Skriptum zur Vorlesung

Praxisorientierte Einführung in die Numerik für Lehramtskandidaten

Sommersemester 2018

Martin Burger
Institut für Numerische und Angewandte Mathematik
Westfälische Wilhelms-Universität Münster
martin.burger@wwu.de

Basierend auf Teilen der Skripte von: Christian Schmeiser, Angewandte Mathematik für Lehramtskandidaten Frank Wübbeling, Numerik 1 und 2

Inhaltsverzeichnis

Kapitel 1

Einleitung

Das Ziel dieser Vorlesung ist es einen Überblick über verschiedene Aspekte der angewandten und numerischen Mathematik zu geben, die zum Teil auch relevant für den Schulunterricht sind. Unser vorrangiges Ziel ist es nicht uns auf den Schulstoff zu beschränken, wichtiger ist es grundlegende Ideen und systematische Herangehensweisen der angewandten Mathematik zu vermitteln. Besonders wichtige Aspekte sind dabei einerseits die saubere mathematische Modellierung realer Probleme, d.h. die Übersetzung einer Fragestellung aus der Praxis in Gleichungen oder andere berechenbare Modelle inklusive der verwendeten Modellannahmen, und andererseits die numerische Lösung der entstehenden Probleme. In letzterem Fall legen wir weniger Gewicht auf die eigentliche Programmierung sondern eher auf das Verständnis der grundlegenden Algorithmen.

Wir beginnen mit einem Beispiel zur Illustration: Drei Familien machen gemeinsam Urlaub und jede davon übernimmt gewisse Beträge. So kauft jede der Familien einmal für gemeinsames Essen im Supermarkt ein, Familie eins besorgt die Eintrittskarten für den gemeinsamen Besuch einer Attraktion und Famile zwei kauft einmal beim Metzger ein. Nun stellt sich am Ende des Urlaubs die Frage wie die Kosten gerecht aufzuteilen sind. Eine einfache ad-hoc Lösung wäre die Kosten zu addieren und zu dritteln. Jedoch stellt sich bei genauerer Betrachtung die Frage nach der Gerechtigkeit eines solchen Ansatzes. Implizit macht dieser die Annahme, dass alle Familien dann auch gleich an allen Ausgaben partizipieren, was jedoch nicht klar ist und eigentlich nicht ohne weitere Daten zu beantworten ist. Es stellt sich etwa die Frage ob alle Familien gleich groß sind. Wenn es eine verschiedene Anzahl von Kindern gibt ist zu klären ob die Preise der Tickets für Erwachsene und Kinder gleich sind. Und bezüglich des Metzger-Einkaufs könnte man die Frage stellen, ob es in den Familien Vegetarier gibt, die an dem Einkauf gar nicht partizipieren. Wir sehen also, dass für die Genauigkeit der Beantwortung einer einfachen Frage verschiedene Modelle und Modellparameter nötig sind. Im Wesentlichen sind bei der Modellierung und Lösung eines praktischen Problems folgende Schritte nötig:

1. Präzisierung und mathematische Formulierung der Fragestellung
2. Überprüfung der Vollständigkeit der Angaben /Daten, nötige Annahmen
3. Erstellung eines mathematischen Modells
4. (Numerische) Lösung des Modells, Berechnung eines Resultats

5. Formulierung des Resultats in der Sprache der Aufgabenstellung

Für das obige Beispiel bedeutet der erste Punkt die Frage was eine gerechte Aufteilung ist u.a. wie groß die einzelnen Familien sind, dies führt dann sofort auch auf die Frage nach Vollständigkeit der Daten wie Preise für Kindertickets. Dann werden immer noch Annahmen getroffen, auf die sich im Zweifel die Familien einigen müssen, etwas dass jeder ungefähr gleich viel von den Einkäufen profitiert oder man die Kinder weniger gewichtet. Darauf basierend kann nun das mathematische Modell für die Aufteilung formuliert und berechnet werden, was in diesem Fall vermutlich immer noch recht elementar ist. Zum Abschluss muss das Resultat noch formuliert werden, d.h. wieviel bezahlt jeder, eventuell beinhaltet dies auch Nachverarbeitungsschritte wie das Runden auf ganze Euro- oder Centbeträge.

Kapitel 2

Schulbeispiele

Im Folgenden diskutieren wir einige Beispiele aus dem gymnasialen Schulstoff und deren strukturierte Lösung.

Beispiel 2.1. Das Zugbeispiel: Um 15:00 fährt in Wien ein Zug Richtung Salzburg ab mit der Geschwindigkeit 80 km/h. Um 15:30 fährt im 300km entfernten Salzburg ein Zug Richtung Wien ab mit der Geschwindigkeit 100km/h. Wo begegnen die beiden Züge einander ?

Präzisierung als mathematische Fragestellung: Wir suchen den Abstand X von Wien, bei dem die beiden Züge einander begegnen.

Annahmen, Daten: Wir kennen die Zeiten, Abstände und Geschwindigkeiten, dies sollte, wir werden dies im Rahmen von Lösung 2 noch strukturierter sehen. Problematisch wären Zwischenhalte sowie Brems- und Beschleunigungsphasen, in denen von der Durschnittsgeschwindigkeit abgewichen wird. Wir nehmen also an, dass keine Zwischenhalte existieren.

Modell und Lösung 1: Um den Abstand X berechnen zu können, müssen wir wissen wann die beiden Züge aneinander vorbeifahren. Wir nennen diese Zeit T und da der zweite Zug um 15:30 abfährt und die 300 km in drei Stunden zurücklegt wissen wir $T \in [15.5, 18.5]$. Es gilt dann für die Abstände gleich Geschwindigkeit mal Zeit seit Abfahrt, d.h.

$$X = 80(T - 15), \qquad 300 - X = 100(T - 15.5).$$

Damit haben wir ein lineares Gleichungssystem für zwei Unbekannte T und X, das wir z.b. mit Eliminationsmethode lösen können. Dies ergibt $T \approx 156$.

Modell und Lösung 2: Wir legen Wien (W) und Salzburg (S) auf eine Koordinatenachse und berechnen der Ort $x_1(t)$ bzw. $x_2(t)$ der beiden Züge zur Zeit t. Es gilt

$$x_1(t) = W + v_1(t - t_1), \qquad x_2(t) = S - v_2(t - t_2),$$

wobei v_1, v_2 die Geschwindigkeiten und t_1, t_2 die Anfahrtszeiten sind. Wir sehen nun welche Daten wir brauchen, nämlich W, S, v_1, v_2, t_1, t_2. da wir das Koordinatensystem beliebig legen können und nur die relative Entfernung zu Wien berechnen wollen, kann $W = 0$ gesetzt werden und damit ist $S = 300$ bekannt. Zur Lösung suchen wir nun X und T mit

$$x_1(T) = x_2(T) = X,$$

dies können wir wie oben lesen.

Formulierung des Resultats: Die Züge treffen ca. 156 km von Wien entfernt aufeinander, für den Weichensteller sollte man dann auf der Karte nachschlagen wo dies genau passiert.

Beispiel 2.2. Das Badewannenbeispiel: Fünf Minuten lang läuft Wasser mit 15 Grad Celsius in die Badewanne, dann zehn Minuten lang mit 50 Grad Celsius. Welche Temperatur hat das Badewasser am Ende ?
Präzisierung als mathematische Fragestellung: Wir suchen die Temperatur T des Badewassers nach 15 Minuten.
Annahmen, Daten: Wir kennen die Zeiten und Temperaturen, dies scheint bei geeigneter Vereinfachung zu genügen. Bei genauerer Betrachtung müssen wir auch die Durchflussgeschwindigkeiten (gleich bei Heiss- und Kaltwasser ?) und auch die mögliche Wärmeabgabe an die Luft innerhalb der 15 Minuten betrachten. Wir entscheiden uns dafür diese Effekte zu ignorieren.
Modell und Lösung 1: Als einfaches Modell nehmen wir an, dass bei gleichen Durchflussgeschwindigkeiten die selbe Menge an Wasser in 15 Minuten mit der Temperatur geflossen wäre, wenn wir den Wasserhahn gleich darauf eingestellt haben. Es gilt also

$$15 * T = 5 * 15 + 10 * 50, \qquad \text{also} \qquad T = 38,3.$$

Modell und Lösung 2: Wir benutzen die Notation t_1, t_2 für die beiden Zeiten (5 und 10 Minuten) und T_1, T_2 für die beiden Temperaturen. Das grundlegende physikalische Prinzip, dass wir benutzen wollen ist die Erhaltung der Energie, diese ist proportional zu Masse M und absoluter Temperatur $\theta = \theta 0 + T$, wobei $\theta_0 = 273,15$. Dies ergibt

$$E = \kappa(M_1 + M_2)(\theta_0 + T) = \kappa M_1(\theta_0 + T_1) + \kappa M_2(\theta_0 + T_2)$$

mit einer Proportionalitätskonstante κ. Die eingelassene Masse ergibt sich aus dem Produkt von Flussgeschwindigkeit F (Masse pro Zeit) und Zeit. Damit ist $M_1 = Ft_1$, $M_2 = Ft_2$. Setzen wir dies oben ein und formen um, sehen wir, dass es nicht nötig ist F, κ und θ_0 zu kennen, wir erhalten wieder

$$T = \frac{t_1T_1 + t_2T_2}{t_1 + t_2}.$$

Formulierung des Resultats: Die Gesamttemperatur erhalten wir als gewichteten Mittelwert der beiden Einlasstemperaturen mit den jeweiligen (relativen) Zeiten, also in diesem Fall $38,3$ Grad Celsius.

Beispiel 2.3. Geburtstagswahrscheinlichkeit: Wir suchen die Wahrscheinlichkeit, dass wir bei einer Party mit 30 Teilnehmern zwei davon finden, die am selben Tag Geburtstag haben.
Präzisierung als mathematische Fragestellung: Eine mathematisch präzisere Formulierung ist die Frage mit welcher Wahrscheinlichkeit p bei einer Stichprobe von 30 Leuten, die wir als unabhängig betrachten, mindestens zwei am gleichen Tag Geburtstag haben.

Annahmen, Daten: Nun sollten wir die Wahrscheinlichkeit für einzelne Tage kennen, eine einfache Annahme ist $\frac{1}{365}$, die nicht ganz stimmt aber eine gute Approximation liefert. Schaltjahre, in denen die Wahrscheinlichkeiten zu modifizieren sind, wollen wir ebenfalls ignorieren für eine erste Approximation.
Modell und Lösung : Wie oft in der Wahrscheinlichkeitsrechnung ist es leichter, die Gegenwahrscheinlichkeit $1 - p$ zu berechnen, die die Wahrscheinlichkeit dafür beschreibt, dass alle 30 Teilnehmer an verschiedenen Tagen Geburtstag haben. Dafür können wir, da alle Geburtstage die gleiche Wahrscheinlichkeit einfach die Anzahl der günstigen durch die Anzahl der möglichen Fälle rechnnen. Die Anzahl der möglichen ist 356^30, die günstigen können wir folgendermaßen berechnen: Nummerieren wir die 30 Teilnehmer durch, so hat der erste zunächst

365 Möglichkeiten, der zweite jeweils nur 364 günstige (der Geburtstag des ersten abgezogen), für den dritten 363 (ohne die Geburtstage der ersten beiden) und so weiter. Damit gilt

$$1 - p = \frac{365 * 364 * \ldots * 336}{365 * 365 * \ldots * 365} \approx 0,694.$$

Kapitel 3

Verzinsung und Kredite

Im Folgenden betrachten wir einige Aspekte von Verzinsung als dynamischen Prozess. Wir beginnen unsere Beschreibung in einer Einheit von Jahren mit einer m-maligen Verzinsung pro Jahr (also jährliche Zinsen $m = 1$, monatliche Zinsen $m = 12$ etc.). Die jährliche Zinsrate bezeichnen wir mit r, bei jeder einzelnen Verzinsung wird dementsprechend die Rate $\frac{r}{m}$ verwendet. Damit ist die kleinste Zeiteinheit die wir betrachten $\frac{1}{m}$. Der Wert des Kapitals $P_m(t)$ zur Zeit t bei m-maliger Verzinsung und fixer (jährlicher) Einzahlung x_m kann am einfachsten durch die Änderung in diesem Zeitschritt beschrieben werden. Es gilt

$$P_m(t + \frac{1}{m}) = (1 + \frac{r}{m})P_m(t) + \frac{1}{m}x_m,$$

d.h. das Kapital wird verzinst (auf das $1 + \frac{r}{m}$-fache) und dann wird die neue Einzahlung addiert. Analog gilt bei Schulden P_m

$$P_m(t + \frac{1}{m}) = (1 + \frac{r}{m})P_m(t) - \frac{1}{m}x_m,$$

da die offene Restschuld durch eine Rückzahlung x_m ja reduziert wird.

Die obigen Gleichungen für den Wert des Kapitals bzw. der Schuld können wir nun in einem zweiten Schritt lösen, entweder numerisch (durch ein Computerprogramm, dass genau die obigen Formeln implementiert) oder explizit, was wir im Schuldenfall nun tun würden. Wir beginnen mit dem homogenen Fall, d.h. $x_m = 0$, hier haben wir

$$P_m(t + \frac{1}{m}) = (1 + \frac{r}{m})P_m(t) = (1 + \frac{r}{m})^2 P_m(t - \frac{1}{m}) = (1 + \frac{r}{m})^3 P_m(t - \frac{3}{m}) = \dots.$$

Damit erhalten wir am Ende die homogene Lösung

$$P_m(t) = (1 + \frac{r}{m})^{mt} P_m(0).$$

Zur Lösung im Fall $x_m \neq 0$ verwendet man die Methode der Variation der Konstanten, d.h. den Ansatz

$$P_m(t) = (1 + \frac{r}{m})^{mt} Q_m(t).$$

Der Name kommt daher, dass bei der homogenen Gleichung $Q_m(t)$ ja konstant wäre, jetzt aber variiert wird. Setzen wir diesen Ansatz ein, so sehen wir, dass

$$Q_m(t + \frac{1}{m}) = Q_m(t) - \frac{1}{m}(1 + \frac{r}{m})^{-mt-1} x_m$$

gilt. Nun erhalten wir

$$Q_m(t) = Q_m(t-\frac{1}{m}) - \frac{1}{m}(1+\frac{r}{m})^{-mt}x_m = Q_m(t-\frac{2}{m}) - \frac{1}{m}(1+\frac{r}{m})^{-(mt-1)}x_m - \frac{1}{m}(1+\frac{r}{m})^{-mt}x_m = \dots.$$

Daraus erhalten wir

$$Q_m(t) = Q_m(0) - \frac{1}{m}x_m \sum_{k=1}^{mt}(1+\frac{r}{m})^{-k}$$

und somit (da $P_m(0) = Q_m(0)$)

$$P_m(t) = (1+\frac{r}{m})^{mt}P_m(0) - \frac{1}{m}x_m \sum_{k=1}^{mt}(1+\frac{r}{m})^{mt-k}.$$

Nun benennen wir noch den Index in der letzten Summe um zu $j = mt - k$, d.h.

$$P_m(t) = (1+\frac{r}{m})^{mt}P_m(0) - \frac{1}{m}x_m \sum_{j=0}^{mt-1}(1+\frac{r}{m})^{j}$$

und erhalten mit der Formel für die geometrische Reihe sowie dem Anfanswert der Schulden P^0

$$P_m(t) = (1+\frac{r}{m})^{mt}P^0 - x_m\frac{1}{r}\left((1+\frac{r}{m})^{mt} - 1\right).$$

3.1 Kreditrückzahlung

Wird wenden uns nun einer typischen Frage aus der Praxis zu: Wir wollen einen Kredit mit dem Umfang P^0 aufnehmen und ihn in der Zeit T zurückzahlen, wobei eine Zahlung und Verzinsung m-mal im Jahr mit fixer Zinsrate und fixem Rückzahlungsbetrag x_m stattfinden. Nun stellt sich die Frage wie groß der Betrag x_m sein muss, damit wir die Rückzahlung tatsächlich bis zur Zeit T erledigt haben. Eine Fragestellung dieser Art bezeichnet man in der angewandten Mathematik als inverses Problem, da man die leichter zu modellierende Fragestellung umkehrt. Anstatt wie oben bei gegebener Rückzahlung die Restschuld zur Zeit T zu berechnen wollen wir nun aus gegebener Restschuld (in diesem Fall gleich Null) die Rückzahlung bestimmen. Dementsprechend ist es auch leichter sich nicht direkt eine Lösung für das inverse Problem zu versuchen sondern erst einmal das Vorwärtsproblem abhängig von x_m zu formulieren wie wir es oben getan haben. Aus der expliziten Lösungsformel erhalten wir mit $P_m(T) = 0$ direkt

$$x_m = r\frac{(1+\frac{r}{m})^{mT}}{(1+\frac{r}{m})^{mT} - 1}P^0.$$

Bei komplizierteren Modellen ist es oft nicht möglich eine explizite Lösungsformel anzugeben und diese dann auch für das inverse Problem zu verwenden. Dann muss man eine numerische Lösungsmethode einsetzen, die wir später in der Vorlesung noch genauer diskutieren werden. Wir beachten, dass auch ohne Lösungsschritte unsere Modellierung folgendes Gleichungssystem liefert:

$$\begin{aligned} P_m(0) &= P^0 \\ P_m(t+\frac{1}{m}) &= (1+\frac{r}{m})P_m(t) - \frac{1}{m}x_m, \qquad t = 0, \dots T - \frac{1}{m} \\ P_m(T) &= 0. \end{aligned}$$

Dies sind $mT + 2$ lineare Gleichungen für die $mT + 2$ Unbekannten $P_m(\frac{k}{m})$, $k = 0, \ldots, mT$ und x_m. Mit einem geeigneten Lösungsverfahren für lineare Gleichungssysteme können wir also auch daraus die gesuchte Rückzahlung x_m bestimmen.

3.2 Kontinuierliche Verzinsung

Ein interessanter Grenzwert der Verzinsung erhalten wir bei sehr kleinen Intervallen, d.h. m sehr groß. In diesem Fall können wir den Grenzwert zu einer kontinuierlichen Verzinsung betrachten. Dazu stellen wir uns vor wir erweitern P_m von den diskreten Zeitschritten $\frac{k}{m}$ zu einer Funktion auf dem Intervall $[0, T]$, etwa indem wir den Wert zwischen den Zeitschritten linear interpolieren, d.h.

$$P_m(t) = m(t - \frac{k}{m})P_m(\frac{k+1}{m}) + m(\frac{k+1}{m} - t)P_m(\frac{k}{m}), \qquad t \in [\frac{k}{m}, \frac{k+1}{m}].$$

Damit haben wir ein gemeinsames Intervall auf dem alle Funktionen P_m definiert sind und können den Grenzwert der Funktionenfolge für $m \to \infty$ betrachten. Wir nehmen an, dass es einen Grenzwert P_∞ gibt und wollen eine Gleichung für P_∞ herleiten. Dazu schreiben wir die Rekursion für P_m als

$$\frac{P_m(t + \Delta t) - P_m(t)}{\Delta t} = rP_m(t) - x_m$$

mit $\Delta t = \frac{1}{m}$. Wir erwarten, dass der Grenzwert P_∞ diese Gleichung mit einer Grenzeinzahlung x_∞ bis auf einen kleinen Rest erfüllt. Nun haben wir es mit dem Grenzwert $\Delta t \to 0$ zu tun, in dem die linke Seite gegen eine Ableitung konvergiert, d.h. die Gleichung für P_∞ ist $P'_\infty(t) = rP_\infty(t) - x_\infty$. Dies ist eine gewöhnliche Differentialgleichung, da neben der Funktion auch ihre Ableitung auftritt (eine eindimensionale gewöhnliche Ableitung im Gegensatz zu partiellen Ableitungen im Mehrdimensionalen).

Gewöhnliche Differentialgleichungen können ähnlich gelöst werden wie die Rekursion oben. Betrachten wir zunächst die homogene Gleichung

$$P' = rP$$

, so sehen wir sofort, dass die Lösung durch $P(t) = P(0)e^{rt}$ gegeben ist. Für die inhomogene Gleichung benutzen wir wieder Variation der Konstanten mit dem Ansatz $P(t) = Q(t)e^{rt}$. Setzen wir dies in die Gleichung

$$P' = rP + x$$

ein so folgt

$$Q' = e^{-rt}x$$

und aus Integration mit $Q(0) = P(0) = P^0$

$$Q(t) = P^0 + \int_0^t e^{-rs}x \; ds = P^0 + \frac{1}{r}(e^{-rs} - 1)x.$$

Damit ist die Lösung P gegeben durch

$$P(t) = e^{rt}P^0 - (e^{rt} - 1)x.$$

Kapitel 4

Bilder und Tomographie

Im Folgenden wollen wir eine kurze Einführung in die mathematische Behandlung von Bildern und ihre Rekonstruktion aus indirekten Messungen, wie man es in der Computertomographie findet, geben. Dazu beginnen wir zunächst mit der Frage was ein mathematisches Bild, oder zunächst ein Signal überhaupt ist.

Die Signalverarbeitung ist eines der ältesten Gebiete der angewandten Mathematik, klassischerweise unterscheidet man zwischem einem kontinuierlichen (analogen) Signal $s : [0, T] \to \mathbb{R}$ und einem diskreten (digitalen) Signal $S^N \in \mathbb{R}^N$. Der Zusammenhang zwischen einem kontinuierlichen und einem diskreten Signal ist die Abtastung an verschiedenen Zeitpunkten $t_n \in [0, T]$, $n = 1, \ldots, N$. Wir erhalten dann als n-ten Eintrag $S^N_n = s(t_n)$.

Eine kanonische Klasse von Signalen die man dabei im Kopf hat sind Audiosignale, die typischerweise in Wellenform auftreten. Hier ist s die Intensität (Lautstärke) und die Wellenform von s in der Zeit bestimmt die Intensität. Als klassisches Beispiel denken wir an Wellen der Form $s(t) = \sin(\pi k \frac{t}{T})$ mit $k \in \mathbb{N}$. Hier bestimmt k die Frequenz, bei einer akustischen Welle wächst die Tonhöhe mit k. Wir sehen dabei auch, dass die Abtastung mit der Wellenlänge in Zusammenhang steht. Ist N nicht deutlich grösser als k, so kann das Signal nicht richtig wiedergegeben werden. Wir betrachten dazu das Beispiel $N = k$ und $T = 1$ mit den kanonischen Stützstellen $t_n = \frac{n}{N}$. In diesem Fall ist $s(t_n) = \sin(\pi N \frac{n}{N}) = 0$, d.h. statt eines Wellensignals tasten wir immer die Nullstellen ab.

