißt wenn $P(x)$ gleich $Q(x)$ ist (das folgt aus der Stetigkeit der Funktion $(P(x) - Q(x))^2$: wäre $P(x) \neq Q(x)$ für ein x, dann gäbe es eine Umgebung U von x und ein $\varepsilon > 0$ mit $(P(y) - Q(y))^2 > \varepsilon$ für alle $y \in U$, woraus $\int_I (P(x) - Q(x))^2\,\mathrm{d}x > 0$ folgen würde). □

Der Fall, der uns hier interessiert, sind trigonometrische Polynome. Sei $n = 2m + 1$ für ein $m \ge 1$ und $I = [-\pi, \pi]$. Für $1 \le j \le m$ sei $\varphi_j(x) = \cos jx$ und $\varphi_{m+j}(x) = \sin jx$. Weiters sei $\varphi_0(x) = 1$. Wir erhalten dann $P(x) = a_0 + \sum_{j=1}^{m} a_j \cos jx + \sum_{j=1}^{m} a_{m+j} \sin jx$.

Wir berechnen die Matrix $R = (R_{uv})_{0 \le u,v \le 2m}$. Es gilt $\int_{-\pi}^{\pi} \cos jx\,\mathrm{d}x = \int_{-\pi}^{\pi} \sin jx\,\mathrm{d}x = 0$ für $j \ge 1$ und $\int_{-\pi}^{\pi} \mathrm{d}x = 2\pi$. Damit ist bereits $R_{0v} = R_{v0} = 0$ für $v \ge 1$ und $R_{00} = 2\pi$ gezeigt. Wir setzen $\delta_{uv} = 0$, wenn $u \neq v$, und $\delta_{uv} = 1$, wenn $u = v$. Wegen der Identitäten $\cos\alpha\cos\beta = \frac{1}{2}\cos(\alpha+\beta) + \frac{1}{2}\cos(\alpha-\beta)$, $\sin\alpha\sin\beta = \frac{1}{2}\cos(\alpha-\beta) - \frac{1}{2}\cos(\alpha+\beta)$ und $\sin\alpha\cos\beta = \frac{1}{2}\sin(\alpha+\beta) + \frac{1}{2}\sin(\alpha-\beta)$ erhalten wir

$$\begin{aligned}&\textstyle\int_{-\pi}^{\pi} \cos ux\cos vx\,\mathrm{d}x = \frac{1}{2}\int_{-\pi}^{\pi} \cos(u+v)x\,\mathrm{d}x + \frac{1}{2}\int_{-\pi}^{\pi} \cos(u-v)x\,\mathrm{d}x = \pi\delta_{uv} \quad \text{für} \quad u, v \ge 1 \\ &\textstyle\int_{-\pi}^{\pi} \sin ux\sin vx\,\mathrm{d}x = \frac{1}{2}\int_{-\pi}^{\pi} \cos(u-v)x\,\mathrm{d}x - \frac{1}{2}\int_{-\pi}^{\pi} \cos(u+v)x\,\mathrm{d}x = \pi\delta_{uv} \quad \text{für} \quad u, v \ge 1 \\ &\textstyle\int_{-\pi}^{\pi} \sin ux\cos vx\,\mathrm{d}x = \frac{1}{2}\int_{-\pi}^{\pi} \sin(u+v)x\,\mathrm{d}x + \frac{1}{2}\int_{-\pi}^{\pi} \sin(u-v)x\,\mathrm{d}x = 0 \quad \text{für} \quad u, v \ge 1\end{aligned}$$

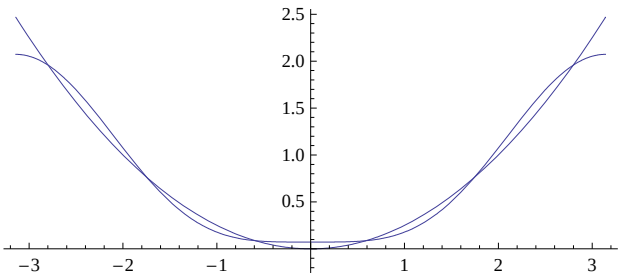
Somit gilt $R_{00} = 2\pi$, $R_{uu} = \pi$ für $u \ge 1$ und $R_{uv} = 0$ für $u \neq v$. Die Matrix R ist eine Diagonalmatrix und das Gleichungssystem $R\mathbf{a} = \mathbf{q}$ leicht lösbar. Wegen $q_0 = \int_{-\pi}^{\pi} g(x)\,\mathrm{d}x$ und da $q_j = \int_{-\pi}^{\pi} g(x)\cos jx\,\mathrm{d}x$ und $q_{m+j} = \int_{-\pi}^{\pi} g(x)\sin jx\,\mathrm{d}x$ für $1 \le j \le m$ gilt, haben wir