Bei einem Bild haben wir es im wesentlichen mit einem zweidimensionalen Signal zu tun, also einer Funktion $I : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ (typischerweise ist Ω ein Rechteck) wobei $I(x)$ die Helligkeit (Grauwert) im Punkt x beschreibt. Ein Farbbild ist dann einfach eine Kombination von drei solchen Bildern für die Helligkeit verschiedener Farbanteile, zum Beispiel rot, grün und blau. Ein wichtiger Unterschied von Bildern zu Audiosignalen ist, dass hier nicht Wellenanteile sondern vor allem Kanten besonders wichtig sind. Die Schärfe in einem Bild wird meist nach der Qualität der Kanten, d.h. der Unstetigkeitsstellen in einem Bild, bestimmt. Dabei sind besonders größere Übergänge zwischen verschiedenen Grauwerten interessant, während sehr kleine Bereiche mt abweichendem Grauwert eher als störendes Rauschen wahrgenommen werden. Die Entrauschung ist eines der Grundprobleme der Signal- und Bildverarbeitung, hier versucht man das Rauschen, meist zufällige Störungen in einzelnen Bereichen zu entfernen, etwa durch Mittelung über benachbarte Bildpunkte. Weitere typische Aufgaben der Bildverarbeitung sind die Bildrestaurierung (das Einfüllen benachbarter Information in gestörte Bereiche), die Segmentierung (das Auffinden von Kanten oder Objekten in Bildern) oder die Bewegungsschätzung in Folgen von Bildern (Videos).

Wie im Signalfall können wir auch ein diskretes Bildmodell verwenden, das mit dem kontinuierlichen zusammenhängt. Üblicherweise geht man hier nicht von einer Abtastung, sondern eher von einer Mittelung ab. Dazu wird das Gebiet Ω in kleine (disjunkte) Rechtecke P_{jk} zerteilt, die sogenannten Pixel, $\Omega = \cup_{j,k} P_{j,k}$. Hier ist j der Index zur Zählung in x-Richtung und k in y-Richtung. Der diskrete Wert I_{jk} ist dann der Mittelwert im Pixel, d.h.

$$I_{jk} = \frac{\int_{P_{jk}} I(x)\ dx}{\int_{P_{jk}} 1\ dx}.$$

Dies ist auch ein sinnvolles physikalisches Modell für die Aufnahme in Digitalkameras (CCD), die genau aus einem Gitter aus Detektoren (entsprechend den Pixeln) bestehen, dass jeweils den Mittelwert an Helligkeit in der Detektorfläche registrieren kann (bzw. die Photonen zählt die innerhalb des Detektors ankommen). Dementsprechend kann ein diskretes Bild einfach als Matrix interpretiert werden. Umgekehrt kann aus einem diskreten Bild auch ein kontinuierliches erstellt werden, in dem man einfach eine stückweise konstante Funktion konstruiert, die in P_{jk} den Wert I_{jk} annimmt. Ein besonderer Vorteil der Möglichkeit zwischen diskret und kontinuierlich hin und her zu transformieren ist damit auch die Möglichkeit Bilder mit verschiedener Auflösung, d.h. verschiedener Anzahl an Pixel zu vergleichen. Dazu kann man einfach beide auf das kontinuierliche Bild transformieren, oder eines davon zunächst zu einem kontinuierlichen und dann wieder zur Auflösung des anderen. Algorithmen für Bildverarbeitungsaufgaben funktionieren üblicherweise dann robust bei sich ändernder Auflösung wenn sie auch ein vernünftiges Äquivalent auf kontinuierlichen Bildern haben.

Ein elementares Problem im Zusammenhang mit Bildern ist die Rekonstruktion von Bildern aus indirekten Daten. Dies ist besonders wichtig in der modernen medizinischen Bildgebung, in der man Bilder aus dem Körperinneren gewinnen möchte ohne ihn zu öffnen. Das klassische Beispiel dafür ist die Computertomographie, die wir nun mathematisch in einem einfachen zweidimensionalen Fall modellieren wollen. Dabei ist das Bild $I(x, y)$ die zweidimensionale Dichte im Körper und wir schicken entlang verschiedener Richtungen Röntgenstrahlen mit Ausgangsintensität R_0 durch den Körper, auf der gegenüberliegenden Seite messen wir deren Abschwächung. Um dies zu modellieren betrachten wir zunächst nur Strahlen parallel zur x-Achse, mit einem Koordinatensystem so, dass die Emission bei $x = 0$ und die Detektion bei $x = L$ passiert. Also ist $R(0, y) = R_0$ und $R(L, y)$ sind die gemessenen Daten. Auch hier haben wir wieder ein inverses Problem, leichter ist es bei gegebenem Bild I das Entstehen von $R(L, y)$ zu modellieren. Dazu betrachten wir wieder kleine Änderungen der Position von x zu $x + \Delta x$. Die Abschwächung eines Röntgenstrahls ist proportional zu seiner eigenen Intensität R und der Menge an Material auf dem kleinen Weg, also

$$R(x + \Delta x, y) - R(x, y) \approx -R(x, y)I(x, y)\Delta x.$$

Im Grenzwert $\Delta x \to 0$ erhalten wir wieder eine Differentialgleichung

$$\partial_x R(x, y) = -R(x, y)I(x, y).$$

Hier haben wir zwar eine partielle Ableitung, aber die zweite Variable y spielt eigentlich keine besondere Rolle, wir erhalten für jeden Wert von y eine gewöhnliche Differentialgleichung, mit Kettenregel

$$-\partial_x(\log R(x, y)) = I(x, y)$$

Integrieren wir dies von 0 bis L so folgt

$$-\log R(L,y) + \log R_0 = \int_0^L I(x,y)\ dx.$$

Da wir die linke Seite aus den Messungen für das inverse Problem berechnen können, kennen wir also alle Linienintegrale in x-Richtung des Bildes x. Führt man nun auch noch eine Drehung ein, d.h. misst man die Schwächung in anderen Richtungen, so hat man am Ende das inverse Problem die Funktion I aus Linienintegralen in alle verschiedenen Richtungen zu berechnen.

Zur numerischen Lösung kann man das Problem wieder diskretisieren, setzt man das diskrete Bildmodell als stückweise konstante Funktion ein, so ist das Linienintegral eine Summe übe die Bildwerte in den Pixel multipliziert mit der Länge die die jeweilige Linie durch das Pixel läuft. Damit erhält man ein großes lineares Gleichungssystem für die Werte des diskreten Bildes, das man wieder mit passenden Algorithmen lösen kann.

Kapitel 5

Populationsdynamik

Populationsdynamik, d.h. die Beschreibung der Entwicklung der Anzahl von Mitgliedern verschiedener Arten, ist ein klassisches Anwendungsgebiet der Mathematik in der Biologie. Wir beginnen hier mit einem sehr einfachen Fall in dem nur eine Spezies existiert, die sich in Generationen fortpflanzt, deren Überlappung vernachlässigbar ist. In einem solchen Fall können wir eine zeitdiskrete Beschreibung verwenden, d.h. wir leiten eine Gleichung für N_k, die Anzahl an Individuen in der k-ten Generation her.

Das einfachste Modell ist von der Form

$$N_{k+1} = f(N_k),$$

d.h. die Anzahl der nächsten Generation ergibt sich durch direkte Fortpflanzung aus der vorherigen mit einer linearen oder nichtlinearen Funktion f. Das einfachste Beispiel ist wieder eine konstante Geburtenrate r, d.h. $f(N_k) = rN_k$. Wir beachten, dass wir hier auch r als nichtnatürliche Zahl zulassen können, was dann auch auf relle Werte für N_k führt. Dies ist zunächst erstaunlich, wird aber klarer wenn wir den Prozess genauer als zufällig modellieren und N_k als die durschnittliche Anzahl interpretieren, die sich dabei ergibt, ein Durchschnitt natürlicher Zahlen muss natürlich nicht mehr natürlich sein. Die Rekursion ist die gleiche wie im Fall der Verzinsung, wir erhalten $N_k = r^k N_0$ und sehen ein Verhalten abhängig von r. Wenn $r < 1$ ist, stirbt die Population langsam aus, wenn $r > 1$ wächst sie exponentiell. Dies kann für sehr große Zeit kein realistisches Modell sein, da eine zu große Population auch extrem große Resourcen brauchen würde, also muss man hier eine nichtlineare Modifizierung durchführen.

5.1 Die Fibonacci-Folge

Wir diskutieren nun ein einfaches historisches Beispiel aus dem Jahre 1202. Damals veröffentlichte Leonardo von Pisa, genannt Fibonacci, ein Modell für die Fortpflanzung von Kaninchen. Die Annahme dabei war, dass jedes Paar von Kaninchen nach einem Monat geschlechtsreif ist und dann nach einem weiteren Monat ein neues Kaninchenpaar zur Welt bringt, das später auch geschlechtsreif wird etc. Sei nun k die Anzahl der Monate, J_k die Anzahl an noch nicht geschlechtsreifen Paaren von Jungkaninchen nach k Monaten, und R_k die Anzahl geschlechtsreifer Paare. Dann gelten nach diesem Modell die Rekursionen

$$R_{k+1} = R_k + J_k, \qquad J_{k+1} = R_k.$$

Für die Gesamtanzahl $R_k + J_k$ erhält man daraus durch Einsetzen

$$N_{k+1} = N_k + N_{k-1},$$

eine zweistufige Rekursion. Beginnt man mit einem Paar Neugeborener, d.h. $J_0 = 1$ und $R_0 = 0$, damit $J_1 = 0$ und $R_1 = 1$, so ergibt sich $N_0 = N_1 = 1$ und daraus für N_k die bekannte Fibonacci-Folge $1, 1, 2, 3, 5, 8, 13, \ldots$.

Wir wollen nun wieder eine explizite Formel für N_k herleiten. Wir sehen, dass die Lösung durch die beiden Anfangswerte N_0 und N_1 eindeutig festgelegt ist, deshalb muss der gesamte Lösungsraum zweidimensional sein. Sobald wir also zwei linear unabhängige Lösungen gefunden haben können wir jede Lösung als Linearkombination dieser beiden darstellen. Um eine solche Lösung zu finden, machen wir ähnlich zur einstufigen Rekursion einen Ansatz $N_k = \lambda^k$, d.h.

$$\lambda^{k+1} = \lambda^k + \lambda^{k-1}.$$

Da wir nur an nichttrivialen Lösungen interessiert sind, können wir $\lambda = 0$ ausschliessen und durch λ^{k-1} teilen. Lösungen der obigen Form erhalten wir also genau aus den beiden Nullstellen λ_1 und λ_2 der quadratischen Gleichung

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

Diese sind gegeben durch

$$\lambda_1 = \frac{1+\sqrt{5}}{2}, \quad \lambda_2 = \frac{1-\sqrt{5}}{2}.$$

Damit können wir eine Lösung der Rekursion als

$$N_k = c_1\lambda_1^k + c_2\lambda_2^k$$

bestimmen, mit geeigneten Konstanten c_1 und c_2. Diese bestimmen wir aus den Anfangswerten, es gilt

$$N_0 = 1 = c_1 + c_2, \quad N_1 = 1 = c_1\lambda_1 + c_2\lambda_2$$

und damit $c_1 = c_2 = \frac{1}{2}$. Die Lösung kann dann als

$$N_k = \frac{(1+\sqrt{5})^k}{2^{k+1}} + \frac{(1-\sqrt{5})^k}{2^{k+1}}$$

geschrieben werden.

5.2 Die logistische Abbildung

Die Fibonacci Folge war ein Beispiel für ein lineares dynamisches System, wir diskutieren nun noch ein kanonisches Beispiel für ein nichtlineares System. Dazu betrachten wir ein einfaches Populationsmodell mit einfacher Fortpflanzung und Geburtenrate r, das wir zunächst als

$$N_{k+1} = N_k + rN_k$$

schreiben würden. Solange N_k klein ist scheint dies vernünftig zu sein, wird die Population aber zu groß so muss man auch die limitierten Resourcen betrachten. Ein Ökosystem kann

nicht ein beliebige Anzahl an Individuen ernähren, sodass das Wachstum mit steigender Anzahl an Individuen verlangsamt wird. Ist $N_{\max}$ die maximale Anzahl, die das System ernähren kann, so modifiziert man den Wachstumsterm bei der logistischen Abbildung zu

$$N_{k+1} = N_k + rN_k(1 - \frac{N_k}{N_{\max}}).$$

Skalieren wir zu $u_k = \frac{N_k}{N_{\max}}$, so erhalten wir die vereinfachte Rekursion

$$u_{k+1} = u_k + ru_k(1 - u_k).$$

Durch den letzten quadratischen Term in u_k haben wir in diesem Fall also ein nichtlineares dynamisches System. Nehmen wir an, dass die Rekursion konvergiert, also $u_k \to \overline{u}$ für $k \to \infty$, so erhalten wir im Grenzwert

$$\overline{u} = \overline{u} + r\overline{u}(1 - \overline{u}).$$

Eine Lösung dieser Gleichung nennt man stationäre Lösung, man sieht dass der Anfangswert $u_0 = \overline{u}$ zu einer konstanten Dynamik in der Zeit führt, $u_k = \overline{u}$ für alle k. In diesem Fall haben wir zwei stationäre Lösungen $\overline{u} = 0$ und $\overline{u} = 1$. Es stellt sich also die Frage welche davon nach langer Zeit angenähert wird oder nicht bzw. was bei kleinen Störungen von $\overline{u}$ passiert. Dies führt auf die Begriffe von Stabilität und Instabilität, die wir im Folgenden diskutieren werden.

5.3 Stabilität und Instabilität von stationären Lösungen

Im folgenden betrachten wir wieder ein kanonisches dynamisches System

$$u_{k+1} = f(u_k).$$

Wir nennen $\overline{u}$ eine stationäre Lösung, falls $\overline{u} = f(\overline{u})$. Nun fragen wir uns ob die Dynamik mit einem Anfangswert nahe $\overline{u}$ dann auch in der Nähe der stationären Lösung bleibt oder diese sogar weiter annähert. Dies machen wir in der folgenden Definition mathematisch exakt:

Definition 5.1. Ein stationärer Punkt $\overline{u}$ eines dynamischen Systems heisst **stabil**, wenn für jedes $\epsilon > 0$ ein $\delta > 0$ existiert, sodass für alle Anfangswerte u_0 mit $|u_0 - \overline{u}| < delta$ gilt, dass $|u_k - \overline{u}| < \epsilon$ für alle k. Ist dies nicht der Fall, so heisst $\overline{u}$ **instabil**.

Definition 5.2. Ein stationärer Punkt $\overline{u}$ eines dynamischen Systems heisst **asymptotisch stabil**, wenn ein $\delta > 0$ existiert, sodass für alle Anfangswerte u_0 mit $|u_0 - \overline{u}| < delta$ gilt, dass $u_k \to \overline{u}$ für $k \to \infty$.

Asymptotische Stabilität können wir zum Beispiel aus den Bedingungen des Banach'schen Fixpunktsatzes erhalten. Wir nehmen an $f : \mathbb{R} \to \mathbb{R}$ ist kontraktiv, d.h. es existiert ein $\eta < 1$, sodass

$$|f(u) - f(v)| \leq \eta|u - v|, \qquad \forall\ u, v \in \mathbb{R}$$

gilt. Dann rechnen wir wegen $\overline{u} = f(\overline{u})$ leicht nach, dass

$$|u_{k+1} - \overline{u}| = |f(u_k) - f(\overline{u})| \leq \eta|u_k - \overline{u}|$$

gilt. Durch Rückeinsetzen bis $k = 0$ folgt dann

$$0 \leq |u_k - \overline{u}| \leq \eta^k |u_0 - \overline{u}|.$$

Da die rechte Seite auch gegen Null konvergiert, folgt $u_k \to \overline{u}$.

Kontraktivität ist eine globale Bedingung und wie wir aus dem obigen Argument sehen konvergiert die Dynamik auch global, d.h. für jeden Anfangswert u_0 gegen $\overline{u}$. Im Fall einer differenzierbaren Funktion bedeutet die Kontraktivität, dass der Absolutwert der Ableitung $|f'(u)|$ überall kleiner als eins ist. Ein natürlicher Schritt zu einer lokalen Bedingung ist, dass man $|f'(u)| < 1$ nur in einer Umgebung von $\overline{u}$ fordert. Tatsächlich genügt schon $|f'(\overline{u}| < 1$, wie das folgende Resultat zeigt:

Satz 5.3. *Sei $f : \mathbb{R} \to \mathbb{R}$ zweimal stetig differenzierbar und $\overline{u} = f(\overline{u})$ ein stationärer Punkt, der $|f'(\overline{u}| < 1$ erfüllt. Dann ist $\overline{u}$ asymptotisch stabil. Ist $\overline{u}$ ein stationärer Punkt mit $|f'(\overline{u}| > 1$, dann ist $\overline{u}$ instabil.*

Beweis. Wir beginnen im ersten Fall $|f'(\overline{u}| < 1$ und führen zunächst eine Taylorentwicklung um den Punkt $\overline{u}$ durch. Es gilt wegen der zweimaligen stetigen Differenzierbarkeit von f

$$u_{k+1} - \overline{u} = f(u_k) - f(\overline{u}) = f'(\overline{u})(u_k - \overline{u}) + \frac{1}{2} f''(\tilde{u}_k)(u_k - \overline{u})^2$$

mit einem $\tilde{u}_k \in [u_k, \overline{u}]$. Nun zeigen wir induktiv, dass $|u_k - \overline{u}| \leq \delta$ für alle k falls dies für $k = 0$ erfüllt ist. Sei $\delta < 1$ und

$$C := \max_{x \in [u_0 - 1, u_0 + 1]} |f''(x)|.$$

Wir beachten, dass C endlich ist, da wir f'' und damit auch $|f''|$ als stetig vorausgesetzt haben und eine stetige Funktion auf einem kompakten Intervall beschränkt ist. Nun gilt mit der obigen Identität sowie der Dreiecksungleichung

$$|u_{k+1} - \overline{u}| = |f'(\overline{u})(u_k - \overline{u}) + \frac{1}{2} f''(\tilde{u}_k)(u_k - \overline{u})^2| \leq |f'(\overline{u})(u_k - \overline{u})| + |\frac{1}{2} f''(\tilde{u}_k)(u_k - \overline{u})^2|.$$

Mit unserer Abschätzung für die zweite Ableitung folgt insbesondere

$$|u_{k+1} - \overline{u}| \leq |f'(\overline{u})| \ |u_k - \overline{u}| + \frac{C}{2} |u_k - \overline{u}|^2.$$

Da $|u_k - \overline{u}| \leq \delta$ folgt insbesondere

$$|u_{k+1} - \overline{u}| \leq (|f'(\overline{u})| + \frac{C}{2}\delta) \ |u_k - \overline{u}| \leq\leq (|f'(\overline{u})| + \frac{C}{2}\delta)\delta.$$

Wegen $|f'(\overline{u})| < 1$ ist für δ hinreichend klein die rechte Seite immer noch kleiner gleich δ und damit gilt auch $|u_{k+1} - \overline{u}| \leq \delta$. Weiters erhalten wir aus der Abschätzung auch

$$|u_{k+1} - \overline{u}| \leq (|f'(\overline{u})| + \frac{C}{2}\delta) \ |u_k - \overline{u}|$$

woraus wir induktiv folgern, dass

$$0 \leq |u_k - \overline{u}| \leq (|f'(\overline{u})| + \frac{C}{2}\delta)^k \ |u_0 - \overline{u}|$$

gilt. Da $|f'(\overline{u})| + \frac{C}{2}\delta$ kleiner als eins ist, konvergiert die rechte Seite gegen Null, also auch u_k gegen $\overline{u}$, also ist die stationäre Lösung asymptotisch stabil.

Im Fall $|f'(\overline{u}| > 1$ gehen wir ähnlich vor, verwenden aber die zweite Dreiecksungleichung

$$|u_{k+1} - \overline{u}| =\geq |f'(\overline{u})(u_k - \overline{u})| - |\frac{1}{2}f''(\tilde{u}_k)(u_k - \overline{u})^2|.$$

Nun schätzen wir wieder die zweite Ableitung ab und erhalten für $|u_k - \overline{u}| \leq \delta$

$$|u_{k+1} - \overline{u}| \leq (|f'(\overline{u})| - \frac{C}{2}\delta)\ |u_k - \overline{u}|.$$

Für δ hinreichend klein ist nun der Faktor $|f'(\overline{u})| - \frac{C}{2}\delta$ immer noch größer als eins, damit wächst die Norm und es wird ab einem endlichen Index k_* gelten $|u_{k_*} - \overline{u}| > \delta$. Somit ist $\overline{u}$ instabil. □

Kapitel 6

Optimale Steuerung im Fischfang

Wir betrachten nun ein einfaches Modell für die Entwicklung einer Fischpopulation in einem Gebiet mit begrenzten Resourcen (z.B. Teich). $N(t)$ sei die erwartete Anzahl an Fischen zur Zeit t und r die Differenz aus Geburten- und Sterberate (pro Fisch pro Zeit), so ist ein einfaches Modell gegeben durch

$$N'(t) = rN(t).$$

Wenn $r > 0$ ist, so ist dies für kleine Populationen, die noch nicht durch die Resourcen begrenzt sind, sicher vernünftig, allerdings wächst die Populationsgröße jedoch exponentiell mit e^{rt}. Für grössere Populationen sollte das Wachstum jedoch gehemmt werden bzw. bei einer kritischen Größe N_c sogar gestoppt. Damit ist es naheliegend die effektive Wachstumsrate linear mit N abnehmen zu lassen, sodass sie bei N_c gleich null wird, also r durch $r(1 - \frac{N}{N_c})$ zu ersetzten. Dies führt auf die logistische Differentialgleichung

$$N'(t) = rN(t)(1 - \frac{N(t)}{N_c}).$$

Wir sehen, dass wenn $N(0) < N_c$ gilt, auch $N(t) \leq N_c$ folgt, da $N' = 0$ wird, sobald $N = N_c$ erreicht ist.

Ein interessantes Problem ist nun die Steuerung der Population durch Fischfang. Ist $h > 0$ die Fangrate pro Zeit, so müssen wir offensichtlich die Gleichung zu

$$N'(t) = rN(t)(1 - \frac{N(t)}{N_c}) - h(t)$$

modifizieren. Interessant ist es nun auf jeden Fall eine *nachhaltige* Fangrate zu finden, sodass die Population nicht ausstirbt, d.h. $N(t) > 0$ für alle t. Aus wirtschaftlichen Gründen möchte man den Profit in einer gewissen Zeit, etwa dem Intervall $(0, T)$ maximieren. Nimmt man an, dass die Nachfrage groß genug und der Preis pro Fisch P_f konstant bleibt, so ist der effektive Profit pro Zeit zur Zeit t gleich $P(t) = P_f h(t) e^{-\rho t}$, wobei ρ die Inflationsrate ist. Wir beachten, dass wegen der Inflation die Preise anderer Güter analog zu Zinsen mit $e^{\rho t}$ steigen, sodass der Wert bei konstantem Preis natürlich genau in dem Maße abnimmt. Wollen wir nun die Population optimal steuern, so haben wir folgendes Problem: bestimme h so, dass

$$\int_0^T h(t) e^{-\rho t} \, dt \to \max$$

unter den Nebenbedingungen

$$N(t) > 0, \qquad N'(t) = rN(t)(1 - \frac{N(t)}{N_c}) - h(t).$$

Dies ist ein optimales Steuerungsproblem für eine Differentialgleichung. Besondere Schwierigkeiten dabei sind, dass einerseits die Unbekannte der Optimierung h eine Funktion ist und andererseits die Nebenbedingung eine Differentialgleichung ist. Solche Optimalsteuerungsprobleme können durch Optimierungsmethoden in unendlichdimensionalen Banachräumen gelöst werden. Wir verfolgen hier aber einen einfacheren Ansatz und betrachten nur eine konstante Fangrate, d.h. $h(t) = FN(t)$. Damit reduziert das Optimierungsproblem auf eine einfache Gleichung der Form

$$N'(t) = rN(t)(1 - \frac{N(t)}{N_c}) - FN(t) = rN(t)(1 - \frac{N(t)}{N_c} - \frac{F}{r}).$$

Wir sehen sofort, dass sich die maximale Populationsgröße zu $N_c(1-\frac{F}{r})$ ändert. Damit wir eine vernünftige Lösung erhalten können sollte immer $F \leq r$ gelten, sonst stirbt die Population sicher aus. In diesem Fall erhalten wir eine positive Lösung, falls $N(0) > 0$ ist, also erfüllen wir die Nebenbedingung $N(t) > 0$ automatisch.

Wegen der Existenz einer eindeutigen Lösung N_F der Differentialgleichung

$$N_F'(t) = rN_F(t)(1 - \frac{N_F(t)}{N_c} - \frac{F}{r}), \qquad N_F(0) = N_0$$

für gegebenes F, können wir eindeutig eine Abbildung $F \mapsto N_F$ und damit auch eine Abbildung $J : \mathbb{R} \to \mathbb{R}$,

$$F \mapsto J(F) = \int_0^T FN_F(t)e^{-\rho t}\ dt.$$

definieren. Wir wollen also eine eindimensionale Funktion J über dem Intervall $[0, r]$ maximieren, ein auf den ersten Blick einfaches Problem. Die Schwierigkeit liegt darin, dass die Funktion J nicht explizit gegeben ist, sondern erst durch Lösung einer Differentialgleichung berechnet werden muss, ein Problem, das man häufig in realen Optimierungsszenarien findet. Insbesondere ist auch die Berechnung der Ableitung von J nach F ein schwieriges Problem.

Wir wollen die Berechnung durch Differenzenquotienten durchführen. Für kleines ϵ gilt

$$\begin{aligned} \frac{J(F+\epsilon) - J(F)}{\epsilon} &= \int_0^T \frac{(F+\epsilon)N_{F+\epsilon}(t) - FN_F(t)}{\epsilon} e^{-\rho t}\ dt \\ &= \int_0^T \left(N_{F+\epsilon}(t) + F\frac{N_{F+\epsilon}(t) - N_F(t)}{\epsilon} \right) e^{-\rho t}\ dt. \end{aligned}$$

Im Grenzwert $\epsilon \to 0$ erhalten wir

$$J'(F) = \int_0^T \left(N_F(t) + F\tilde{N}_F(t) \right) e^{-\rho t}\ dt,$$

wobei

$$\tilde{N}_F(t) = \lim_{\epsilon \to 0} \frac{N_{F+\epsilon}(t) - N_F(t)}{\epsilon},$$

d.h. die Ableitung von N_F nach F, ist. $\tilde{N}_F(t)$ können wir nicht explizit berechnen, aber wir können eine Differentialgleichung dafür herleiten. Da wir den Anfangswert fest halten, gilt $N_{F+\epsilon}(0) - N_F(0) = 0$ und damit im Grenzwert $\tilde{N}_F(0) = 0$. In der Gleichung erhalten wir

$$\tilde{N}'_F(t) = r\tilde{N}_F(1 - \frac{F}{r} - 2\frac{N_F(t)}{N_c}) - N_F(t).$$

Wollen wir nun die Lösung berechnen, so erhalten wir ein System von Differentialgleichungen. Setzen wir noch

$$j_F(t) = \int_0^t \left(N_F(s) + F\tilde{N}_F(s)\right) e^{-\rho s}\, ds,$$

so ist $J'(F) = 0$ gleichbedeutend mit $j_F(T) = 0$. Weiters gilt

$$\begin{aligned}
N'_F(t) &= rN_F(t)(1 - \frac{F}{r} - \frac{N_F(t)}{N_c}) \\
\tilde{N}'_F(t) &= r\tilde{N}_F(t)(1 - \frac{F}{r} - 2\frac{N_F(t)}{N_c}) - N_F(t) \\
j'_F(t) &= \left(N_F(t) + F\tilde{N}_F(t)\right) e^{-\rho t}.
\end{aligned}$$

Da wir die Anfangswerte $N_F(0) = N_0$, $\tilde{N}_F(0) = 0$, $j_F(0) = 0$ und den Endwert $j_F(T) = 0$ gegeben haben, ist dies insgesamt ein Randwertproblem für das System an Differentialgleichungen und die Rate F. Dieses kann zumindest numerisch gelöst werden.

Kapitel 7

Partielle Differentialgleichungen: Die Wellengleichung

Als kleine Einführung in partielle Differentialgleichungen wollen wir uns im Folgenden mit Wellenphänomenen beschäftigen. Man unterscheidet zwischen zwei verschiedenen Arten von Wellen:

- *Longitudinalwellen:* hier erfolgt eine lokale Bewegung in Richung der Wellenausbreitung, wie etwa bei einer schunkelnden Menge.
- *Transversalwellen:* hier erfolgt eine lokale Bewegung orthogonal zur Bewegungsrichtung, wie etwa bei einer La Ola Welle in einem Stadion.

Wir beginnen mit einer einfachen Modellierung einer Longitudinalwelle, dazu nehmen wir an, dass zu Beginn Teilchen mit kleinem regelmässigen Abstand Δx in einer Reihe angeordnet sind, die sich dann in Zeitschritten der Größe Δt bewegen. Als Variable zur Beschreibung der Welle wählen wir die Auslenkung $u(x,t)$ eines Teilchens zur Zeit t, das sich zur Zeit 0 im Punkt x befindet. Bei einer Welle, die von links nach rechts läuft, ist die Auslenkung eines Teilchens gleich der Auslenkung des von ihm links liegenden Teilchens im Zeitschritt davor, d.h. es gilt

$$u(x,t) = u(x - \Delta x, t - \Delta t).$$

Analog gilt für eine von rechts nach links gehende Welle

$$u(x,t) = u(x + \Delta x, t - \Delta t).$$

Nehmen wir nun an, dass es eine glatte Funktion gibt, die die kontinuierliche Welle beschreibt, so folgt aus der Taylor-Entwicklung

$$u(x,t) = u(x \pm \Delta x, t - \Delta t) \approx u(x,t) \mp \frac{\partial u}{\partial x}(x,t)\Delta x - \frac{\partial u}{\partial t}(x,t)\Delta t.$$

Damit erhalten wir für eine nach rechts laufende Welle u_R die Gleichung

$$\frac{\partial u_R}{\partial t}(x,t) + c\frac{\partial u_R}{\partial x}(x,t) = 0,$$

mit der Wellengeschwindigkeit $c = \frac{\Delta x}{\Delta t}$. Analog haben wir für eine nach links laufende Welle u_L die Gleichung

$$\frac{\partial u_L}{\partial t}(x,t) + c\frac{\partial u_L}{\partial x}(x,t) = 0.$$

Für eine Transversalwelle können wir ähnlich vorgehen, indem wir als Variable die orthogonale Auslenkung $u(x,t)$ wählen und erhalten die selben Gleichungen für einfache Wellen. Ein verbleibender Punkt ist die Überlagerung von Wellen. Der einfachste Fall ist eine lineare Überlagerung von Wellen in verschiedene Richtungen, d.h. die Gesamtauslenkung u ist gegeben durch $u = u_L + u_R$. Wollen wir nun eine Gleichung für u finden, können wir höhere Ableitungen berechnen. Zunächst gilt

$$\frac{\partial u}{\partial t} = \frac{\partial u_L}{\partial t} + \frac{\partial u_R}{\partial t} = c\frac{\partial u_L}{\partial t} - c\frac{\partial u_R}{\partial t}$$

und

$$\frac{\partial u_L}{\partial t} - \frac{\partial u_R}{\partial t} = c\frac{\partial u_L}{\partial t} + c\frac{\partial u_R}{\partial t}.$$

Berechnen wir nun weitere Ableitungen (der ersten Gleichung bezüglich x und der zweiten bezüglich t), so gilt

$$\frac{\partial^2 u}{\partial t^2} = c\frac{\partial^2}{\partial x \partial t}(u_L - u_R) = c^2\frac{\partial^2 u}{\partial x^2}.$$

Dies ist die lineare Wellengleichung

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}.$$

Im Fall einer nichtlinearen Überlagerung ist die Herleitung schwieriger und führt dann zu einer nichtlinearen partiellen Differentialgleichung. Ein Beispiel ist die Überlagerung bei einer La Ola Welle, die eher von der Form

$$u(x,t) = \max\{u_L(x,t), u_R(x,t)\}$$

ist.

Kapitel 8

Fehleranalyse bei numerischen Rechnungen

Zunächst wollen wir überlegen, welche Arten von Fehlern uns in der numerischen Mathematik begegnen. Im Wesentlichen können wir vier verschiedene Fehlerquellen unterscheiden:

1. Modellierungsfehler: Typischerweise legt man ein (vereinfachtes) mathematisches Modell zugrunde, dass die Wirklichkeit nicht vollständig beschreiben kann, d.h. das Modell entspricht nicht exakt der Anwendung.

2. Diskretisierungsfehler: Implementiert wird nicht die exakte mathematische Formel sondern eine Approximation, wodurch Fehler entstehen. Beispiel: der Differenzenquotient als Approximation für die Ableitung,

$$f'(x) \sim \frac{f(x+h) - f(x)}{h}, \quad \text{für } h \text{ “klein”}.$$

3. Messfehler: Die gemessenen Daten haben in der Praxis nur endliche Genauigkeit oder, anders ausgedrückt, sind fehlerbehaftet. Diese Fehler führen dann unglücklicherweise auch zu Fehlern in unseren numerischen Methoden.

4. Rechenfehler: Computer müssen beim Rechnen Runden, was zu Fehlern in den Berechnungen führt.

Interessant für uns sind die Punkte 3 und 4, wobei typischerweise der Messfehler den Rechenfehler überwiegt.

8.1 Messfehler und Fehlerverstärkung

Wir wollen, zunächst an einem Beispiel, die Auswirkungen von Messfehlern auf das Ergebnis einer numerischen Methode untersuchen. Sei beispielsweise eine stetig differenzierbare Funktion $f\colon \mathbb{R} \to \mathbb{R}$ gegeben, und nehmen wir einmal an, wir wollen diese an einer Stelle $x \in \mathbb{R}$ auswerten. Unglücklicherweise kennen wir diese Stelle aber nicht, sondern nur eine Approximation $\tilde{x} \in \mathbb{R}$, so dass der Abstand zwischen x und $\tilde{x}$,

$$|\Delta x| = |x - \tilde{x}|,$$

klein ist. Es bleibt uns also nicht viel anderes übrig, als f an der Stelle $\tilde{x}$ auszuwerten, aber wir fragen uns: Wie groß ist der Fehler im Ergebnis $f(\tilde{x})$? Dazu definieren wir zunächst, was genau ein Fehler ist.

Definition 8.1. Sei $x \in \mathbb{R}^n$ ein gegebener Vektor und $\tilde{x} \in \mathbb{R}^n$ eine Näherung, $\|\cdot\|$ eine Norm auf $\mathbb{R}^n$. Dann definieren wir

1. den *absoluten* Fehler:
$$\|\Delta x\| = \|x - \tilde{x}\|,$$
2. den *relativen* Fehler bzgl. x (für $x \neq 0$):
$$\frac{\|\Delta x\|}{\|x\|} = \frac{\|x - \tilde{x}\|}{\|x\|},$$
3. den *relativen* Fehler bzgl. $\tilde{x}$ (für $\tilde{x} \neq 0$):
$$\frac{\|\Delta x\|}{\|\tilde{x}\|} = \frac{\|x - \tilde{x}\|}{\|\tilde{x}\|}.$$

Selbstverständlich ist für kleinen Unterschied zwischen x und $\tilde{x}$ auch der Unterschied in den relativen Fehlern gering, das heißt
$$\frac{\|\Delta x\|}{\|x\|} \sim \frac{\|\Delta x\|}{\|\tilde{x}\|}.$$

Nun zurück zu unserem Problem von oben. Gesucht ist eine Abschätzung für den maximalen, relativen Fehler in $f(\tilde{x})$, bei gegebenem Fehler $|\Delta x|$ in $\tilde{x}$. Anders ausgedrückt: Bei gegebenem (relativem) Fehler im Input, wie groß ist der maximale (relative) Fehler im Output, d.h. wir suchen eine Abschätzung der Form ($f(x) \neq 0$, $x \neq 0$)
$$\frac{|f(x) - f(\tilde{x})|}{|f(x)|} \leq C \frac{|x - \tilde{x}|}{|x|}.$$

Um so eine Abschätzung zu erhalten, nutzen wir die Taylorentwicklung von f, bzw. den Mittelwertsatz. Sei ohne Einschränkung der Allgemeinheit $\tilde{x} \geq x$, dann existiert nach dem MWS ein $\xi \in [x, x + \Delta x] = [x, \tilde{x}]$ so dass
$$f(x) = f(\tilde{x}) + f'(\xi)(x - \tilde{x}) \quad \Leftrightarrow \quad f(x) - f(\tilde{x}) = f'(\xi)(x - \tilde{x}).$$

Damit rechnen wird (für $f(x) \neq 0, x \neq 0$)
$$\begin{aligned}\frac{|f(x) - f(\tilde{x})|}{|f(x)|} &= \left|\frac{f'(\xi)(x - \tilde{x})}{f(x)}\right| = \left|\frac{f'(\xi)x}{f(x)}\frac{(x - \tilde{x})}{x}\right| = \left|\frac{f'(\xi)x}{f(x)}\right|\left|\frac{(x - \tilde{x})}{x}\right| \\ &\leq \max_{\xi' \in [x,\tilde{x}]} \left|\frac{f'(\xi')x}{f(x)}\right|\left|\frac{(x - \tilde{x})}{x}\right| = M\left|\frac{(x - \tilde{x})}{x}\right|.\end{aligned}$$

Die Zahl M wird *Konditionszahl* oder Verstärkungsfaktor genannt und besagt, um welchen Faktor ein Fehler im Input *maximal* erhöht wird. Da das $\xi \in [x, \tilde{x}]$ aus der Taylorentwicklung im Allgemeinen nicht bekannt ist, nehmen wir das Maximum über alle ξ' in diesem Intervall.

Allerdings ist typischerweise auch das Intervall $[x, \tilde{x}]$ nicht genau bekannt (sonst würden wir ja auch x kennen). Für (kleinen) Fehler $|\Delta x|$ approximiert man daher häufig $f'(\xi) \sim f'(x)$, so dass

$$M = \max_{\xi' \in [x,\tilde{x}]} \left| \frac{f'(\xi')x}{f(x)} \right| \sim \left| \frac{f'(x)x}{f(x)} \right|.$$

Dieses M werden wir im Folgenden benutzen.

Beispiel 8.2. *Sei $f(x) = x^\alpha$. Dann ist $f'(x) = \alpha x^{\alpha-1}$ und damit gilt für die Konditionszahl*

$$M \sim \left| \frac{f'(x)x}{f(x)} \right| = \left| \frac{\alpha x^{\alpha-1} x}{x^\alpha} \right| = |\alpha|.$$

Das heißt, je größer der Exponent, desto größer der maximal Fehler im Output. Dass das stimmt, ist leicht einzusehen: Je größer der Exponent, desto steiler steigt f an, so dass sich $f(x)$ stark verändert, auch wenn wir x nur leicht verändern; für kleinen Exponenten ist f sehr flach und die Fehler sind wesentlich geringer.

Um die Fehlerverstärkung von anderen numerischen Verfahren, zum Beispiel der Addition, zu untersuchen, müssen wir die obige Abschätzung zunächst auf allgemeine Vektoren $x \in \mathbb{R}^n$ und Funktionen $f\colon \mathbb{R}^n \to \mathbb{R}$ erweitern. Sei also $f\colon \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{R}^n$ und $\tilde{x} \in \mathbb{R}^n$ eine Näherung an x. Dann existiert (wieder mit dem Mittelwertsatz) ein $\xi \in \mathbb{R}^n$, so dass

$$f(x) = f(\tilde{x}) + \langle f'(\xi), x - \tilde{x} \rangle = f(\tilde{x}) + \sum_{i=1}^{n} \frac{\partial f(\xi)}{\partial x_i}(x_i - \tilde{x}_i).$$

Damit erhalten wir mit nahezu analoger Rechnung:

$$\frac{|f(x) - f(\tilde{x})|}{|f(x)|} = \left| \frac{\sum_{i=1}^n \frac{\partial f(\xi)}{\partial x_i}(x_i - \tilde{x}_i)}{f(x)} \right| \leq \sum_{i=1}^{n} \left| \frac{\frac{\partial f(\xi)}{\partial x_i} x_i}{f(x)} \frac{x_i - \tilde{x}_i}{x_i} \right| \leq \sum_{i=1}^{n} M_i \left| \frac{x_i - \tilde{x}_i}{x_i} \right|,$$

wobei

$$M_i = \max_{\xi' \in [x,\tilde{x}]} \left| \frac{\frac{\partial f(\xi')}{\partial x_i} x_i}{f(x)} \right|.$$

Hier bezeichnet $[x, \tilde{x}]$ allerdings kein Intervall, sondern die Verbindungsstrecke zwischen x und $\tilde{x}$ im $\mathbb{R}^n$, d.h. $[x, \tilde{x}] = \{\lambda x + (1 - \lambda)\tilde{x} \mid \lambda \in [0, 1]\}$. Auch hier approximieren wir wieder das Maximum einfach mit dem Wert an der Stelle x, so dass

$$M_i \sim \left| \frac{\frac{\partial f(x)}{\partial x_i} x_i}{f(x)} \right|.$$

Wir haben jetzt das Handwerkszeug, um die Multiplikation und Addition auf ihre Fehlerverstärkung zu untersuchen.

Beispiel 8.3. *(Multiplikation) Sei $f\colon \mathbb{R}^2 \to \mathbb{R}$, $x = (x_1, x_2) \mapsto x_1 x_2$. Dann gilt*

$$\frac{\partial f(x)}{\partial x_1} = x_2, \quad \frac{\partial f(x)}{\partial x_2} = x_1,$$

und damit für die Verstärkungsfaktoren M_1 und M_2

$$M_1 = \left| \frac{\frac{\partial f(x)}{\partial x_1} x_1}{f(x)} \right| = \left| \frac{x_2 x_1}{x_1 x_2} \right| = 1 = M_2.$$

Da $M_1 = M_2 = 1$, gibt es somit keine *Verstärkung des Inputfehlers, die Multiplikation ist somit kaum fehleranfällig.*

Ganz anders sieht es bei der Addition aus.

Beispiel 8.4. *Addition Sie $f\colon \mathbb{R}^2 \to \mathbb{R}$, $x = (x_1, x_2) \mapsto x_1 + x_2$. Hier ist*

$$\frac{\partial f(x)}{\partial x_1} = 1, \quad \frac{\partial f(x)}{\partial x_2} = 1,$$

so dass z.B. für M_1 gilt

$$M_1 = \left| \frac{\frac{\partial f(x)}{\partial x_1} x_1}{f(x)} \right| = \left| \frac{x_1}{x_1 + x_2} \right|.$$

Problematisch wird es jetzt, wenn $x_1 + x_2$ nahe bei Null ist, d.h. $x_1 \sim -x_2$, und x_1 groß ist, denn dann wird M_1 extrem groß. Tatsächlich, z.B. für $x_1 = 1$ und $x_2 = -1 - 10^{-12}$ erhalten wir

$$M_1 = \frac{1}{1 - 1 + 10^{-12}} = 10^{12}.$$

Das obige Phänomen bei der Addition heißt *Auslöschung*, und ist eine zu beachtende Fehlerquelle in der numerischen Mathematik. Eine berechtigte Frage ist, warum wir so eine Rechnung überhaupt durchführen sollten, wenn sie so schlecht funktioniert. Eine Situation, wo das Problem auftritt, ist die Approximation der Ableitung durch einen Differenzenquotienten,

$$f'(x) \sim \frac{f(x+h) - f(x)}{h}, \quad h \ll 1.$$

Führt man diese Approximation numerisch durch, z.b. für $h = 10^{-6}$, wird man eine gute Näherung erhalten. Man vermutet jetzt, dass man für noch kleineres h eine noch bessere Approximation erhält; das ist allerdings nur richtig, so lange h nicht *zu* klein wird. Wenn das passiert, ist $f(x + h)$ *zu* nah an $f(x)$, d.h. es gilt $f(x + h) \sim f(x)$ und wir erhalten als Ableitung z.B. einfach 0 und damit unter Umständen einen extrem großen Fehler.

Wir haben uns jetzt mit den Auswirkungen des Inputfehlers beschäftigt, und setzen uns jetzt mit den Problemen beim Rechnen auf einem Computer auseinander. Genauer: mit dem Runden.