$$\begin{aligned}a_0 &= \tfrac{1}{2\pi}\textstyle\int_{-\pi}^{\pi} g(x)\,\mathrm{d}x \\ a_j &= \tfrac{1}{\pi}\textstyle\int_{-\pi}^{\pi} g(x)\cos jx\,\mathrm{d}x \quad \text{für} \quad 1 \le j \le m \\ b_j &= a_{m+j} = \tfrac{1}{\pi}\textstyle\int_{-\pi}^{\pi} g(x)\sin jx\,\mathrm{d}x \quad \text{für} \quad 1 \le j \le m\end{aligned}$$

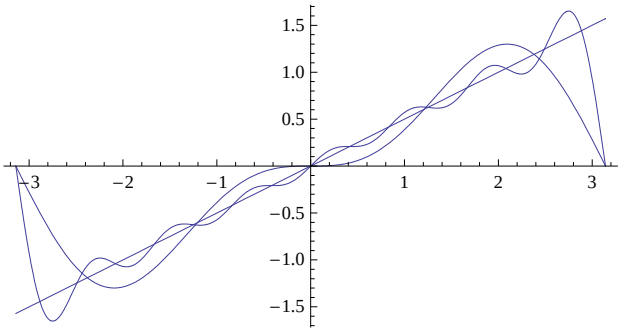
Somit ist das trigonometrische Polynom $P(x) = a_0 + \sum_{j=1}^{m} a_j \cos jx + \sum_{j=1}^{m} b_j \sin jx$ gefunden, das die Funktion g auf dem Intervall $[-\pi, \pi]$ am besten approximiert, wobei wir ab jetzt b_j statt a_{m+j} schreiben. Diese Approximation nennt man Fourierapproximation. Die Koeffizienten a_j für $j \ge 0$ und b_j für $j \ge 1$ heißen Fourierkoeffizienten der Funktion g.

Wir berechnen die Fourierapproximation für einige Funktionen $g : [-\pi, \pi] \to \mathbb{R}$. Ist g eine gerade Funktion, das heißt $g(-x) = g(x)$ für $x \in [-\pi, \pi]$, dann ist $x \mapsto g(x)\sin jx$ eine ungerade Funktion und es gilt $b_j = 0$ für alle $j \ge 1$. Ist g eine ungerade Funktion, das heißt $g(-x) = -g(x)$ für $x \in [-\pi, \pi]$, dann ist $x \mapsto g(x)\cos jx$ ebenfalls eine ungerade Funktion und es gilt $a_j = 0$ für alle $j \ge 0$.