8.2 Maschinenzahlen und Rundungsfehler

Um uns mit den Fehlern beim Runden auf dem Computer zu befassen, müssen wir zunächst wissen, wie ein Computer zahlen darstellt und speichert. Angenommen wir wollen die Zahl $\sqrt{2} = 1.414213...$ auf dem Computer darstellen. Die Idee ist hier, eine *Basis* b zu wählen, bezüglich derer die Zahl dargestellt werden soll, und die Anzahl der Stellen hinter dem Komma p (Mantissenlänge), die gespeichert werden sollen. Die Basis $b = 2$ wird auf dem Computer benutzt (Binary Format), wir benutzen die Basis $b = 10$ (Human Format). Wir geben zunächst ein Beispiel.

Beispiel 8.5. *Wir wollen immer noch $\sqrt{2}$ darstellen, z.b. bzgl. der Basis $b = 10$ und mit einer Nachkommastelle. Das führt uns zu der (wenig überzeugenden) Approximation*

$$\sqrt{2} \sim (0.14)_{10} \cdot 10^1.$$

Gemeint mit der Darstellung, die Gleitkommadarstellung heißt, ist

$$(0.14)_{10} \cdot 10^1 = 0 \cdot 10^1 + 1 \cdot 10^0 + 4 \cdot 10^{-1},$$

wie man vom "normalen" Rechnen kennt. Machen wir das Gleiche bezüglich der Basis $b = 2$, erhalten wir eine Computerdarstellung

$$\sqrt{2} \sim (0.1)_2 \cdot 2^1 = 0 \cdot 2^1 + 1 \cdot 2^0 = 1.$$

Hier sieht man schon das Problem: eine Nachkommastelle reicht in der Basis $b = 2$ lediglich aus, um eine sehr schlechte Approximation an $\sqrt{2}$ darzustellen. Für eine Mantissenlänge von $p = 4$ sieht das Ganze schon besser aus,

$$\sqrt{2} \sim (0.1011)_2 \cdot 2^1 = 0 \cdot 2^1 + 1 \cdot 2^0 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2} + 1 \cdot 2^{-3} = 1.375.$$

Um das obige Konzept zu vereinheitlichen, machen wir die folgende Definition.

Definition 8.6. (Maschinenzahlen) Sei eine Basis $b \geq 2$, die Mantissenlänge $p \geq 1$ und die Exponentlänge $r \geq 1$ fest gewählt. Dann ist die Menge $\mathcal{M}$ der Maschinenzahlen gegeben durch

$$\mathcal{M} = \left\{ \pm \left(\sum_{k=1}^{p} m_k b^{-k} \right) b^{\pm e} : \ m_k \in \mathbb{N}_0, 0 \leq m_k \leq b-1, e \in \mathbb{Z}, |e| < b^r \right\}.$$

Die Zahlen $m \in \mathcal{M}$ haben die *b-adische*, normalisierte Darstellung

$$m = \pm 0.m_1 m_2 m_3 \dots m_p \cdot b^{\pm e},$$

wobei $m_1 \neq 0$, außer wenn $m = 0$.

Eine kurze Erklärung:

1. Die Bedingung $m_k \in \mathbb{N}_0, 0 \leq m_k \leq b-1$ sagt ganz einfach, dass die Zahlen m_k bezüglich der Basis existieren. Zum Beispiel für $b = 2$ gilt $m_k \in \{0, 1\}$, so dass die Darstellung nur aus Nullen und Einsen bestehen darf, im Human Format für $b = 10$ stehen uns jedoch die Zahlen $\{0, \dots, 9\}$ als Nachkommazahlen zur Verfügung.

2. Die Bedingung $e \in \mathbb{Z}$ besagt, dass der Exponent der Basis ganzzahlig sein muss. Das ist nötig, da wir z.B. für rationale Exponenten erst rechnen müssten, und wir definieren uns ja gerade erst Zahlen zum Rechnen.

3. Die Bedingung $|e| < b^r$ ist eine Bedingung an die Länge des Exponenten, da dieser auch Speicherplatz in Anspruch nimmt und daher nicht beliebig lang sein darf (so wie bei der Mantisse). Diese Bedingung beachten wir aber zunächst nicht.

Wir geben noch ein Beispiel an, das die nächsten Probleme beleuchten wird. Sei die Basis $b = 10$ gewählt, mit einer Nachkommastelle, also $p = 1$. Dann können wir z.B. $x = 2$ darstellen, als

$$2 = (0.2)_{10} \cdot 10^1.$$

Wir fragen uns jetzt, was die nächstgrößere Zahl in $\mathcal{M}$ ist. Offensichtlich können wir nur zwei Dinge tun: Entweder, wir erhöhen die erste Nachkommastelle um 1, oder den Exponenten um 1. Da zweiteres die Zahl $(0.2)_{10} \cdot 10^2 = 20$ liefert, muss die nächstgrößere Zahl also

$$(0.3)_{10} \cdot 10^1 = 3$$

sein. Das heißt wir können die Zahlen $\{1, 2, \ldots, 10\}$ darstellen, dann aber nur die Zahlen $\{10, 20, 30, \ldots, 100\}$ usw. Interessanterweise (aber nicht überraschend) werden die Abstände zwischen den Zahlen also größer, je größer der Exponent gewählt werden muss. Um z.B. die Zahl 11 darzustellen, brauchen wir damit mindestens zwei Nachkommastellen, so dass

$$11 = 0.11 \cdot 10^2.$$

Das führt zu einem weiteren Problem beim Addieren von zwei Zahlen. Für die oben gewählte Basis und Mantissenlänge gilt

$$m = 1 = 0.1 \cdot 10^1 \in \mathcal{M}, \quad n = 1000 = 0.1 \cdot 10^4 \in \mathcal{M},$$

d.h. sowohl m als auch n sind darstellbar. Jetzt ist allerdings

$$m + n = 1001 = 0.1001 \cdot 10^4 \notin \mathcal{M},$$

da wir hierfür 4 Nachkommastellen benötigen. $\mathcal{M}$ ist also nicht abgeschlossen bzgl. der Addition. Was passiert also, wenn allgemein eine Zahl $x \in \mathbb{R}$ nicht darstellbar ist in $\mathcal{M}$? Wir müssen *Runden.*

Definition 8.7. (Rundungsfunktion) Eine Funktion $\mathrm{rd}\colon \mathbb{R} \to \mathcal{M}$ heißt Rundungsfunktion, wenn

$$|\mathrm{rd}(x) - x| = \min_{m \in \mathcal{M}} |m - x|.$$

In anderen Worten, und genau wie man erwartet, rundet eine Rundungsfunktion eine Zahl $x \in \mathbb{R}$ auf die nächstgelegene Zahl $m \in \mathcal{M}$. Achtung: Runden ist nicht eindeutig! Das kann man leicht am Folgenden Beispiel sehen.

Beispiel 8.8. *Sei wieder* $b = 10$ *und* $p = 1$. *Dann ist* $0.2 = 0.2 \cdot 10^0$ *Maschinenzahl, d.h.* $0.2 \in \mathcal{M}$. *Die nächstgrößere Zahl bzgl. dieser Darstellung ist wie oben* 0.3, *die nächstkleinere ist* 0.1. *Damit folgt, dass z.B. für alle* $x \in (0.25, 0.35)$ *gilt* $\mathrm{rd}(x) = 0.3$, *für alle* $x \in (0.15, 0.25)$ *gilt* $\mathrm{rd}(x) = 0.2$. *Für* $x = 0.25$ *jedoch kommen zwei Lösungen infrage, nämlich* $\mathrm{rd}(x) \in \{0.2, 0.3\}$, *da beide Abstand* 0.05 *zu* x *haben. Hier ist das Runden dann Definitionssache, und muss explizit angegeben werden für die Routine* rd.

Wir betrachten nun noch den Fehler beim Runden, in diesem Fall den relativen Fehler $\frac{|\mathrm{rd}(x)-x|}{|rd(x)|}$. Offensichtlich kürzt sich dabei der Exponent b^{pme} und der Fehler beim Runden auf die Maschinenzahl muss kleiner als $0.5b^{-p}$ sein. Andererseits ist per Definition $m_1 \geq 1$, d.h. $|rd(x)| \geq b^{-1}$. Also folgt

$$\frac{|\mathrm{rd}(x) - x|}{|rd(x)|} \leq 0.5b^{-p+1} := \mathrm{eps}.$$

Die Zahl eps heißt dann Maschinengenauigkeit.

Kapitel 9

Eliminationsverfahren zur Lösung linearer Gleichungssystem

Wir wollen ein Gleichungssystem $Ax = b$ lösen mit $A \in \mathbb{R}^{n\times n}$, $x \in \mathbb{R}^n$ und $b \in \mathbb{R}^n$. Angenommen, wir können $A = BC$ schreiben, wobei B und C invertierbare $\mathbb{R}^{n\times n}$-Matrizen sind, dann können wir

$$A^{-1} = C^{-1}B^{-1}$$

schreiben. Das ermöglicht womöglich eine deutlich schnellere Berechnung für schöne Matrizen (unitäre oder orthogonale Matrizen). In dem Fall gewinnen wir

$$x = A^{-1}b = C^{-1}B^{-1}b$$

als schnelle Lösungsweg. Das Ziel vieler numerischer Lösungsverfahren ist es, Techniken zu finden, mit denen wir eine Matrix A gut und systematisch aufteilen können. Diese kannn dann als Algorithmus am Computer implementieren.

Zur Berechnung einer Lösung kennen wir auch die Cramersche Regel:

$$x_i = \frac{\det(A_i)}{\det(A)} \ , \ A_i := (a_1 \ \dots \ b \ \dots \ a_n)$$

d.h. die i-te Spalte wurde durch b ersetzt. Um dies zu nutzen, müssen wir zunächst die Determinanten berechnen. Dazu nutzen wir den Entwicklungssatz:

$$\det(A) = a_{1,1}\det(A_1') - a_{1,2}\det(A_2')\dots$$

Für eine 2×2-Matrix brauchen wir zwei Rechenoperationen. Für 3×3-Matrix ist der Aufwand bereits größer gleich 6 Rechenoperationen. Für 4×4-Matrizen ist er größer gleich $4\cdot3\cdot2$. Für $n\times n$-Matrizen wird die Laufzeit größer gleich $n!$, was viel zu groß ist. Dieser Algorithmus ist somit unbrauchbar in der Praxis.

9.1 Gauß-Elimination

Wir betrachten hier das Additionsverfahren (Eliminationsverfahren) aus der Schule an einem Beispiel:

$$\begin{aligned} 3x_1 + 2x_2 + x_3 &= 8 \\ 6x_1 + 5x_2 - 4x_3 &= 12 \\ -3x_1 + x_2 - 2x_3 &= -3 \end{aligned}$$

Dies sei $A^{(1)}x = b^{(1)}$. Wir rechnen $II - 2I$ und $III + I$:

$$\begin{aligned} 3x_1 + 2x_2 + x_3 &= 8 \\ x_2 - 6x_3 &= -4 \\ 3x_2 - x_3 &= 5 \end{aligned}$$

Dies ist nun $A^{(2)}x = b^{(2)}$. Wir rechnen weiter $III - 3II$:

$$\begin{aligned} 3x_1 + 2x_2 + x_3 &= 8 \\ x_2 - 6x_3 &= -4 \\ 17x_3 &= 17 \end{aligned}$$

Dies ist $A^{(3)}x = b^{(3)}$. Damit ist die Lösung da und wäre auch für jede größere Matrix so anwendbar. Wir bringen das Verfahren zu Ende, indem wir aus $17x_3 = 17$ folgern, dass $x_3 = 1$. Einsetzen in die anderen Gleichungen liefert $x_2 = 2$ und $x_1 = 1$. Wir nennen dies „rückwärts einsetzen“. Wir betrachten nun die Frage, wie genau wir $A^{(2)}$ aus $A^{(1)}$ gewinnen:

1) übernehme die erste Zeile.

2) Für die i-te Zeile mit $i > 1$: Subtrahiere von der i-ten Zeile das $l_{i,1}$-fache der ersten Zeile. Dies ist der Koeffizient von X_1 in der i-ten Gleichung: $(A^{(1)})_{i,1}$.
Der Koeffizient vor x_1 in der ersten Gleichung ist $(A^{(1)})_{1,1}$.
Der Koeffizient vor x_1 in der Gleichung im Ergebnis ist $(A^{(1)})_{i,1} - l_{i,1} \cdot (A^{(1)})_{1,1} = 0$.
Wir wählen daher

$$l_{i,1} = \frac{(A^{(1)})_{i,1}}{(A^{(1)})_{1,1}}$$

Wir stellen nun die Frage, wie wir von $A^{(k)}$ zu $A^{(k+1)}$ gelangen.

1) Behalte die ersten k Zeilen von $A^{(k)}$.

2) Subtrahiere für $i > k$ das $l_{i,k}$-fache der k-ten Zeile von der i-ten Zeile, d.h.

$$l_{i,k} = \frac{(A^{(k)})_{i,k}}{(A^{(k)})_{k,k}}$$

Wir wandeln dies in ein kleines Pseudoprogramm um:

1) $A^{(1)} := A$, $b^{(1)} := b$.

2) Für $k = 1, \ldots, n-1$:

2.1) Für $j = 1, \ldots, k$

$$A^{(k+1)}(j,:) := A^{(k)}(j,:)$$

2.2) Für $i = k+1, \ldots, n$

$$l_{i,k} := \frac{A^{(k)}(i,k)}{A^{(k)}(k,k)}$$
$$A^{(k+1)}(i,:) := A^{(k)}(i,:) - l_{i,k} A^{(k)}(k,:)$$

Damit haben wir bis zu einem gewissen Grad formalisiert, was bereits aus der Schule bekannt ist. Wir wollen nun den Rechenaufwand analysieren. Dazu betrachten wir, wie viele Rechenoperationen wir im k-ten Schritt benötigen.

- Für $i = k+1, \ldots, n$: 1 Division, n Multiplikationen, n Additionen (zusammen n Rechenoperationen). Dies müssen wir jedoch noch leicht korrigieren: Im ersten Schritt beispielsweise wird der erste Koeffizient sowieso Null, daher bleibt ein Vektor der Länge $n-1$. Im zweiten Schritt bleiben nur noch $n-2$ Koeffizienten relevant. D.h. wir brauchen im k-ten Schritt genau genommen $n-k$ Rechenoperationen.[1]
- Da wir dies in $n-k$ Zeilen durchführen, haben wir insgesamt $(n-k)(n-k)$ Rechenoperationen.
- Für alle k Schritte erhalten wir dann insgesamt

$$\sum_{k=1}^{n-1} (n-k)^2 = \sum_{k=1}^{n-1} k^2 = \frac{n^3}{3} + O(n^2)$$

  Die Anzahl der Divisionen ist $\frac{(n-1)n}{2}$, was auch in der Größenordnung von n^2 liegt. Divisionen werden in der Regel ca. fünfmal langsamer ausgeführt, dies ist aber nur eine Skalierung und verändert die Größenordnung nicht.

Wir folgern: Für eine Matrix mit $n = 10^6$ Zeilen müssen wir 10^{18} Rechenoperationen durchführen. An dieser Stelle haben wir aber noch nicht rückwärts eingesetzt, dieser Aufwand kommt noch hinzu.

Satz 9.1. *Die Durchführung der Gauß-Elimination für ein GLS $Ax = b$ mit $A \in \mathbb{R}^{n \times n}$ invertierbar, $b \in \mathbb{R}^n$, benötigt $\frac{n^3}{3} + O(n^2)$ Rechenoperationen.*

Selbst auf modernen Rechnern ist dieser Aufwand unheimlich groß und kann für große n nicht geleistet werden.

Wir halten fest:

1. Die Gauß-Elimination ist durchführbar, falls $(A^{(k)})_{k,k} \neq 0$.
2. Falls dies nicht der Fall ist, aber $(A^{(k)})_{j,k} \neq 0$ für ein $j > k$, so vertausche die j-te und k-te Zeile (elementare Zeilenumformung).

[1] Man würde de facto dem Rechner explizit sagen, dass er für die $k+1$-te Matrix die entsprechenden Koeffizienten gar nicht auszurechnen braucht. Der Pseudocode oben ist also für die Laufzeit noch nicht optimal.

3. Falls auch das nicht geht, dann sieht man, dass die k-te Spalte offenbar nur aus Nullen besteht. Dann ist die Matrix nicht invertierbar, was aber vorausgesetzt wurde. Der Fall kann also nicht auftreten.

4. Wie sieht nun $R := A^{(n)}$ aus? R ist eine rechte obere Dreiecksmatrix! D.h. $(R)_{i,k} = 0$ für $i > k$.

Wir halten die oben eingeführten Begriffe in einer Definition fest.

Definition 9.2. Sei $R \in \mathbb{R}^{n\times n}$, $(R)_{i,k} = 0$ für $i > k$, dann heißt R rechte obere Dreiecksmatrix. Sei $L \in \mathbb{R}^{n\times n}$, sei $(L)_{i,k} = 0$ für $i < k$, dann heißt L linke untere Dreiecksmatrix. Falls zusätzlich $(L)_{i,i} = 1$ für alle i gilt, so heißt L normiert.

Diese Matrizen haben schöne Eigenschaften: Seien R_1, R_2 rechte obere Dreiecksmatrizen, so ist $R_1 \cdot R_2$ ebenfalls eine rechte obere Dreiecksmatrix (wird in den übungszetteln bewiesen). Zusätzlich gilt: $(R_1 \cdot R_2)_{i,i} = (R_1)_{i,i} \cdot (R_2)_{i,i}$ für alle i.

Wir nennen diese Strategie **Spaltenpivotsuche**. Wähle im k-ten Schritt die Anordnung der Gleichungen durch Vertauschen der Gleichungen k bis n so, dass $(A)_{k,k}$ bis $(A)_{n,n}$ betraglich möglichst groß sind.

Zwei Bemerkungen zur Geschwindigkeit:

1. Die Geschwindigkeit der Gauß-Elimination ist $\frac{n^3}{3} + O(n^2)$. Bei einigen speziellen Matrizen mag der Aufwand jedoch deutlich geringer sein, wenn sie beispielsweise schon in einer schönen Form vorliegen.

2. Man könnte sich fragen, ob das Problem nicht der Algorithmus selbst ist – womöglich gibt es viel schnellere Algorithmen. Tatsächlich gibt es solche, die dann allerdings schlechtere Ergebnisse liefern. Dazu gehört als einer der ersten Strassen (1969). Die aktuell schnellste Laufzeit ist $O(n^{\log_2(7)})$.

Ein kleiner Nachteil des Gauß-Algorithmus ist, dass er noch ein wenig unsystematisch wirkt („Addiere diese Zeile“). Kann man das besser machen?

Definition 9.3. 1) Sei $R \in \mathbb{R}^{n\times n}$, $(R)_{i,k} = 0$ für $i > k$, dann heißt R rechte obere Dreiecksmatrix. Sei $L \in \mathbb{R}^{n\times n}$, sei $(L)_{i,k} = 0$ für $i < k$, dann heißt L linke untere Dreiecksmatrix. Falls zusätzlich $(L)_{i,i} = 1$ für alle i gilt, so heißt L normiert.

2) Sei $L \in \mathbb{R}^{n\times n}$ normierte linke untere Dreiecksmatrix. L heißt Elementarmatrix oder Frobeniusmatrix, falls nur in einer Spalte unterhalb der Hauptdiagonalen Elemente ungleich Null stehen, d.h. es gibt ein i, sodass $(L)_{j,k} = 0$ für $j > k$ und falls $i \neq j$.

3) Eine Matrix $P \in \mathbb{R}^{n\times n}$ heißt Permutationsmatrix, falls in jeder Zeile und jeder Spalte genau eine 1 auftritt und alle anderen Elemente 0 sind.

Beispiele für Frobeniusmatrizen werden wir in den Übungen sehen. Ein banales Beispiel für eine Permutationsmatrix P ist die Einheitsmatrix, aber auch

$$P_2 = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & \ddots & 0 & \dots & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix}$$

Es gilt $P^T P = I$ und somit $P^T = P^{-1}$. Es gilt zudem $\det(P) = \pm 1$ wegen

$$1 = \det(I) = \det(P^T P) = \det(P) \cdot \det(P^T) = \det(P^2) \ .$$

Wir betrachten, was eine Frobeniusmatrix L_i mit einem Vektor macht. Wir berechnen:

$$L_i \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ x_{i+1} - l_{i+1,i} \cdot x_i \\ x_{i+2} - l_{i+2,i} \cdot x_i \\ \vdots \\ x_n - l_{n,i} \cdot x_n \end{pmatrix}$$

Dies weist eine Parallele zum Gauß-Algorithmus aus: Wir haben ein Vielfaches einer bestimmten Zeile weiter unten abgezogen. Dies Matrix ist somit nützlich zur Berechnung von b. Unter Matrixmultiplikation erhalten wir:

$$L_i \cdot A = \begin{pmatrix} a_1 \\ \vdots \\ a_i \\ a_{i+1} - l_{i+1,i} \cdot a_i \\ \vdots \end{pmatrix}$$

mit $a_1, \ldots, a_n$ als Zeilen der Matrix A. Wir haben also einen Weg gefunden, den Gauß-Algorithmus mathematisch zu beschreiben, indem wir ihn als Multiplikation mit Matrizen ausdrücken. Wir wissen von übungsblatt 4, dass L_i^{-1} auch eine Frobeniusmatrix ist, wobei unter der i-ten Zeile sämtliche Einträge ein anderes Vorzeichen haben. Ferner wissen wir, dass $L_i L_{i+1}$ nun genau unter der i-ten und $i+1$-ten Zeile die jeweiligen Einträge von L_i und L_{i+1} haben:

$$\begin{pmatrix} 1 & 0 & & & & & \\ 0 & 1 & & & & & \\ 0 & (L_i)_{i+1,i} & 1 & & & & \\ 0 & (L_i)_{i+2,i} & (L_{i+1})_{i+2,i+1} & 1 & & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & (L_i)_{n,i} & (L_{i+1})_{n,i+1} & 0 & 0 & 1 \end{pmatrix}$$

Wir betrachten nun Permutationsmatrizen. Es gilt (wobei P_2 die Permutationsmatrix weiter oben ist)

$$P_2 x = \begin{pmatrix} x_n \\ x_1 \\ \vdots \\ x_{n-1} \end{pmatrix}$$

Die Einträge des Vektors werden also permutiert.

9.2 LR-Zerlegung

Satz 9.4 (LR-Zerlegung)**.** *Sei A eine $n \times n$-Matrix. Dann gibt es eine Permutationsmatrix P, eine normierte linke untere Dreiecksmatrix L und eine rechte obere Dreiecksmatrix R, sodass $PA = LR$.*

Beweis. Es gibt eine Anordnung der Gleichungen in $Ax = b$, so dass das Gaußsche Eliminationsverfahren durchführbar ist. Diese Anordnung werde realisiert durch die Permutationsmatrix P, d.h. PA ist die gewünschte Anordnung. Wir erhalten das neue LGS:

$$PAx = Pb$$

Wir fangen an mit $A^{(1)} = PA$ und $b^{(1)} = Pb$. Erinnerung: Wir gelangen von $A^{(k)}$ zu $A^{(k+1)}$, indem wir

$$A^{(k+1)} = L_k A^{(k)}$$

gerechnet haben, d.h. im Grunde haben wir mit einer Elementarmatrix L_k multipliziert. Wir wissen weiter, dass $R = A^{(n)}$ eine rechte obere Dreiecksmatrix ist, die sich durch $L_{n-1}A^{(n-1)}$ ergibt. Dies lässt sich rekursiv zurückführen auf

$$R = A^{(n)} = L_{n-1}L_{n-2} \cdot \ldots \cdot L_1 A^{(1)} \ .$$

Durch Invertierung folgt:

$$PA = A^{(1)} = (L_1)^{-1} \cdot \ldots \cdot L_{n-1}^{-1} R$$

Wir wissen nun, dass das Produkt der inversen Elementarmatrizen wieder eine Elementarmatrix ist, genauer sogar eine normierte linke untere Dreiecksmatrix L. Die Einträge unter den jeweiligen Einträgen der Hauptdiagonale sind genau die der einzelnen L_i^{-1}. Damit haben wir genau die behauptete Form hergestellt. □

Zur Interpretation der LR-Zerlegung: Die Permutationsmatrix sorgt dafür, dass die Zeilen die gewünschte Form haben. Die Elementarmatrix stellt dann die Zeilenumformungen (das Additionsverfahren) dar.

Ein paar Anmerkungen:

1) Sei $PA = LR$. Bestimme x, sodass $Ax = b$. Dies geht ohne Weiteres nicht, wir multiplizieren zunächst beide Seiten mi P: $PAx = Pb$. Nun ist $PA = LR$, also somit $LRx = Pb$. Wir setzen nun $y = Rx$ und erhalten:

   $$Ly = Pb \qquad Rx = y$$

   Wir wissen nun, dass L normierte linke untere Dreiecksmatrix ist – dies wollen wir nutzen. In der ersten Zeile von $L \cdot y = Pb$ steht nur $y_1 = (Pb)_1$. In der zweiten Zeile steht

   $$(L)_{2,1}y_1 + y_2 = (Pb)_2$$

   und somit $y_2 = b_2 - (L)_{2,1}y_1$. Dies nennen wir **Vorwärtseinsetzen**. Dies führen wir für alle y_i durch. In $Rx = y$ lautet die letzte Zeile

   $$(R)_{n,n}x_n = y_n$$

und somit $x_n = \frac{y_n}{R_{n,n}}$. Für die zweitletzte Spalte haben wir

$$(R)_{n-1,n-1}x_{n-1} + (R)_{n-1,n}x_n = y_{n-1}$$

und als Auflösung

$$x_{n-1} = \frac{1}{(R)_{n-1,n-1}} \left(y_{n-1} - (R)_{n-1,n}x_n\right)$$

Dies nennen wir **Rückwärtseinsetzen**. Mit diesem Schema lassen sich mittels der LR-Zerlegung LGSe lösen.

2) Nicht jede invertierbare Matrix besitzt eine LR-Zerlegung, die ohne Permutation funktioniert. Ein Beispiel wäre:

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

Da links oben eine Null steht, ist das Gauß-Verfahren nicht durchführbar ohne Zeilenvertauschung.

3) Auch singuläre Matrizen besitzen eine LR-Zerlegung. Sogar die Nullmatrix lässt sich schreiben mittels $P \cdot (0) = L \cdot (0)$, denn: Wir haben für die rechte obere Dreiecksmatrix nicht vorausgesetzt, dass sie normiert ist.

Satz 9.5. *Sei $A \in \mathbb{R}^{n\times n}$ invertierbar. Es gebe L und R so, dass $A = LR$ mit den üblichen Eigenschaften. Weiter sei $A = L'R'$ eine zweite solche Zerlegung. Dann gilt $L = L'$ und $R = R'$.*

Beweis. Es gilt $LR = A = L'R'$. Da A invertierbar ist, sind auch alle Zerlegungsmatrizen invertierbar. Daraus folgt nun, dass $(L')^{-1}L = R'R^{-1}$. Wir wissen nun bereits, dass Inverse von normierten unteren Dreiecksmatrizen wieder solche sind und ebenso Produkte von normierten linken unteren Dreiecksmatrizen. Somit ist $(L')^{-1}L$ eine linke untere Dreiecksmatriz, analog ist $R'R^{-1}$ eine rechte obere Dreiecksmatrix. Da zwischen diesen beiden Matrizen nun Gleichheit besteht, folgt, dass auf der Hauptdiagonalen nur Einsen stehen. Ansonsten können nur Nullen in der Matrix stehen (andernfalls wären es vorher keine Dreiecksmatrizen gewesen). Somit folgt $(L')^{-1}L = I$, woraus $L = L'$ folgt. Analog für R und R'. Das war zu zeigen. □

Kapitel 10

Fehleranalyse bei der Lösung linearer Gleichungssysteme

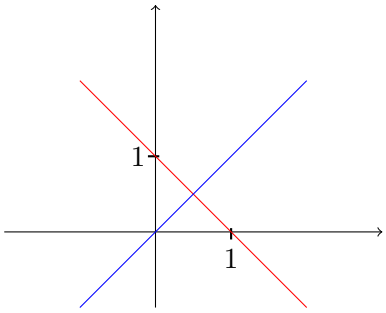
Sei $Ax = b$ gegeben, es sei A invertierbar. Statt A steht nur eine Näherung $\widetilde{A} = A + dA$ zur Verfügung, ebenso gibt es für b nur eine Annäherung $\widetilde{b} = b + db$. Wir müssen daher $\widetilde{A}\widetilde{x} = \widetilde{b}$ mit $\widetilde{x} = x + dx$ lösen. Nun ist die Frage, ob

$$\frac{\|dx\|}{\|x\|} \leq K \cdot \left(\frac{\|dA\|}{\|A\|} + \frac{\|db\|}{\|b\|} \right)$$

gilt, d.h. ob eine solche Fehlerabschätzung existiert. Sei beispielsweise das LGS

$$\begin{aligned} x + y &= 1 \\ x - y &= 0 \end{aligned}$$

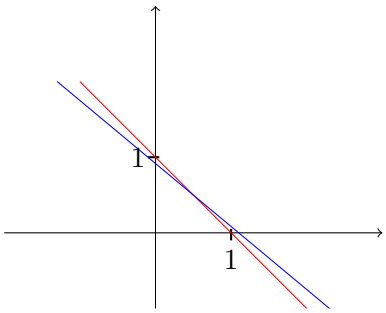
und dessen graphische Realisierung gegeben durch (in rot dargestellt die Gerade, die sich aus Zeile 1 ergibt, in blau die durch Zeile 2 erzeugte Gerade):



Die Lösung ist offensichtlich $(1/2, 1/2)$. Ersetzt man nun im LGS 1 durch 1.2, so rechnet man leicht nach, dass der Fehler in der Lösung ähnlich wenig abweicht wie 1 von 1.2. In folgendem System

$$\begin{aligned} x + y &= 1 \\ x + 1.2y &= 1.1 \end{aligned}$$

ist das nicht mehr der Fall. Die Lösung ist wieder $(1/2, 1/2)$.



Zeichnet man die entsprechenden Gerade nur leicht ungenau, so ergibt sich direkt ein grober Fehler, da sich die Geraden auf einer stark unterschiedlichen Höhe schneiden. Möglicherweise habe ich auch die Matrix falsch gemessen, beispielsweise

$$\begin{aligned} x + y &= 1 \\ x + y &= 1.1 \end{aligned}$$

und somit wäre das LGS gar nicht mehr lösbar, denn $\widetilde{A}$ ist in diesem Fall nicht mehr invertierbar. Für die folgenden Betrachtungen erinnern wir an die geometrische Reihe: Sei $q \in \mathbb{R}$, dann gilt

$$(1-q)^{-1} = \sum_{i=1}^{\infty} q^k \ ,$$

wenn $|q| < 1$. Es ist vom Aufwand her deutlich einfacher, eine Potenzen zu addieren als eine Division durchzuführen. Daher wenden wir dieses Prinzip nun auf Matrizen an:

$$(I-T)^{-1} = \sum_{k=0}^{\infty} T^k$$

mit $\|T\| < 1$.

Satz 10.1 (Neumannsche Reihe)**.** *Sei $(V, \|\cdot\|)$ vollständig, also Banachraum. Sei $T : V \to V$ linear und es gelte $\|T\| < 1$ mit der induzierten Matrixnorm. Dann ist $(I-T)$ invertierbar und es gilt*

$$(I-T)^{-1} = \sum_{k=0}^{\infty} T^k \ .$$

Beweis. Sei $V = \mathbb{R}^n$ und $T = \mathbb{R}^{n \times n}$. Wir zeigen nun, dass wir mit der Summe eine Cauchyfolge vorliegen haben, d.h.

$$\| \sum_{k=r}^{s} T^k \| \leq \sum_{k=r}^{s} \|T^k\| \leq \sum_{k=r}^{s} \|T\|^k \leq \|T\|^r \cdot \sum_{k=0}^{s-r} \|T\|^k$$

aufgrund der Dreiecksungleichung. Wir schätzen weiter

$$\|T\|^r \cdot \sum_{k=0}^{s-r} \|T\|^k \leq \|T\|^r \cdot \frac{1}{1-\|T\|} \overset{k \to \infty}{\longrightarrow} 0$$

mittels der geometrischen Summenformel. Somit konvergiert $\sum_{k=0}^{\infty} T^k$ als Cauchyfolge, da die reellen Zahlen vollständig sind – den Grenzwert nennen wir B. Wir betrachten nun

$$\begin{aligned}(I-T)\cdot B &= \lim_{n\to\infty}(I-T)\cdot\sum_{k=0}^{n}T^k \\ &= \lim_{n\to\infty}(I-T)(I+T-T+T^2+\ldots-T^n) \\ &= \lim_{n\to\infty} I-T^{n+1}\end{aligned}$$

unter Anwendung der Teleskopsumme (alle T schmeißen sich gegenseitig raus). Wir wissen nun wegen $\|T^{k+1}\| \leq \|T\|^{k+1} \to 0$, dass $\lim_{n\to\infty} I - T^{n+1} = 1$. Daraus folgt, dass $B = (I-T)^{-1}$. □

Man kann dieses Resultat auch so interpretieren: Haben wir die Einheitsmatrix und stören sie ein wenig, so bleibt die Matrix nach wie vor invertierbar, sofern die Störung (in obigem thm T) nicht allzu groß wird.

Korollar 10.2. *Sei V ein Banachraum, sei $T : V \to V$ invertierbar. Es sei $dT : V \to V$ ebenfalls eine Matrix, es sei $q = \|T^{-1}\| \cdot \|dT\| < 1$. Dann ist $T + dT$ invertierbar und es gilt*

$$\|(T+dT)^{-1}\| \leq \|T^{-1}\| \cdot \frac{1}{1-q} \ .$$

Zur Interpretation: Wir haben eine invertierbare Matrix, die wir ein wenig stören und wir sehen, dass sie noch immer invertierbar ist. Zusätzlich können wir die Norm der Annäherung abschätzen.

Beweis. Es ist $(T+dT) = T\cdot(I-(-T^{-1}dT))$. Wir wissen, dass T invertierbar ist. Der restliche Teil ist genau dann invertierbar, wenn seine Norm echt kleiner als 1 ist. Dies ist trivialerweise der Fall wegen

$$\| - T^{-1}dT\| \leq \|T^{-1}\| \cdot \|dT\| = q < 1 \ .$$

Wir erhalten weiter

$$(T+dT)^{-1} = \left(\sum_{k=0}^{\infty}(-1)^k(T^{-1}dT)^k\right)\cdot T^{-1}$$

nach obigem thm. Legen wir darum die Norm und wenden auf die Reihe die Dreiecksungleichung an, so können wir die Reihe gegen die geometrische Reihe $\sum_{k=0}^{\infty} q^k$ abschätzen und erhalten die Behauptung. □

Satz 10.3. *Sei $A \in \mathbb{R}^{n\times n}$ invertierbar. Sei $b \in \mathbb{R}^n$ und $x \in \mathbb{R}^n$ und $Ax = b$. Sei $dA \in \mathbb{R}^{n\times n}$ und $db \in \mathbb{R}^n$. Wir bezeichnen mit*

$$K(A) := \|A\| \cdot \|A^{-1}\|$$

die Kondition von A. Es gelte

$$q = \|dA\| \cdot \|A^{-1}\| = \|dA\| \cdot \frac{K(A)}{\|A^{-1}\|} = K(A) \cdot \frac{\|dA\|}{\|A\|} < 1 \ .$$

Dann ist $A + dA$ invertierbar. Sei $\widetilde{x} = x + dx$ die Lösung von $(A + dA)(x + dx) = (b + bx)$. Dann gilt

$$\frac{\|dx\|}{\|x\|} \leq \frac{K(A)}{1-q}\left(\frac{\|dA\|}{\|A\|} + \frac{\|db\|}{\|b\|}\right) .$$

Beweis. Es ist

$$\begin{aligned}(A + dA)(x + dx) = Ax + dAx + (A + dA)dx = b + db\\ \Leftrightarrow (A + dA)dx = db - dAx\\ \Leftrightarrow dx = (A + dA)^{-1}(db - dAx)\end{aligned}$$

Uns interessiert allerdings die Norm, daher folgern wir

$$\begin{aligned}\frac{\|dx\|}{\|x\|} &\leq \frac{\|(A + dA)^{-1}\| \cdot (\|db\| + \|dA\| \cdot \|x\|)}{\|x\|}\\ &\leq \|A^{-1}\| \cdot \frac{1}{1-q} \cdot \left(\frac{\|db\|}{\|x\|} + \|dA\|\right)\\ &\leq \frac{\|A\| \cdot \|A^{-1}\|}{1-q} \cdot \left(\frac{\|db\|}{\|A\| \cdot \|x\|} + \frac{\|dA\|}{\|A\|}\right)\\ &\leq \frac{K(A)}{1-q}\left(\frac{\|db\|}{\|b\|} + \frac{\|dA\|}{\|A\|}\right)\end{aligned}$$

mittels der üblichen Regeln für Normen und wegen des obigen Korollars. □

10.1 Fehler bei der Gauss Elimination

Wir betrachten beispielhaft das folgende LGS:

$$\begin{aligned}10^{-4}x_1 + x_2 &= 1 + 10^{-4}\\ x_1 + x_2 &= 2\end{aligned}$$

Die Lösung ist natürlich $x_1 = x_2 = 1$. Der unvermeidbare Fehler ist in etwa

$$\frac{2 \cdot 1}{1-q} \cdot (\text{eps} + \text{eps})$$

mit eps als relativem Fehler, den der Computer automatisch macht. Im human format (2 Dezimalstellen) ist eps ≈ 0.05, d.h. der Fehler sollte in der Größenordnung 10% liegen. Wir haben nun:

$$\begin{aligned}10^{-4}x_1 + x_2 = 1 \oplus 10^{-4} &= 1\\ x_2(1 - 10^4) = 2 \ominus 10^4(1 \oplus 10^{-4})&\\ -10^4 x_2 = -10^4&\end{aligned}$$

Wir lösen von hinten auf und erhalten $x_2 = 1$. Weiter:

$$x_1 = 10^{-4}(1 - x_2) = 0 .$$

Das ist ein sehr schlechtes Ergebnis, denn der Fehler in x_1 ist somit 100%. Wir vertauschen nun die Gleichungen:

$$\begin{aligned} x_1 + x_2 &= 2 \\ 10^{-4}x_1 + x_2 &= 1 + 10^{-4} \end{aligned}$$

Dies formen wir um zu

$$\begin{aligned} x_1 + x_2 &= 2 \\ x_2(1 - 10^{-4}) &= 1 + 10^{-4} - 2 \cdot 10^{-4} \end{aligned}$$

Im Computer liefert die zweite Zeile aufgrund von entsprechenden Rundungen:

$$\begin{aligned} x_2(1 \ominus 10^{-4}) &= 1 \oplus 10^{-4} \ominus 2 \cdot 10^{-4} \\ x_2 \cdot 1 &= 1 \end{aligned}$$

Durch Rückeinsetzen erhalten wir nun auch ganz gewöhnlich $x_1 = 1$.

Weiterhin halten wir fest:

6) Wir führen eine Fehleranalyse durch. Wir erinnern uns, dass der unvermeidbare Fehler abgeschätzt wird durch

$$\frac{\|dx\|}{\|x\|} \leq \frac{\|A\|\|A^{-1}\|}{1-q} \left(\frac{\|dA\|}{\|A\|} + \frac{\|db\|}{\|b\|} \right)$$

Die letzte Gleichung hat die Form $(A^{(n)})_{n,n}x_n = (b^{(n)})_n$. Die erste Zeile lautet

$$(A^{(n)})_{1,1}x_1 + \ldots + (A^{(n)})_{1,n}x_n = (b^{(n)})_1$$

Zu diesem Zeitpunkt sind $x_2, \ldots, x_n$ bereits bekannt und wir erhalten

$$x_1 = \frac{1}{(A^{(n)})_{1,1}} \left((b^{(n)})_1 - (A^{(n)})_{1,2}x_2 - \ldots - (A^{(n)})_{1,n}x_n \right)$$

Wir wollen Auslöschung vermeiden und wollen daher, dass der Term in den Klammern möglichst groß im Betrag ist. Wir wissen, dass der Klammerterm genau $(A^{(n)})_{1,1}x_1$ ist. Eine Idee wäre es, die Anordnung der Gleichungen so zu wählen, dass $(A^{(n)})_{1,1}$ betraglich möglichst groß ist.

Kapitel 11

Unter- und überbestimmte lineare Gleichungssysteme

Wir haben bisher invertierbare, d.h. insbesondere quadratische Matrizen betrachtet. Dies ist jedoch nicht trivialerweise der Fall. Wir untersuchen dies hier näher: Sei $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$. Zu lösen sei $Ax = b$ mit Lösung $x \in \mathbb{R}^n$. Dann können drei Fälle eintreten:

- $m > n$: Das LGS ist überbestimmt.
- $m < n$: Das LGS ist unterbestimmt.
- $m = n$: Hat das LGS nun zusätzlich vollen Rang, so ist das LGS eindeutig lösbar.

In der Anwendung treten diese Fälle durchaus auf. Ein Ingenieur sucht beispielsweise drei Variablen x_1, x_2, x_3. Er nimmt zunächst eine Messung vor, die liefert:

$$x_1 + 2x_2 + 3x_3 = 5$$

Es gibt einen zweidimensionalen Lösungsraum. Der Ingenieur jedoch braucht eine konkrete (sachbezogene) Lösung. Ihm hilft es dabei nicht, einfach irgendeine Lösung aus dem Lösungsraum zu konkretisieren. Es ist demnach gefragt, aus dem groen Vorrat an Lösungen eine auszuzeichnen. Um voranzukommen, misst der Ingenieur unter gleichen Parametern noch einmal und erhält:

$$x_1 + 2x_2 + 3x_3 = 6$$

Offensichtlich widersprechen sich die beiden Gleichungen. Dies ist darauf zurückzuführen, dass der Ingenieur sich irgendwo vermessen hat. Aus der Sicht der theoretischen Mathematik steht man nun in einer Sackgasse, denn hier gibt es keine exakte Lösung. Die Aufgabe der Numerik besteht darin, einen Lösungserthm zu finden. Misst der Ingenieur nun unter anderen Einstellungen drei weitere Male, so werden wir in der Regel dennoch kein LGS schaffen, dass eine (theoretische) Lösung besitzt. Man sieht auch: Es ist ganz natürlich, dass mehr Messwerte vorhanden sind, als für eine analytische Lösung nötig sind.

Beispiel 11.1 (Polynominterpolation)**.** Sei $\mathbb{P}_n$ der Vektorraum der Polynome von Grad kleiner gleich n. Gegeben seien Messwerte $y_k \in \mathbb{R}$ zu Messpunkten t_k mit $k = 1, \dots m$ und t_k

paarweise verschieden. Wir haben also eine Wertetabelle, die man in einem Koordinatensystem als Punkte eintragen kkann. Wir suchen nun $p \in \mathbb{P}_l$ mit $p(t_j) = y_j$ für $j = 1, \ldots, m$. Allgemein ist

$$p(x) = a_0 + a_1 x + \ldots + a_l x^l \ ,$$

d.h. äquivalent zur Suche von p können wir die Koeffizienten a_i bestimmen für $i = 1, \ldots, l$, sodass

$$\begin{aligned} y_1 = p(t_1) &= a_0 + t_1 a_1 + \ldots + t_1^l a_l \\ &\vdots \\ y_m = p(t_m) &= a_0 + t_m a_1 + \ldots + t_m^l a_l \end{aligned}$$

Wir können dieses Gleichungssystem in eine Matrix fassen:

$$\begin{pmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^l \\ \vdots & & & & \vdots \\ 1 & t_m & t_m^2 & \ldots & t_m^l \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ \vdots \\ a_l \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$$

Die linke Matrix (im Folgenden A genannt) ist eine reelle $(m \times (l+1))$-Matrix. Matrizen von solcher Form nennen wir Vandermonde-Matrix.