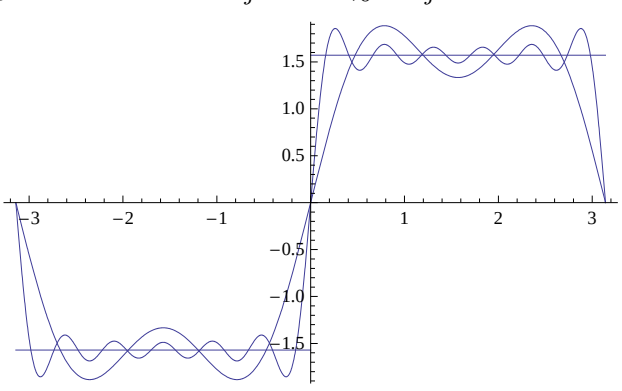
Beispiel: Sei $g : [-\pi, \pi] \to \mathbb{R}$ durch $g(x) = \frac{x^2}{4}$ definiert. Wir berechnen die Fourierkoeffizienten der Funktion g. Da g eine gerade Funktion ist, haben wir $b_j = 0$ für alle $j \geq 1$. Wegen $\int_{-\pi}^{\pi} \frac{x^2}{4} \, \mathrm{d}x = \frac{\pi^3}{6}$ gilt $a_0 = \frac{\pi^2}{12}$. Schließlich berechnen wir noch mit zweimaliger partieller Integration $\int_{-\pi}^{\pi} \frac{x^2}{4} \cos jx \, \mathrm{d}x = \frac{x^2}{4j} \sin jx \Big|_{-\pi}^{\pi} + \frac{x}{2j^2} \cos jx \Big|_{-\pi}^{\pi} - \frac{1}{2j^3} \sin jx \Big|_{-\pi}^{\pi} = \frac{2\pi(-1)^j}{2j^2}$, wobei verwendet wurde, dass $\sin j\pi = 0$ und $\cos j\pi = \cos(-j\pi) = (-1)^j$ für alle $j \in \mathbb{Z}$ gilt. Es folgt $a_j = \frac{(-1)^j}{j^2}$ für $j \geq 1$. Damit sind die Fourierkoeffizienten berechnet und die m-te Fourierapproximation $P_m(x) = a_0 + \sum_{j=1}^{m} a_j \cos jx + \sum_{j=1}^{m} b_j \sin jx = \frac{\pi^2}{12} + \sum_{j=1}^{m} \frac{(-1)^j}{j^2} \cos jx$ auf dem Intervall $[-\pi, \pi]$ für die Funktion g ist gefunden. In obiger Zeichnung sind die Funktionen g und P_2 gezeichnet. Man sieht, dass die zweite Fourierapproximation P_2 bereits recht gut approximiert.



Beispiel: Sei $g : [-\pi, \pi] \to \mathbb{R}$ durch $g(x) = \frac{x}{2}$ definiert. Da g ungerade ist, gilt $a_j = 0$ für alle $j \geq 0$. Wir berechnen $\int_{-\pi}^{\pi} \frac{x}{2} \sin jx \, \mathrm{d}x = -\frac{x}{2j} \cos jx \Big|_{-\pi}^{\pi} + \frac{1}{2j^2} \sin jx \Big|_{-\pi}^{\pi} = \frac{\pi(-1)^{j+1}}{j}$ mit partieller Integration, ganz analog wie im letzten Beispiel. Es ergibt sich $b_j = \frac{(-1)^{j+1}}{j}$ für $j \geq 1$. Damit sind die Fourierkoeffizienten berechnet und die m-te Fourierapproximation $P_m(x) = \sum_{j=1}^{m} \frac{(-1)^{j+1}}{j} \sin jx$ auf dem Intervall $[-\pi, \pi]$ für die Funktion g ist gefunden. In nebenstehender Zeichnung sind zusammen mit g die Funktionen P_2 und P_7 gezeichnet. Man sieht, dass die Approximation nicht so gut ist wie im letzten Beispiel.



Beispiel: Sei $g : [-\pi, \pi] \to \mathbb{R}$ durch $g(x) = \frac{\pi}{2} \operatorname{sign} x$ definiert, das heißt g ist gleich $-\frac{\pi}{2}$ auf $[-\pi, 0)$ und gleich $\frac{\pi}{2}$ auf $(0, \pi]$. Da g ungerade ist, haben wir $a_j = 0$ für alle $j \geq 0$. Wir berechnen $\int_{-\pi}^{\pi} \frac{\pi}{2} \operatorname{sign} x \sin jx \, \mathrm{d}x = 2 \int_{0}^{\pi} \frac{\pi}{2} \sin jx \, \mathrm{d}x = -\frac{\pi}{j} \cos jx \Big|_{0}^{\pi} = \frac{\pi}{j}(1 - (-1)^j)$. Es folgt $b_j = \frac{1}{j}(1 - (-1)^j)$ für $j \geq 1$, das heißt $b_j = 0$ für gerades $j \geq 1$ und $b_j = \frac{2}{j}$ für ungerades $j \geq 1$. Damit sind die Fourierkoeffizienten berechnet. Die $2k$-te Fourierapproximation $P_{2k}(x) = \sum_{l=1}^{k} \frac{2}{2l-1} \sin(2l-1)x$ auf dem Intervall $[-\pi, \pi]$ für die Funktion g ist gefunden. In nebenstehender Zeichnung sind zusammen mit g die Funktionen P_4 und P_{12} gezeichnet. Man sieht, dass die Approximation nicht so gut ist wie im vorletzten Beispiel.