Satz 11.2. *Seien $t_k \in \mathbb{R}$ paarweise verschieden, es seien $y_k \in \mathbb{R}$ mit $k = 1, \ldots, m$.*

1) *Falls $l + 1 = m$, so gibt es genau ein Polynom $p \in \mathbb{P}_l$ mit $p(t_k) = y_k$.*

2) *Falls $m > l + 1$, so gibt es höchstens ein, im Allgemeinen gar kein Polynom $p \in \mathbb{P}_l$ mit $p(t_k) = y_k$.*

3) *Falls $m < l + 1$, so gibt es unendlich viele Polynome $p \in \mathbb{P}_l$ mit $p(t_k) = y_k$.*

Beweis. Sei $l + 1 = m$. Dies ist ohnehin der einzige Fall, in dem wir Invertierbarkeit erwarten können. Nun ist A invertierbar genau dann, wenn $\ker(A) = \{0\}$: In diesem Fall nämlich ist die Matrix A, als lineare Abbildung verstanden, injektiv und somit auch bijektiv. Ist nun $(a_0 \quad \ldots \quad a_l)^T \in \ker(A)$, so gilt

$$A \cdot \begin{pmatrix} a_0 \\ \vdots \\ a_l \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

und somit $p(t_1) = 0$, $p(t_2) = 0$ und so fort bis $p(t_m) = 0$. Es ergibt sich das Polynom $p(x) = a_0 + a_1 x + \ldots + a_l x^l$. Nun ist $p \in \mathbb{P}_l$, jedoch hat dieses offensichtlich $m = l+1$ Nullstellen. Das Polynom hat somit mehr Nullstellen als möglich. Daraus folgt, dass das gesuchte Polynom gerade das Nullpolynom ist, denn somit $\ker(A) = \{0\}$. Damit ist A invertierbar, daher lässt sich genau eine Lösung finden für die Koeffizienten des Polynoms. Somit gibt es ein eindeutiges Interpolationspolynom. □

Bemerkung 11.3. Für $l = 1$ und $m = 2$ bedeutet dies: Kennen wir zwei Punkte in einem Koordinatensystem, so gibt es genau ein lineares Polynom (d.h. eine Gerade), die durch diese beiden Punkte läuft. Füge ich einen weiteren Messpunkt hinzu, so gibt es genau eine quadratische Funktion durch diese drei Punkte. Fügt man noch mehr Punkte hinzu, so gibt es nur noch zwei Fälle: Die neuen Punkte liegen bereits auf dem interpolierten Polynom oder eben nicht - dann haben wir ein Problem.

Im Allgemeinen sagt der obige thm: Haben wir zu viele Gleichungen, so finden wir in der Regel keine Lösung – haben wir zu wenige, so haben wir zu viel Auswahl.

Beispiel 11.4 (Lineare Interpolation/Regression)**.** Wir suchen ein $p \in \mathbb{P}_1$, d.h. wir suchen eine lineare Funktion mit $p(t_k) = y_k$ mit $k = 1, \ldots, l$. Wir nehmen an, dass $m \geq 2$, d.h. wir haben zumindest ausreichend viele Messwerte. Es wäre nun naiv, direkt eine Gerade durch die ersten beiden Messwerte zu legen, denn dann trifft man in der Regel viele Messpunkte nicht. Sinnvoller wäre wohl eher eine Ausgleichsgerade. Wir müssen uns daher über Gütekriterien Gedanken machen, die eine solche Gerade erfüllen sollte. Statt $Ax = b$ zu lösen verfahren wir daher wie folgt: Suche $x \in \mathbb{R}^n$, sodass $\|Ax - b\|_2$ möglichst klein wird. Da es nicht relevant ist, in welcher Form wir die Norm klein machen, betrachten wir fortan $\|Ax - b\|_2^2$. Eine solche Lösung nannte schon Gau eine **kleinste Quadrate-Lösung**.

Definition 11.5. Sei $A \in \mathbb{R}^{m \times n}$, sei $b \in \mathbb{R}^m$. Ein $\overline{x} \in \mathbb{R}^n$ heit kleinste Quadrate-Lösung von $Ax = b$ (least squares) genau dann, wenn

$$\|A\overline{x} - b\|_2^2 \leq \|Ax - b\|_2^2$$

für alle $x \in \mathbb{R}^n$.

Bemerkung 11.6. Es handelt sich hier zunächst um eine Verallgemeinerung eines Lösungsbegriffs. Falls $m = n$ und A invertierbar, so ist $x = A^{-1}b$ eine kleinste Quadrate-Lösung (d.h. der alte Lösungsbegriff bleibt bestehen). In diesem Fall ist die Lösung auch eindeutig.

Mit diesen neuen Begriffen können wir dem Ingenieur aus der Einleitung in dieses Kapitel nun eine Lösung anbieten, indem wir ihm eine Gerade liefern, die möglichst kleinen Abstand zu allen Punkten hat. Problematisch bleibt die Frage, wie man das eigentlich ausrechnen soll. Wie lässt sich somit die Lösung $\overline{x}$ charakterisieren?

Ein kurzer Einschub:

$$\begin{aligned} \ker(A) &= \{x \in \mathbb{R}^n \mid Ax = 0\} \\ \mathrm{Bild}(A) &= \{Ax \mid x \in \mathbb{R}^n\} \\ \mathrm{rang}(A) &= \dim \mathrm{Bild}(A) \\ \mathrm{rang}(A) &= \mathrm{rang}(A^T) \\ \dim \ker(A) + \dim \mathrm{Bild}(A) &= n \end{aligned}$$

Lemma 11.7. *Sei $A \in \mathbb{R}^{m \times n}$.*

1) $\mathbb{R}^m = \mathrm{Bild}(A) \oplus \ker(A^T)$.

2) $\ker(A^T A) = \ker(A)$.

Beweis. Sei $x \in \ker(A^T)$. Es ist $A^T \in \mathbb{R}^{n\times m}$, d.h. $x \in \mathbb{R}^m$. Sei $z \in \text{Bild}(A)$. Es gibt somit ein y mit $z = Ay$. Nun ist

$$(z, x) = (Ay, x) = (y, A^T x) = (y, 0) = 0$$

und somit sind z und x orthogonal. Somit folgt weiter $\ker(A^T) \subset \text{Bild}(A)^\perp$. Sei nun $x \in \text{Bild}(A)^\perp$. Dann ist insbesondere $(x, AA^T x) = 0$, da $AA^T x \in \text{Bild}(A)$. Nun gilt

$$(x, AA^T x) = (A^T x, A^T x) = \|A^T x\|_2^2 = 0$$

und somit ist bereits $A^T x = 0$. Daraus folgt $x \in \ker(A^T)$, d.h. $\text{Bild}(A)^\perp \subset \ker(A^T)$. Somit folgt $\text{Bild}(A)^\perp = \ker(A^T)$. Damit folgt

$$\mathbb{R}^m = \text{Bild}(A) \oplus \ker(A^T) = \text{Bild}(A) \oplus \text{Bild}(A)^\perp \ .$$

Der zweite Teil wurde schon in den bungen bewiesen. □

Satz 11.8. *Ein $x \in \mathbb{R}^n$ ist genau dann eine kleinste Quadrate-Lösung von $Ax = b$, falls $A^t Ax = A^t b$.*

Die Gleichung $A^T Ax = A^T b$ heit Gausche Normalgleichung.

Beweis. Wir treffen für den Beweis mit Hin- und Rückrichtung zunächst einige Vorbereitungen: Sei $Ax = b$ gegeben, schreibe $b = b_1 + b_2$ mit $b_1 \in \text{Bild}(A)$ und $b_2 \in \text{Bild}(A)^\perp = \ker(A^T)$. Sei $y \in \mathbb{R}^n$. Betrachte nun

$$\|Ay - b\|_2^2 = \|Ay - b_1 - b_2\|_2^2 = \|Ay - b_1\|_2^2 + \|b_2\|_2^2$$

mit $Ay - b_1 \in \text{Bild}(A)$ und $b_2 \in \text{Bild}(A)^\perp$. **Zentrales Argument** für die zweite Gleichheit ist der thm des Pythagoras: Die Vektoren $Ay - b_1$ und b_2 stehen senkrecht zueinander und bilden mit $Ay - b_1 - b_2$ als Hypotenuse ein rechtwinkliges Dreieck. Wir notieren: $\|Ay - b\|_2^2$ ist genau dann minimal, wenn $Ay = b_1$ $(\star)$.
„$\Rightarrow$“ Wir können wegen $b_1 \in \text{Bild}(A)$ sicher sein, dass $Ay = b_1$ eine Lösung hat. Sei ohne Einschränkung $Ay = b_1$ eine solche Lösung, dann folgt

$$A^T b = A^T b_1 + A^T b_2 = A^T Ay$$

wegen $A^T b_2 = 0$, da $b_2 \in \ker(A^T)$.

„$\Leftarrow$“ Sei $z \in \mathbb{R}^n$ mit $A^T Az = A^T b$. Setze weiter $Ay = b_1$, dann folgt, dass $A^T Ay = A^T b_1$. Es gilt nun

$$A^T A(z - y) = A^T A(b - b) = 0$$

und so folgt $(z - y) \in \ker(A^T A)$ und somit nach obigem Lemma $(z - y) \in \ker(A)$. Damit haben wir

$$A(z - y) = 0 \ \Leftrightarrow \ Az = Ay = b_1 \ .$$

Mit $(\star)$ folgt die Behauptung. □

Bemerkung 11.9. Wir halten fest:

1) Jede Gleichung $Ax = b$ besitzt eine kleinste Quadrate-Lösung.

2) Seien x, z kleinste Quadrate-Lösungen. Dann

$$A^T Ax = A^T b = A^T Az$$

und somit $A^T A(x - z) = 0$, also auch $A(z - x) = 0$. Somit unterscheiden sich z und x nur um ein $w \in \ker(A)$, d.h. $z = x + w$.

3) Ist $A = 0$, so sind alle $x \in \mathbb{R}^n$ eine kleinste Quadrate-Lösung.

Beispiel 11.10. Zu messen sei die Länge L. Wir bekommen Messwerte $l_1, \ldots, l_5$. Diese übersetzen wir in ein LGS:

$$\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot L = \begin{pmatrix} l_1 \\ \vdots \\ l_5 \end{pmatrix}$$

und setzen $A = (1 \quad \ldots \quad 1)^T$ und $b = (l_1 \quad \ldots \quad b_5)^T$. Nun muss $A^T AL = A^T b$ von L erfüllt werden. Wir erhalten somit die Bedingung:

$$\begin{pmatrix} 1 & \ldots & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot L = \begin{pmatrix} 1 & \ldots & 1 \end{pmatrix} \cdot \begin{pmatrix} l_1 \\ \vdots \\ l_5 \end{pmatrix}$$

und somit $5L = l_1 + \ldots + l_5$. Stellt man dies äquivalent um, so erhält man

$$L = \frac{l_1 + l_2 + l_3 + l_4 + l_5}{5}$$

den Mittelwert als beste Lösung für $n = 5$ Messungen.

Beispiel 11.11. Gegeben haben wir die Messungen

t_i	-2	0	1	1
y_i	-2	-4	4	6

und wir vermuten, dass $y(t) = at + b$ die Situation abbildet. Wir wollen nun den Unterschied zwischen der Gerade und den Punkten minimieren. Gesucht sind also a und b, sodass

$$Ax := \begin{pmatrix} -2 & 1 \\ 0 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -2 \\ -4 \\ 4 \\ 6 \end{pmatrix} =: c$$

erfüllt ist. Wir fordern nun wieder $A^T A \cdot (a \; b)^T = A^T c$, d.h.:

$$\begin{pmatrix} -2 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} -2 & 1 \\ 0 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -2 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} -2 \\ -4 \\ 4 \\ 6 \end{pmatrix}$$

Wir gewinnen daraus das einfache LGS

$$\begin{pmatrix} 6 & 0 \\ 0 & 4 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 14 \\ 4 \end{pmatrix}$$

und lösen dieses auf zu $b = 1$ und $a = 7/3$.

Bemerkung 11.12. Achtung: Die kleinste Quadrate-Lösung ist im Allgemeinen nicht eindeutig, sondern nur genau dann, wenn A injektiv ist. **Beweis:** Sind x, z kleinste Quadrate-Lösungen, so ist $(x - z) \in \ker(A^T A)$ wegen $A^T A(x - z) = A^T A(b - b) = 0$. Wir wissen aber auch, dass $(x - z) \in \ker(A)$. Gilt nun für alle solche Lösungen die Eindeutigkeit $x = z$, so ist dies äquivalent zu $x - z = 0$. Dies bedeutet dann nichts anderes, als dass $\ker(A) = \{0\}$, d.h. der Kern besteht nur aus Nullelementen. Dies ist genau dann der Fall, wenn A injektiv ist.

Problem: Welche Lösung wählen wir aus dem affinen Unterraum der kleinsten Quadrate-Lösungen aus? Der affine Unterraum besteht aus allen Punkten auf der Geraden. Wir wollen den Punkt mit der kleinsten Norm betrachten.

Definition 11.13. Sei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. Wir nennen $x^+ \in \mathbb{R}^n$ Minimum-Norm-Lösung von $Ax = b$ genau dann, wenn

1) x^+ kleineste Quadrate-Lösung ist und

2) $\|x^+\| \leq \|y\|$ für alle kleinste Quadrate-Lösungen $y \in \mathbb{R}^n$ erfüllt ist.

Satz 11.14. *x^+ ist genau dann Minimum-Norm-Lösung von $Ax = b$, falls*

1) $A^T A x^+ = A^T b$.

2) $x^+ \in \text{Bild}(A^T)$.

Beweis. Sei $\overline{x}$ eine kleinste Quadrate-Lösung (die Existenz ist bereits bewiesen). Schreibe $\overline{x} = x^+ \oplus y$ mit $x^+ \in \ker(A)^\perp = \text{Bild}(A^T)$ und $y \in \ker(A)$. Es ist x^+ eine kleinste Quadrate-Lösung, denn:

$$A^T A x^+ = A^T A(\overline{x} - y) = A^T A\overline{x} - A^T A y = A^T b$$

und $A^T A\overline{x} = A^T b$, da $\overline{x}$ kleinste Quadrate-Lösung ist, und $A^T A y = 0$ wegen $y \in \ker(A)$. Ist y eine kleinste Quadrate-Lösung, so ist $x^+ - y \in \ker(A)$ und

$$\|y\|_2^2 = \|x^+ + (y - x^+)\|_2^2 \, .$$

Wir betrachten in obiger Norm die Summe von zwei senkrecht aufeinander stehenden Vektoren und wissen nach Pythagoras (vgl. Beweis zur Gau-Normalgleichung), dass:

$$\|x^+ + (y - x^+)\|_2^2 = \|x^+\|_2^2 + \|y - x^+\|_2^2 \geq \|x^+\|_2^2$$

Somit ist x^+ klarerweise eine Minimum-Norm-Lösung. □

Wir geben noch zwei Beispiele an, um die bisherigen Begriffe noch zusätzlich zu motivieren.

Beispiel 11.15. Gegeben sei $A \in \mathbb{R}^{2\times 1}$ mit $A = (1 \quad -1)^T$. Klar ist: Diese Matrix definiert eine lineare Abbildung $A : \mathbb{R} \to \mathbb{R}^2$. Wir erhalten:

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix} x = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

für $b_1 = 1$ und $b_2 = 0$ erhalten wir zwei sich widersprechende Gleichungen und somit keine Lösung. Betrachte $\text{Bild}(A) = \{(x \quad -x)^T \mid x \in \mathbb{R}\}$. Nun gilt $\text{Bild}(A)^\perp = \ker(A^T)$. Ich beschreibe nun $b = b_1 + b_2$ mit $b_1 \in \text{Bild}(A)$ und $b_2 \in \text{Bild}(A)^\perp = \ker(A^T)$. Die Idee ist nun: Löse $Ax = b_1$. Das ist auf jeden Fall lösbar, und somit eine kleinste Quadrate-Lösung. Wir haben damit unseren bisherigen Lösungsbegriff verallgemeinert.

Beispiel 11.16. Sei $A \in \mathbb{R}^{1\times 2}$ mit $A = (1 \quad 1)$. Betrachte nun

$$(1 \quad 1) \cdot \begin{pmatrix} x \\ y \end{pmatrix} = b := 1$$

Somit haben wir vermittels $x + y = 1$ viele Lösungen. Wir betrachten daher den Quellraum: Die Gerade $y = 1 - x$. Wir wollen nun die Gleichung mit der kleinsten Norm auswählen. Alle kleinste Quadrate-Lösungen unterscheiden sich nur durch ein Element aus $\ker(A)$, d.h. $x + y = 0$ bzw. $x = -y$. Es ist nun wieder $\ker(A)^\perp = \text{Bild}(A^T)$ und $A^T\alpha = (\alpha \quad \alpha)^T$. Die Idee ist nun: Wähle x so, dass $x \in \ker(A)^\perp = \text{Bild}(A^T)$, um die Norm von x minimal zu halten.

Sei $A \in \mathbb{R}^{m\times n}$. Wir erinnern an $\text{rang}(A) = \dim \text{Bild}(A)$. Der Dimensionsthm besagt:

$$\dim \text{Bild}(A) + \dim \ker(A) = n \ .$$

Damit ist direkt klar, dass $\text{rang}(A) \leq n$. Analog folgt $\text{rang}(A^T) \leq m$. Klar ist auch, dass $\text{rang}(A) \leq \min\{n, m\}$. Falls $\text{rang}(A) = \min\{n, m\}$, dann hat A vollen Rang. Es gilt weiter

$$\begin{aligned} \text{rang}(A^TA) &= n - \dim \ker(A^TA) \\ &= n - \dim \ker(A) \\ &= \dim \text{Bild}(A) = \text{rang}(A) \\ &= \text{rang}(A^T) = \text{rang}(AA^T) \end{aligned}$$

Definition 11.17 (Pseudoinverse, Moore-Penrose-Inverse)**.** Sei $A \in \mathbb{R}^{m\times n}$. Sei $A^+ : \mathbb{R}^m \to \mathbb{R}^n$ eine Abbildung. Wir definieren $A^+b := x^+$, wobei x^+ die Minimum-Norm-Lösung von $Ax = b$ ist. Dann ist A^+ wohldefiniert und linear.

Satz 11.18. *Sei $A \in \mathbb{R}^{m\times n}$, A habe vollen Rang.*

1) *Falls $n = m$, so ist A invertierbar.*
2) *Falls $m < n$, so ist A surjektiv. In diesem Fall $A^+ = A^T(AA^T)^{-1}$.*
3) *Falls $m > n$, so ist A injektiv. In diesem Fall $A^+ = (A^TA)^{-1}A^T$.*

Beweis. Wir betrachten die Fälle:

1) Aufgrund der Invertierbarkeit existiert eine eindeutige kleinste Quadrate-Lösung, die dann auch Minimum-Norm-Lösung ist.

2) Sei $m < n$. Dann ist $\dim \text{Bild}(A) = m$. Somit ist $\text{Bild}(A) = \mathbb{R}^m$ und daher A surjektiv. Die Gleichung $Ax = b$ ist lösbar für alle $b \in \mathbb{R}^m$, es entstehen also kleinste Quadrate-Lösungen. Wir suchen nun ein $x^+ \in \text{Bild}(A^T)$ mit $Ax^+ = b$. Wir schreiben nun $x^+ = A^T y$ für ein $y \in \mathbb{R}^n$. Somit $AA^T y = b$. Dieses y ist sogar eindeutig bestimmt, denn AA^T ist invertierbar. Es ist nämlich $\text{rang}(A^T) = \text{rang}(A) = m$ und somit A^T injektiv, d.h. $\dim \ker(A^T) = 0$. Damit ist $\dim \ker(AA^T) = 0$ und als quadratische Matrix ist AA^T direkt surjektiv und auch invertierbar. Wir berechnen somit

$$y = (AA^T)^{-1} b$$

und folgern mittels $x^+ = A^T(AA^T)^{-1}b$:

$$A^+ b = A^T(AA^T))^{-1} b \ .$$

3) Sei $m > n$. Somit $\text{rang}(A) = \dim \text{Bild}(A) = n$. Damit ist insbesondere $\dim \ker(A) = n - \dim \text{Bild}(A) = 0$, somit ist A injektiv. Wegen $\ker(A) = \ker(A^T A)$ ist auch $A^T A$ injektiv, und somit ist $A^T A$ auch surjektiv, da $A^T A \in \mathbb{R}^{n \times n}$ quadratische Matrix ist. Somit ist $A^T A$ invertierbar. Die Gleichung $A^T Ax = A^T b$ hat nur eine Lösung, nämlich $x = (A^T A)^{-1} A^T b$. Somit gibt es nur eine kleinste Quadrate-Lösung, welche zugleich Minimum-Norm-Lösung ist. Somit ist

$$A^+ b = x^+ = (A^T A)^{-1} A^T b \ .$$

□

Bei der Berechnung der Minimum-Norm-Lösung müssen wir drei Fälle unterscheiden. Sei dazu $A \in \mathbb{R}^{m \times n}$ und A habe vollen Rang.

1) $m > n$. Dann ist $A^T A \in \mathbb{R}^{n \times n}$, zusätzlich haben wir die Normalgleichung $A^T Ax^+ = A^T b$. Idee:

   - Berechne $A^T A$ und $A^T b$.
   - Berechne die Lösung mit einem beliebigen Verfahren.

   Dabei ergibt sich ein Problem, nämlich der Berechnungsfehler. Der erwartete Fehler ist proportional zu $K(A^T A)$. Im Allgemeinen ist $K(A^T A) \sim K(A)^2$. Für Matrizen mit hoher Kondition $K(A)$ ist der Fehler also sehr hoch. Wir sollten wenn möglich also die Berechnung von $A^T A$ vermeiden, eine Möglichkeit dazu ist die sogenannte QR-Zerlegung, bei der man ähnlich wie bei der LR-Zerlegung vorgeht, aber die Frobenius-Matrizen durch orthogonale Matrizen ersetzt, die jeweils die erste Spalte der Restmatrix auf ein Vielfaches des ersten Einheitsvektors dreht oder spiegelt. Da die Multiplikation mit Orthogonalen Matrizen die Norm erhält, gilt

   $$\|Ax - b\| = \|QRx - b\| = |Q(Rx - Q^T b)\| = |Rx - Q^T b\|,$$

   sodass wir am Ende auch die kleinst-quadrate Lösung eines Systems in Rechtecksform berechnen müssen.

2) $n > m$. Dann ist $x = A^T y$, somit gilt $AA^T y = b$, wenn AA^T invertierbar ist, folgt also $x^+ = A^T y = A^T(AA^T)^{-1} b$. Zur Lösung können wir dann folgendermaßen vorgehen:

- Berechne AA^T .
- Berechne die Lösung y mit einem beliebigen Verfahren.
- Berechne $x = A^T y$

Kapitel 12

Interpolation

Das grundsätzliche Problem der Interpolation ist die Berechnung einer Interpolationsfunktion $g_n : [x_0, x_n] \to \mathbb{R}$ bei gegebenen Werten $y_i = f(x_i)$, $i = 0, \ldots, n$. Der Einfachheit halber nehmen wir an $x_0 < x_1 < \ldots x_n$. Nun gibt es unendlich viele Funktionen g_n, die diese Eigenschaft erfüllen, die Frage ist aber wie gut werden diese die Funktion f approximieren im gesamten Intervall $[x_0, x_n]$. Um eine Eindeutigkeit zu erzwingen schränkt man üblicherweise g_n auf eine Klasse von Funktionen mit nur $n + 1$ unbekannten Parametern ein, also gleich vielen wie die Anzahl der Stützstellen. Der klassische Fall ist die Polynominterpolation aus dem letzten Kapitel. Gegeben x_i, y_i für $i = 0, \ldots, n$. Dann existiert ein Polynom vom Grad kleiner gleich n, $g_n \in \mathbb{P}_n$ mit $p(x_i) = y_i$. Schreiben wir

$$g_n(x) = \sum_{j=0}^{n} a_j x^j,$$

so erhalten wir das $(n + 1) \times (n + 1)$ Gleichungssystem

$$\sum_{j=0}^{n} a_j x_i^j = y_i,$$

in Matrix-Form

$$V_n a = y,$$

mit der Vandermonde-Matrix V_n. Damit ist das Polynom charakterisiert durch die Berechnung seiner Koeffizienten a_j.

Es bleibt die Frage wie gut g_n die Funktion f zwischen den Stützstellen approximiert. Dazu müssen wir Annahmen an die Funktion f machen, etwa über ihre Ableitungen. Betrachten wir dazu ein Beispiel: Die Konstruktion einer Achterbahn. Wichtig ist hier die Beachtung der Fliehkraft, welche die zweite Ableitung $f''(x)$ der Streckenverlaufsfunktion $h(x)$ der Achterbahn ist. Man sollte nun unbedingt fordern, dass die zweite Ableitung stetig ist, denn die Fliehkraft ist nichts anderes als die Beschleunigung. Ist diese nicht stetig, so bedeutet dies in der Sachsituation: Die Geschwindigkeit ändert sich abrupt vom Positiven zum Negativen, und dies ist für den Magen der Fahrgäste absolut nicht verträglich.

Um den Fehler zwischen der Funktion f und dem Polynom g_n zu verstehen betrachten wir zunächst den Fall $n = 1$ mit f zweimal stetig differenzierbar. Die lineare Interpolierende können wir schreiben als

$$g_1(x) = \frac{x - x_0}{x_1 - x_0} y_1 + \frac{x_1 - x}{x_1 - x_0} y_0 = \frac{x - x_0}{x_1 - x_0} f(x_1) + \frac{x_1 - x}{x_1 - x_0} f(x_0).$$

Nun gilt aus der Taylor-Entwicklung

$$f(x) = f(x_0)+f'(x_0)(x-x_0)+\frac{1}{2}f''(\xi_0)(x-x_0), \qquad f(x) = f(x_1)+f'(x_1)(x-x_1)+\frac{1}{2}f''(\xi_1)(x-x_1),$$

mit Zwischenstellen $\xi_0, \xi_1 \in [x_0, x_1]$. Lösen wir diese Gleichungen nach $f(x_0)$ bzw. $f(x_1)$ auf und setzen diese in die Formel für g_1 ein so folgt

$$g_1(x) = f(x) + \frac{(x-x_0)(x_1-x)}{x_1-x_0}(f'(x_1) - f'(x_0)) + R(x),$$

wobei $|R(x)| \leq C|x_1 - x_0|^2$, mit C abhängig von der zweiten Ableitung von f. Da nun auch gilt

$$f'(x_1) - f'(x_0) = f''(\xi_2)(x_1 - x_0)$$

mit einer weiteren Zwischenstelle ξ_2. Damit ist auch $\frac{(x-x_0)(x_1-x)}{x_1-x_0}(f'(x_1) - f'(x_0))$ genauso abschätzbar wir $R(x)$. Insgesamt gilt also

$$|g_1(x) - f(x)| \leq \tilde{C}|x_1 - x_0|^2,$$

mit $\tilde{C}$ abhängig von der zweiten Ableitung von f. Wenn der Abstand zwischen den Stützstellen klein genug ist wird auch der Fehler zwischen g_1 und f an allen Stellen x klein.

Allgemein gilt folgendes Resultat:

Satz 12.1. *Sei $f \in C^{n+1}(\mathbb{R})$, d.h. $f : \mathbb{R} \to \mathbb{R}$ ist $(n+1)$-mal stetig differenzierbar. Seien x_i paarweise verschieden und sei*

$$g_n(x_i) = f(x_i)$$

für $i = 0, \ldots, n$. Dabei ist $g_n \in \mathbb{P}_n$ das Interpolationspolynom zur Funktion f an den Stellen x_i. Dann gibt es für alle $\overline{x}$ ein ξ mit

$$f(\overline{x}) = g_n(\overline{x}) = w(\overline{x})\frac{f^{(n+1)}(\xi)}{(n+1)!}$$

mit $w(x) = (x - x_0) \cdot \ldots \cdot (x - x_n)$. Insbesondere gilt

$$\|f - g_n\|_\infty \leq \frac{\|w\|_\infty}{(n+1)!} \cdot \|f^{(n+1)}\|_\infty$$

bezogen auf ein Intervall $[a, b]$.

Man könnte den Verdacht haben, dass mehr Stützwerte auch bessere Ergebnisse liefern. Dies ist im Allgemeinen falsch, wie die Runge-Funktion

$$f(x) = \frac{1}{1 + 25x^2}$$

zeigt. Erhöhen wir hier die Anzahl der Messwerte, so wird das Ergebnis immer schlechter. Insbesondere passiert dies, wenn die Messwert y_i auch Fehler enthalten, also $y_i = f(x_i) + \epsilon_i$ mit $|\epsilon_i|$ klein. Sei g_n das Interpolationspolynom mit den Daten $f(x_i)$ und $\tilde{g}_n$ jenes mit den gestörten Daten y, dann gilt

$$f(x) - \tilde{g}_n(x) = f(x) - g_n(x) + g_n(x) - \tilde{g}_n(x),$$

und mit der Dreiecksungleichung

$$|f(x) - \tilde{g}_n(x)| \leq |f(x) - g_n(x)| + |g_n(x) - \tilde{g}_n(x)|.$$

Der Fehler kann also abgeschätzt werden durch den Fehler bei exakten Daten und den Fehler zwischen den beiden Interpolationspolynomen. Letzterer ergibt sich dann aus dem Fehler in den Koeffizienten a, für die wir ja oben ein lineares Gleichungssystem hergeleitet haben. Wie wir in Kapitel 10 gesehen haben, ist die Fortpflanzung des Fehlers bei linearen Gleichungssystemen im wesentlichen durch die Konditionszahl der Matrix bestimmt, diese wird bei der Vandermonde-Matrix mit größerem n immer schlechter. Deshalb wird die Polynominterpolation mit steigendem n immer problematischer.

Um die guten Eigenschaften bei kleinem Polynomgrad zu retten und trotzdem eine gute Approximation zu erhalten, stückelt man einfach Polynome zwischen den Stützstellen zusammen, die sogenannten Splines:

Definition 12.2. Seien $s_0 < s_1 < \ldots < s_n$ reelle Zahlen. Eine Funktion $s : [s_0, s_n] \to \mathbb{R}$ heißt Spline mit den Knoten $s_0, \ldots, s_n$, falls:

1) $s \in C^{k-2}([s_0, s_n])$ für $k > 1$.

2) $s([s_i, s_{i+1}]) \in \P_{k-1}$.

Ist dies der Fall, so ist s von der Ordnung k.

Bemerkung 12.3. Ist $k = 2$, so betrachten wir einen Polygonzug. Ist $k = 4$, so interpolieren wir durch kubische Polynome, d.h. kubische Splines.

Der einfachste Fall ist der konstante Spline, d.h. zwischen einzelnen Stützstellen interpoliere ich durch waagerechte Geraden. Dies ist nicht besonders sinnvoll, da die resultierende Funktion in der Regel nicht stetig ist. Besser hingegen sind lineare Splines, d.h. ich verbinde alle Stützpunkte durch Streckenzüge. Diese sind immer noch nicht differenzierbar. Eine deutlich bessere Annäherung sind quadratische Splines, die uns sogar eine insgesamt differenzierbare Funktion erlauben. Optimal sind kubische Splines, die auch in realen Anwendungen zur Geltung kommen.

Bemerkung 12.4. Wir bezahlen die hohe Güte der Interpolation damit, dass die resultierende Interpolationsfunktion nicht unbegrenzt oft differenzierbar ist.

Definition 12.5. Ein kubischer Spline s zu den Knoten $x_0, \ldots, x_n$ und Interpolationspunkten $y_0, \ldots, y_n$ heißt natürlich, falls

$$s''(x_0) = s''(x_n) = 0 \ .$$

Kurzer Exkurs: Das englische Wort *spline* bedeutet übersetzt so viel wie „Straklatte“. Straklatten sind elastische Bretter, die im Schiffbau verwendet werden. Dort ist es praktisch, wenn diese Bretter an den Enden möglichst wenig gebogen werden – ansonsten baut sich starker Druck auf die ganze Konstruktion auf. Wir interpretieren daher nun die zweite Ableitung als Biegeenergie eines Brettes. Dann bedeutet $f'' = 0$ anschaulich, dass das Brett an einer bestimmten Stelle gerade nicht gebogen ist. Sei

$$S := \{f \mid f(x_i) = y_i, f \in C^2([x_0, x_n]), f \in C^\infty((x_i, x_{i+1}))\}$$

die Menge aller Funktionen, die überhaupt für unsere Interpolation infrage kommen. Wir definieren nun die Biegeenergie

$$E(f) := \int_{x_0}^{x_n} f''(x)^2 \, dx \ .$$

Satz 12.6. *Es sei $f \in S$. Wenn*

$$E(f) \leq E(u)$$

für alle $u \in S$ erfüllt ist, dann ist f ein natürlicher Spline.

Beweis. Sei $u \in S$ beliebig, es sei f wie in der Voraussetzung. Ich betrachte nun

$$F(\lambda) = E(f + \lambda h)$$

mit $h = u - f$. Diese Funktion ist nun garantiert eindimensional und hat ein Minimum bei $\lambda = 0$, d.h. $F'(0) = 0$, denn f ist nach Voraussetzung ein Minimum von E. Es gilt weiter

$$h(x_i) = 0$$

für alle i. Somit ist $h \in C^2([x_0, x_n])$ und $h \in C^\infty((x_i, x_{i+1}))$. Damit haben wir

$$\begin{aligned} E(f + \lambda h) &= \int_{x_0}^{x_n} (f'' + \lambda h'')(x)^2 \, dx \\ &= \int_{x_0}^{x_n} (f'')^2(x) + 2\lambda f''(x)h''(x) + \lambda^2 (h'')^2(x) \, dx \end{aligned}$$

Damit folgt:

$$F'(0) = 2 \cdot \int_{x_0}^{x_n} f''(x)h''(x) \, dx = 0$$

Es läge jetzt nahe, partiell zu integrieren. Dies darf man jedoch nicht, da die Funktionen nur aus C^2 kommen. Gehen wir aber auf die einzelnen Intervalle, so erhalten wir:

$$\begin{aligned} F'(0) &= 2 \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f''(x)h''(x) \, dx \\ &= 2 \sum_{i=0}^{n-1} \left[f''(x)h'(x) \right]_{x_i}^{x_{i+1}} - \int_{x_i}^{x_{i+1}} f'''(x)h'(x) \, dx \end{aligned}$$

Dies wiederholen wir:

$$F'(0) = 2 \sum_{i=0}^{n-1} \left[f''(x)h'(x) \right]_{x_i}^{x_{i+1}} - \left[f'''(x)h(x) \right]_{x_i}^{x_{i+1}} + \int_{x_i}^{x_{i+1}} f''''(x)h(x) \, dx$$

Der Integralterm entfällt nun, da f auf $[x_i, x_{i+1}]$ ein kubisches Polynom ist und somit $f'''' = 0$. Bei den übrigen Termen handelt es sich um eine Teleskopsumme (alle inneren Terme heben sich weg), es bleibt nur stehen:

$$-f''(x_0) \cdot h'(x_0) + f''(x_n) \cdot h'(x_n)$$

und dies ist Null, wenn f ein natürlicher Spline ist. Mit ein bisschen mehr Arbeit kann man zeigen, dass kubische Splines tatsächlich die einzigen sind, die diese Bedingung erfüllen. □

Kapitel 13

Iterative Lösung von Gleichungssystemen

Im Folgenden woll wir iterative Verfahren zur Lösung von Gleichungssystemen betrachten. Dazu schreiben wir das System in der sogenannten Fixpunktform

$$x = g(x)$$

und nehmen dies als Basis für ein Iterationsverfahren. Wir wählen ein $x_0 \in \mathbb{R}$ und berechnen eine Folge $x_{k+1} = g(x_k)$. Wenn g stetigist und $x_k \to \overline{x}$ gilt, erhalten wir im Grenzwert $\overline{x} = g(\overline{x})$. Also können wir mit einer endlichen Anzahl an Iterationen einen Fixpunkt $\overline{x}$ gut annähern.

Es gibt verschiedene Motivationen für die Verwendung eines iterativen Verfahrens:

- Bei großen linearen Gleichungssystemen wie zum Beispiel in der Computertomographie kann es sein, dass direkte Verfahren wie die LR-Zerlegung viel zu großen Rechenaufwand benötigen. Diese nutzen auch nicht, wenn in der Matrix viele Nulleinträge stehen, die Matrizen L und R können trotzdem lauter Nichtnulleinträge haben. Bei einer Matrix mit vielen Nulleinträgen (wie in der Computertomographie für alle Pixel durch die der Röntgenstrahl gerade nicht läuft) ist die Multiplikation der Matrix mit einem Vektor aber sehr effizient, da man die Nullen ja nicht mitrechnen muss.
- Bei nichtlinearen Gleichungen (mit Ausnahme quadratischer und kubischer Gleichungen) ist es schlicht nicht möglich eine Lösung direkt zu berechnen, deshalb muss man zu einem approximativen Verfahren greifen.

13.0.1 Iterative Verfahren für lineare Gleichungssysteme

Wir beginnen mit einigen einfachen Beispielen für lineare Probleme. Wir wollen $Ax = b$ lösen, $A \in \mathbb{R}^{n \times n}$ invertierbar. Dazu machen wir zunächst einfache Äquivalenzumformungen. Aus $Ax = b$ erhalten wir $0 = b - Ax$ und durch Multiplikation mit $\tau in \mathbb{R}$ sowie Addition von x auf beiden Seiten die äquivalente Gleichung

$$x = x + \tau(b - Ax) = (I - \tau A)x + \tau b.$$

Führen wir eine Iteration basierend auf diesem Verfahren durch, so erhalten wir das Richardson-Verfahren

$$x^{k+1} = x^k + \tau(b - Ax^k) = (I - \tau A)x^k + \tau b.$$

Um zu verstehen ob dieses Verfahren konvergiert und wie wir τ wählen müssen, betrachten wir den trivialen Fall $n = 1$. Dann ist A eine relle Zahl und

$$x^{k+1} = (1 - \tau A)x^k + \tau b.$$

Wir sehen, dass das Wachstum von x^k schon im homogenen Fall $b = 0$ wie eine geometrische Folge passiert, d.h. mit $(1-\tau A)^k$. Diese Folge bleibt nur dann beschränkt, wenn $-1 \leq 1-\tau A \leq 1$, d.h. $0 \leq \tau A \leq 2$. Die obere Schranke können wir erfüllen wenn der Absolutwert von τ klein genug ist, dies wird auch bei allgemeinem n so sein. Für die untere Schranke muss aber τ das selbe Vorzeichen wie A haben, dies ist schwieriger für eine Matrix zu verallgemeinern.

In gewisser Weise haben wir beim obigen Iterationsverfahren in jedem Schritt die Lösung eines Systems mit der Matrix A durch die viel einfachere Lösung eines Systems mit der Matrix $\frac{1}{\tau}I$ ersetzt. Die Frage des Vorzeichens von τ ist dann auch eine Frage wie gut diese einfache Matrix die ursprüngliche Matrix A approximiert. Nun können wir weitere iterative Verfahren basierend auf folgender Idee konstruieren: einerseits wollen wir nur Systeme mit einer einfachen Matrix lösen, andererseits soll diese A besser approximieren. Besonders einfach ist die Lösung mit Diagonalmatrizen (Inverse ist Diagonalmatrix mit Kehrwerten) und mit unteren Dreiecksmatrizen (durch Vorwärtseinsetzen). Die besten Diagonal- bzw. Dreiecksmatrizen zur Approximation sind vermutlich jene mit den Einträgen von A. Dies führt auf das Jacobi- und das Gauss-Seidel Verfahren, bei denen wir eine Iteration der Form

$$Mx^{k+1} = Nx^k + b$$

konstruieren. Wir zerlegen dazu $A = L + D + R$ mit linker unterer Dreiecksmatrix L, Diagonalmatrix D und rechter oberere Dreiecksmatrix R.

1) Setze $M := D$ und $N := -(L + R)$. Schreibe

$$Dx^{k+1} = -(L + R)x^k + b \;\Leftrightarrow\; x^{k+1} = D^{-1}(b - (L + R)x^k) \; .$$

Komponentenweise ausgeschrieben heißt das:

$$x_i^{k+1} = \frac{1}{a_{i,i}} \cdot \left(b_i - \sum_{j \neq i} a_{i,j} x_j^k \right)$$

Dieses Verfahren heißt Gesamtschrittverfahren oder auch Jacobiverfahren.

2) Setze $M := D + L$ und $N := -R$. Dann

$$(D + L)x^{k+1} = b - Rx^k \;\Leftrightarrow\; x^{k+1} = (D + L)^{-1} \cdot (b - Rx^k) \; .$$

Komponentenweise hingeschrieben:

$$x_i^{k+1} = \frac{1}{a_{i,i}} \cdot \left(b_i - \sum_{j > i} a_{i,j} x_j^k - \sum_{j < i} a_{i,j} x_j^{k+1} \right)$$

Dieses Verfahren heißt Gauss-Seidel oder Einzelschrittverfahren.

Wir sehen, dass wir diese beiden Verfahren völlig analog durchführen können. Der einzig Unterschied ist der Update der Variablen. Der neue Wert von x_1 berechnet sich gleich, wenn wir nun x_2 berechnen nehmen wir bei Jacobi den Wert von x_1 aus der letzten Iteration, bei Gauss-Seidel aber schon den neu berechneten aus der aktuellen Iteration. Dies wird analog für die weiteren x_i durchgeführt. Das Gauss-Seidel Verfahren benötigt weniger Speicher, da wir nur eine Kopie des Vektors x speichern müssen. Sobald wir ein neues x_i berechnet haben, kommt der Wert aus der letzten Iteration nicht mehr vor und kann aus dem Speicher gelöscht werden. Beim Jacobi-Verfahren benötigt man immer zwei Kopien des Vektors x für die letzte und die neue Iteration. Dafür kann das Jacobi-Verfahren leichter auf Parallelrechnern durchgeführt werden, da im Prinzip alle x_i gleichzeitig neu berechnet werden können. Dies ist beim Gauss-Seidel Verfahren nicht der Fall, hier ist die aufsteigende Reihenfolge und eine Berechnung nacheinander zwingend.

Als letztes Verfahren für lineare Systeme, das auch für Kleinstquadrate-Probleme funktioniert, betrachten wir noch das Richardson-Verfahren für die Normalengleichung, also

$$x^{k+1} = x^k + \tau A^T(b - Ax^k) = (I - \tau A^T A)x^k + \tau A^T b.$$

Hier haben wir den Vorteil, dass die Matrix $A^T A$ positiv definit ist. Dies bedeutet $A^T A$ hat nur positive Eigenwerte λ und die Eigenwerte von $I - \tau A^T A$ liegen für A positiv sogar im Intervall $(1 - \tau\lambda_{\max}, 1)$, wobei $\lambda_{\max}$ der größte Eigenwert von $A^T A$ ist. Wir müssen also nur $0 < \tau < \frac{2}{\lambda_{\max}}$ wählen, um alle Eigenwerte der Iterationsmatrix im Intervall $(-1, 1)$ zu haben. In diesem Fall ist die Spektralnorm kleiner eins und wir erhalten Konvergenz, wie wir gleich sehen werden.

Im Allgemeinen sind alle diese Iterationsverfahren von der Form

$$x^{k+1} = Gx^k + c,$$

mit einer Matrix G konstruiert aus A und einem Vektor c konstruiert aus A und b. Setzen wir nun ein, dass eine Lösung des Gleichungssystems $\overline{x}$ auch ein Fixpunkt der Iteration ist, so folgt

$$x^{k+1} - \overline{x} = G(x^k - \overline{x}).$$

Mit einer Vektornorm und zugehöriger Matrixnorm folgt

$$\|x^{k+1} - \overline{x}\| = \|G(x^k - \overline{x})\| \leq \|G\| \, \|x^k - \overline{x}\|$$

und daraus induktiv

$$\|x^k - \overline{x}\| \leq \|G\|^k \|x^0 - \overline{x}\|.$$

Wenn $\|G\| < 1$ ist, dann konvergiert die rechte Seite gegen Null, d.h. $x^k \to \overline{x}$. Damit genügt es für die Konvergenz der Iteration also eine Matrixnorm zu finden in der $\|G\| < 1$ gilt.

13.0.2 Konvergenz von iterativen Verfahren

Wir erarbeiten diese Idee genauer: Sei V ein Vektorraum und sei $D \subset V$. Sei $g : D \to V$. Sei $x_0 \in D$ und habe $x_1 = g(x_0)$, $x_2 = g(x_1)$ und so fort. Falls $x_k \to \overline{x}$ und g stetig ist, so folgt $g(\overline{x}) = \overline{x}$. Damit diese Iteration sinnvoll definiert ist, brauchen wir $g(D) \subset D$, daher wählen wir im Allgemeinen direkt $g : D \to D$.