Die Fourierapproximation $P_m(x) = a_0 + \sum_{j=1}^{m} a_j \cos jx + \sum_{j=1}^{m} b_j \sin jx$ einer Funktion g ist eine stetige Funktion auf $[-\pi, \pi]$, für die $P_m(-\pi) = P_m(\pi)$ gilt. Hat die Funktion g eine Sprungstelle, wie es in letzten Beispiel im Punkt 0 der Fall ist, dann kann dort die Approximation nicht gut sein. Dasselbe gilt in den Endpunkten des Intervalls $[-\pi, \pi]$, wenn $g(-\pi) \neq g(\pi)$ ist. Wegen $P_m(-\pi) = P_m(\pi)$ kann dann die Approximation in der Nähe der Endpunkte nicht gut sein. Das sieht man in den beiden letzten Beispielen.

Die Frage, ob $\lim_{m\to\infty} P_m(x) = g(x)$ gilt, hat keine einfache Antwort. Wir nennen die Funktion g im Punkt x linksmonoton, wenn ein $\delta > 0$ existiert, sodass g auf dem Intervall $(x - \delta, x)$ monoton ist. Wir nennen die Funktion g im Punkt x rechtsmonoton, wenn ein $\delta > 0$ existiert, sodass g auf dem Intervall $(x, x+\delta)$ monoton ist. Ist g im Punkt $x \in (-\pi, \pi)$ stetig und sowohl links- als auch rechtsmonoton, dann kann man $\lim_{m\to\infty} P_m(x) = g(x)$ zeigen. Ist g im Punkt $-\pi$ stetig und rechtsmonoton, im Punkt π stetig und linksmonoton und gilt $g(-\pi) = g(\pi)$, dann kann man $\lim_{m\to\infty} P_m(x) = g(x)$ auch für $x = -\pi$ und $x = \pi$ beweisen. Für eine Funktion g, die auf $[-\pi, \pi]$ stetig und in allen Punkten links- und rechtsmonoton ist und die $g(-\pi) = g(\pi)$ erfüllt, kann man sich daher gute Approximation erwarten. Die Funktion $g(x) = \frac{x^2}{4}$ im ersten der drei Beispiele ist so eine Funktion. Man sieht auch, dass bereits P_2 gut approximiert.

Ein Punkt x heißt Sprungstelle von g wenn $g(x-) = \lim_{y\uparrow x} g(y)$ und $g(x+) = \lim_{y\downarrow x} g(y)$ existieren, aber ungleich sind. Ist $x \in (-\pi, \pi)$ eine Sprungstelle von g und g in x sowohl links- als auch rechtsmonoton, dann kann man $\lim_{m\to\infty} P_m(x) = \frac{g(x-)+g(x+)}{2}$ zeigen. Ähnlichs gilt für $-\pi$ und π. Ist g im Punkt $-\pi$ stetig und rechtsmonoton und im Punkt π stetig und linksmonoton, dann gilt $\lim_{m\to\infty} P_m(-\pi) = \lim_{m\to\infty} P_m(\pi) = \frac{g(-\pi)+g(\pi)}{2}$.

Schließlich sei noch darauf hingewiesen, dass man die Fourierapproximation auch für Funktionen auf einem Intervall $[a, b]$ berechnen kann. Sei $\varphi(x) = \frac{b-a}{2\pi}x + \frac{a+b}{2}$. Das ist eine linare Abbildung, die das Intervall $[-\pi, \pi]$ auf das Intervall $[a, b]$ abbildet. Ist f auf $[a, b]$ definiert, dann bildet man $g(x) = f(\varphi(x))$. Das ist eine Funktion auf dem Intervall $[-\pi, \pi]$, für die man eine Fourierapproximation P_m berechnen kann. Eine Fourierapproximation für die ursprüngliche Funktion f ist dann $P_m(\varphi^{-1}(x))$.

Inhaltsverzeichnis