Bereits für manche sehr einfache Funktion ist dies nicht unproblematisch: $g(x) = \lambda \cdot x(1 - x)$

mit $g : [0,1] \to [0,1]$, falls $\lambda \in [0,4]$. Wir sehen mithilfe des Computers, dass $x = 0$ eine triviale Lösung ist, es aber noch einige weitere Lösungen für $x = g(x)$ gibt. Für größere λ divergieren die Folgen sogar. Wir wollen nun untersuchen, wann eine solche rekursive Folge konvergiert. Betrachte

$$|x_{k+1} - \overline{x}| < |x_k - \overline{x}| \ .$$

Das reicht noch nicht, um Konvergenz gegen $\overline{x}$ zu gewährleisten. Wir steuern daher durch einen zusätzlichen Faktor $q < 1$, d.h.

$$|x_{k+1} - \overline{x}| < q \cdot |x_k - \overline{x}| \ .$$

Wir wissen nun, dass nach Definition

$$|g(x_k) - g(\overline{x})| < q \cdot |x_k - \overline{x}| \ .$$

Um dies zu garantieren, fordern wir $|g(x) - g(y)| < q|x - y|$ für alle $x, y \in D$.

Definition 13.1. Seien X, Y normierte Vektorräume, sei $g : X \to Y$.

1) g heißt kontrahierend, falls es ein $q \in [0,1)$ gibt, sodass

$$\|g(x) - g(y)\| \leq q \cdot \|x - y\|$$

für alle $x, y \in X$. q heißt Kontraktionskonstante.

2) Sei $X = Y$. Falls $g(\overline{x}) = \overline{x}$, so heißt $\overline{x}$ Fixpunkt von g.

Bemerkung 13.2. Ist g kontrahierend, so ist g offensichtlich stetig. Wir müssen daher keine Stetigkeit zusätzlich fordern, sofern g eine Kontraktion ist.

Satz 13.3 (Banach'scher Fixpunktsatz)**.** *Sei X ein vollständiger normierter Vektorraum. Sei $\varnothing \neq D \subset X$ eine abgeschlossene Teilmenge. Sei $g : D \to D$ eine Kontraktion mit Konstante q. Dann hat g genau einen Fixpunkt.*

Bemerkung 13.4. Sei $x_0 \in D$ beliebig. Dann konvergiert die Fixpunktiteration $x^{(k+1)} = g(x^{(k)})$ gegen $\overline{x}$.

Beweis. Wir zeigen zunächst, dass es höchstens einen Fixpunkt gibt. Seien dazu $\overline{x}, \overline{y}$ Fixpunkte. Dann gilt

$$\|\overline{x} - \overline{y}\| = \|g(\overline{x}) - g(\overline{y})\| \leq q \cdot \|\overline{x} - \overline{y}\| \ .$$

Ist nun $\overline{x} \neq \overline{y}$, so folgt $q \geq 1$ – dies ist jedoch nach Definition der Kontraktion nicht möglich. Somit folgt $\overline{x} = \overline{y}$, also die Eindeutigkeit. Wir betrachten nun die Fixpunktiteration aus der Bemerkung:

$$\begin{aligned} \|x^{(k+1)} - x^{(k)}\| &\leq \|g(x^{(k)}) - g(x^{(k-1)})\| \\ &\leq q \cdot \|x^{(k)} - x^{(k-1)}\| \\ &\vdots \\ &\leq q^k \cdot \|x^{(1)} - x^{(0)}\| \ . \end{aligned}$$

Dabei haben wir ganz einfach schrittweise iteriert und die Kontraktion genutzt. Sei nun $l > k$. Dann haben wir

$$\|x^{(l)} - x^{k)}\| \leq \|x^{(l)} - x^{(l-1)}\| + \|x^{(l-1)} - x^{(l-2)}\| + \ldots + \|x^{(k+1)} - x^{(k)}\|$$

nach der Dreiecksungleichung. Wir nutzen jetzt wieder die Kontraktion und erhalten:

$$\begin{aligned}\|x^{(l)} - x^{k)}\| &\leq (q^{l-1} + q^{l-2} + \ldots + q^k) \cdot \|x^{(1)} - x^{(0)}\| \\ &\leq q^k \cdot (1 + q + \ldots + q^{l-1-k}) \cdot \|x^{(1)} - x^{(0)}\| \\ &\leq q^k \cdot \frac{1}{1-q} \cdot \|x^{(1)} - x^{(0)}\|\end{aligned}$$

Dabei haben wir im letzten Schritt verwendet, dass $q < 1$ und dass somit die geometrische Reihe eine Abschätzung nach oben liefert. Für $q < 1$ konvergiert der letzte Term gegen 0 und somit ist $x^{(l)}$ eine Cauchy-Folge. Da V vollständig ist, konvergiert $x^{(l)}$ gegen $\overline{x}$. Aus den Grenzwertsätzen folgt dann $g(\overline{x}) = \overline{x}$, d.h. insbesondere gibt es einen Fixpunkt, dessen Eindeutigkeit wir oben schon gezeigt haben. □

Korollar 13.5. *Sei das Setting wie im Banach'schen Fixpunktsatz. Es gilt:*

$$\|\overline{x} - x^{(k)}\| \leq \frac{q^k}{1-q} \cdot \|x^{(1)} - x^{(0)}\| \ .$$

Es gilt zudem:

$$\|\overline{x} - x^{(k+1)}\| \leq \frac{q}{1-q} \cdot \|x^{(k+1)} - x^{(k)}\| \ .$$

Beweis. Die erste Abschätzung erhalten wir durch $l \to \infty$ in der letzten Ungleichung im Beweis des Fixpunktsatzes. Die zweite Abschätzung beweisen wir hier nicht. □

Bemerkung 13.6. Die erste Abschätzung des Korollars hilft uns wie folgt: Ich wähle einen Startwert und berechne $x^{(1)}$. Wir können nun eine Schranke festlegen, den der Fehler höchstens haben soll, und berechnen k so, dass die rechte Seite dieser Obergrenze entspricht. So können wir im Vorhinein abschätzen, wie oft wir iterieren müssen und vermeiden somit, unnötig lange zu rechnen. Man nennt so etwas eine **a priori Abschätzung**.

Die zweite Abschätzung ist **a posteriori**. Die erste Abschätzung wird womöglich etwas unscharf, d.h. sie liefert nur eine sehr grobe Grenze. Aus diesem Grunde ist es durchaus sinnvoll, die Güte dieser Abschätzung zu prüfen.

Wir wollen nun die Frage beantworten, wie wir erkennen, ob es sich bei einer Funktion um eine Kontraktion handelt.

Lemma 13.7. *Sei $g : D \to \mathbb{R}$ stetig differenzierbar, ferner sei $|g'(x)| \leq q < 1$ für alle $x \in \mathbb{R}$. Dann ist g kontrahierend mit Konstante q.*

Beweis. Seien $x, y \in R$, o.B.d.A. $x > y$. Da g differenzierbar ist, gilt

$$\begin{aligned} |g(x) - g(y)| &= |\int_y^x g'(t)dt| \\ &\leq \int_y^x |g'(t)|dt \\ &\leq \sup_{t\in[0,1]} \|g'(t)\| \int_y^x 1 \; dt \\ &\leq q|x - y| \end{aligned}$$

Somit ist g kontrahierend mit Konstante q. □

Beispiel 13.8. Wir geben ein paar Funktionen an.

1) Die Funktion $g : \mathbb{R} \to \mathbb{R}$ mit $g(x) = \cos(x)$ ist nicht kontrahierend.

2) Die Funktion $g(x) = 0.9 \cdot \cos(x)$ ist kontrahierend, denn $|g'(x)| = |0.9 \cdot (-\sin(x))| \leq 0.9$. Somit wäre $q = 0.9$.

3) Suche $\overline{x}$ mit $\tan(\overline{x}) = \overline{x}$ und wir suchen $\overline{x} \in \left(\frac{\pi}{2}, \frac{3\pi}{2}\right)$. Setze $g(x) = \tan(x)$. Es ist nun

$$g'(x) = \frac{1}{\cos^2(x)} \geq 1 \; ,$$

d.h. die aktuelle Wahl von g ist absolut nicht clever, denn so ist g nicht kontrahierend. Wir wenden daher arctan an und erhalten:

$$\overline{x} - \pi = \arctan(\overline{x}) \; .$$

Dies ist äquivalent zu $\overline{x} = \arctan(\overline{x}) + \pi$. Setze $g(x) := \arctan(x) + \pi$ und wir erhalten

$$g'(x) = \frac{1}{1 + x^2} < 1 \; .$$

4) Setze $g(x) = Bx + c$, dann ist $g'(x) = B$, wobei B eine Matrix ist. Dann ist g kontrahierend genau dann, wenn $\|B\| < 1$. Sei genauer $B \in \mathbb{R}^{n\times}$ mit $\|B\| < 1$ und sei $c \in \mathbb{R}^n$. Sei $x^{(0)} \in \mathbb{R}^n$. Setze

$$x^{(k+1)} := Bx^{(k)} + c$$

und $x^{(k)}$ konvergiert gegen $\overline{x}$ und somit $\overline{x} = g(\overline{x}) = B\overline{x} + c$. Somit folgt $(I - B)\overline{x} = c$ und das bedeutet, dass $x^{(k)}$ gegen $(I - B)^{-1}c$ konvergiert. Man erinnere sich hier an die Neumann-Reihe!

Das Newton-Verfahren

Sei $f : \mathbb{R} \to \mathbb{R}$. Zu bestimmen sei eine Nullstelle $\overline{x}$ von f. Man könnte also

$$f(\overline{x}) + \overline{x} = \overline{x}$$

lösen wollen. Das ist aber nicht so clever, denn es wird sehr schwierig, hier die Kontraktionseigenschaft nachzuweisen. Wir werden somit eher ein iteratives Verfahren konstruieren, das sogenannte Newton-Verfahren. Statt einer Nullstelle der Funktion suchen wir eine Nullstelle der Tangente. Sei dazu $f : \mathbb{R} \to \mathbb{R}$ differenzierbar. Zu bestimmen sei eine Nullstelle $\overline{x}$ von f. Die Tangente an $x^{(0)}$ beschreiben wir durch

$$y = f(x^{(0)}) + f'(x^{(0)}) \cdot (x - x^{(0)})$$

und dies soll Null sein. Dann haben wir

$$x = x^{(0)} - f'(x^{(0)})^{-1} \cdot f(x^{(0)}) \ .$$

Definiere

$$g(x) := x - f'(x)^{-1} \cdot f(x)$$

und weiter

$$x^{(k+1)} := x^{(k)} - f'(x^{(k)})^{-1} \cdot f(x^{(k)}) \ .$$

Diese Folge ist die **Newtonfolge**. Wir nehmen an, dass diese Folge gegen ein $\widetilde{x}$ konvergiert, sodass $\widetilde{x} = g(\widetilde{x}) = \widetilde{x} - f'(\widetilde{x})^{-1} \cdot f(\widetilde{x})$. Gehen wir nun davon aus, dass f' als Jacobimatrix invertierbar ist, so ist $f'^{-1} \neq 0$. Subtrahieren wir $\widetilde{x}$, so ist $f'(\widetilde{x})^{-1} \cdot f(\widetilde{x}) = 0$ und es folgt $f(\widetilde{x}) = 0$, d.h. wir haben unsere gesuchte Nullstelle.

Lemma 13.9. *Sei $U \subset \mathbb{R}^m$ offen. Sei $g : U \to \mathbb{R}^n$ kontrahierend. Sei $\overline{x} = g(\overline{x})$ ein Fixpunkt von g. Dann existiert eine abgeschlossene Umgebung D von $\overline{x}$ mit der Eigenschaft $g : D \to D$.*

Beweis. Sei ϵ so klein, dass $D := \overline{B_\epsilon(\overline{x})} \subset U$. Sei nun $x \in D$, dann ist $\|g(x) - \overline{x}\| = \|g(x) - g(\overline{x})\| \leq q \cdot \|x - \overline{x}\| \leq q \cdot \epsilon < \epsilon$ wegen $q < 1$. Somit $g(x) \in D$. □

Dieses Lemma hilft uns tatsächlich, denn: Wissen wir, dass es einen Fixpunkt gibt, so wissen wir auch, dass wir g zu einer Selbstabbildung machen können.

Korollar 13.10. *Sei $g : U \to \mathbb{R}^n$ stetig differenzierbar, sei $U \subset \mathbb{R}^m$. Es sei $\overline{x} \in \mathring{U}$ mit $g(\overline{x}) = \overline{x}$. Zusätzlich gelte, dass $\|g'(\overline{x})\| < 1$. Dann gibt es eine Umgebung V von $\overline{x}$, sodass $g|_V$ kontrahierend ist.*

Dieses Korollar besagt somit, dass die Fixpunktfolge auf einer kleinen Umgebung von $\overline{x}$ konvergiert.

Beweis. g' ist stetig, weiter $\|g'(\overline{x})\| < 1$. Wähle q mit

$$\|g'(\overline{x})\| < q < 1$$

und da g' stetig ist, gibt es eine Umgebung V von $\overline{x}$, sodass $\|g'(x)\| < q$ für alle $x \in V$. Somit ist g kontrahierend auf V und nun greift der Banach'sche Fixpunktsatz. □

Beispiel 13.11 (Newtonverfahren)**.** Sei $f(x) = x^2 - a$ mit $a > 0$. Gesucht ist die Nullstelle von f, d.h. $\overline{x}$ mit $f(\overline{x}) = 0$. Wir definieren nun

$$g(x) := x - \frac{f(x)}{f'(x)} = x - \frac{x^2 - a}{2x} = \frac{1}{2} \cdot \left(x + \frac{a}{x}\right) \ .$$

Gilt nun $g(\overline{x}) = \overline{x}$, so haben wir

$$\frac{1}{2}\left(\overline{x} + \frac{a}{\overline{x}}\right) = \overline{x} \;\Leftrightarrow\; \overline{x} = \frac{a}{\overline{x}} \; .$$

Wir wollen nun zeigen, dass

$$x^{(k+1)} = g(x^{(k)})$$

konvergiert. Wir betrachten

$$g'(x) = \frac{1}{2}\left(1 - \frac{a}{x^2}\right)$$

und sehen, dass $g'(\sqrt{a}) = 0$ und dass g bei $x = \sqrt{a}$ ein Minimum hat. Die Folge ist somit lokal konvergent, aber mehr wissen wir gerade nicht. An dieser Stelle kamen wir in der Vorlesung leider wegen eines Denkfehlers nicht weiter.

Beispiel 13.12. Betrachte

$$f(x) = \frac{1}{x} - a$$

für $a > 0$. Wir wollen das Newtonverfahren anwenden und wandeln unser Problem zunächst in ein Fixpunktproblem um:

$$\begin{aligned} g(x) = x - \frac{f(x)}{f'(x)} &= x + \left(\frac{1}{x} - a\right) \cdot x^2 \\ &= x + x - ax^2 \\ &= x(2 - ax) \; . \end{aligned}$$

Nun gilt für einen Fixpunkt $\overline{x} = 2\overline{x} - a\overline{x}^2$, was äquivalent ist zu $\overline{x} = 0$ bzw. $\overline{x} = 1/a$. Die zugehörige Newtonfolge ist damit:

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} = x^{(k)} \cdot (2 - ax^{(k)}) \; .$$

Man rechnet leicht nach, dass die Folge monoton und beschränkt ist und dass sie somit gegen $1/a$ konvergiert.

Definition 13.13 (Asymptotische Konvergenzgeschwindigkeit)**.** Sei $x^{(k)} \to \overline{x}$ eine konvergente Folge.

1) Falls es ein $0 < q < 1$ gibt mit

$$\|x^{(k+1)} - \overline{x}\| \leq q \cdot \|x^{(k)} - \overline{x}\| \; ,$$

so heißt die Konvergenz linear oder von der Ordnung 1.

2) Sei $p > 1$. Falls es ein $C > 0$ gibt mit

$$\|x^{(k+1)} - \overline{x}\| \leq C \cdot \|x^{(k)} - \overline{x}\|^p \; ,$$

so heißt $x^{(k)}$ konvergent von der Ordnung p.

Satz 13.14 (Konvergenz des Newtonverfahrens)**.** *Sei $f : \mathbb{R}^n \to \mathbb{R}^n$ zweimal stetig differenzierbar. Sei $\overline{x}$ eine Nullstelle von f.*

1) *Sei $f'(\overline{x})$ invertierbar. Dann gibt es eine Umgebung U von $\overline{x}$, sodass das Newtonverfahren gegen $\overline{x}$ konvergiert, falls $x^{(0)} \in U$. Die Konvergenzgeschwindigkeit ist quadratisch.*

2) *Sei $f'(\overline{x})$ nicht invertierbar, aber $f''(\overline{x})$ invertierbar und $f'(y)$ invertierbar in einer kleinen Umgebung von $\overline{x}$ für $y \neq \overline{x}$. Dann konvergiert das Newtonverfahren gegen $\overline{x}$ mit linearer Geschwindigkeit.*

Beweis. 1) Wir zeigen zunächst die lokale Konvergenz im Fall $n = 1$. Für

$$g(x) = x - \frac{f(x)}{f'(x)}$$

ist zu zeigen, dass $|g'(\overline{x})| < 1$. Dann folgt

$$g'(x) = 1 - \frac{(f')^2 - ff''}{(f')^2} = \frac{ff''}{(f')^2}$$

und wir wissen, dass $f(\overline{x}) = 0$. Somit folgt (da $f'(\overline{x}) \neq 0$)

$$|g'(\overline{x})| = \left| \frac{f''(\overline{x}) f(\overline{x})}{f'(\overline{x})^2} \right| < 1 \ .$$

Aus dem Fixpunktsatz erhalten wir nun die Konvergenz. Wir zeigen nun noch die quadratische Konvergenz. Dazu nutzen wir Taylor. Da $\overline{x}$ eine Nullstelle ist, haben wir:

$$0 = f(\overline{x}) = f(x) + f'(x)(\overline{x} - x) + f''(\xi)(\overline{x} - x)^2$$

für ein passendes ξ. Wir erhalten:

$$f(x^{(k)}) = -f'(x^{(k)})(\overline{x} - x^{(k)}) - f''(\xi)(\overline{x} - x^{(k)})^2 \ .$$

Wir betrachten nun

$$\begin{aligned} \left| x^{(k+1)} - \overline{x} \right| &= \left| x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} - \overline{x} \right| \\ &= \left| x^{(k)} - \overline{x} + \frac{f'(x^{(k)})(\overline{x} - x^{(k)}) + f''(\xi)(\overline{x} - x^{(k)})^2}{f'(x^{(k)})} \right| \\ &= \left| f''(\xi)(\overline{x} - x^{(k)})^2 \cdot \frac{1}{f'(x^{(k)})} \right| \end{aligned}$$

Da f'' stetig ist, ist $|f''(x)| \leq C_1$ in einer kleinen Umgebung von $\overline{x}$. In dieser kleinen Umgebung ist $f'(\overline{x})$ und $f'(x)$ invertierbar. Auf dieser kleinen Umgebung ist $(f'(x))^{-1}$ ebenfalls stetig. Somit ist auch

$$\left| \frac{1}{f'(x)} \right| \leq C_2$$

beschränkt auf dieser Umgebung. Wir folgern

$$\left| x^{(k+1)} - \overline{x} \right| \leq C_1 \cdot C_2 \cdot (\overline{x} - x^{(k)})^2$$

und das ist genau die Definition von quadratischer Konvergenz.

2) Sparen wir uns.

□

Anhang A

Grundlegendes aus der linearen Algebra

Normierte Vektorräume

Wir sparen uns die Vektorraumdefinition. Sei V ein Vektorraum. Eine Abbildung $\|\cdot\| : V \to \mathbb{R}^+$ heißt Norm, falls

- $\|\alpha x\| = |\alpha| \cdot \|x\|$ für alle $\alpha \in \mathbb{R} \vee \mathbb{C}$, $x \in V$ (Linearität).
- $\|x\| = 0$ genau dann, wenn $x = 0$ (Definitheit).
- $\|x + y\| \leq \|x\| + \|y\|$ (Dreiecksungleichung).

Ist dies erfüllt, so heißt $(V, \|\cdot\|)$ normierter Vektorraum. Wir kennen bereits einige Normen, sei hier $V = \mathbb{C}^n$.

- Euklidische Norm: $\|v\|_2^2 = \sum_{k=1}^n |v_k|^2$.
- p-Norm: $\|v\|_p = \sqrt[p]{\sum_{k=1}^n |v_k|^p}$.
- ∞-Norm: $\|v\|_\infty = \max\limits_{1 \leq k \leq n} |v_k|$.

Sei nun $V = C([0,1])$, d.h. der Vektorraum der stetigen Funktionen $[0,1] \to \mathbb{R}$. Dann haben wir

$$\|v\|_p = \left(\int_0^1 |v(x)|^p dx \right)^{1/p}$$

für $p \geq 0$.

Banachräume

Sei $(V, \|\cdot\|)$ normierter Raum. Sei $v_k \in V$, $k \in \mathbb{N}$, $v \in V$. Wir schreiben $v_k \to v$ (d.h. eine Folge v_k konvergiert gegen v), falls $\|v_k - v\| \to 0$. Falls jede Cauchyfolge in V konvergiert, so heißt V vollständig oder Banachraum.

Sei nun $V = \mathbb{Q}$. Dann existiert eine Folge $v_k \in \mathbb{Q}$, deren Grenzwert nicht in $\mathbb{Q}$ liegt (man denke an den Grenzwert $\sqrt{2}$). Für $V = C([0,1])$ gibt es eine Folge v_k, die gegen eine nicht stetige Funktion konvergiert. Wir haben somit zwei nicht vollständige Räume kennengelernt. Vollständig hingegen sind beispielsweise die reellen Zahlen. Weiterhin wäre $V = C([0,1])$ vollständig, wenn wir den Raum mit der Supremumsnorm versehen, denn jeder Grenzwert einer stetigen Funktion bezüglich der Supremumsnorm ist wieder eine stetige Funktion.

Skalarprodukt

Sei V ein Vektorraum (endlichdimensional). Eine Funktion

$$(\cdot,\cdot) : V \times V \to \mathbb{R} \vee \mathbb{C}$$

heißt Skalarprodukt, falls

- $(\lambda u + \mu v, w) = \lambda(u,w) + \mu(v,w)$ für alle $\lambda, \mu \in \mathbb{R} \vee \mathbb{C}$ und $u, v, w \in V$.
- $(u,v) = \overline{(v,u)}$ für alle $u, v \in V$. Mit $\overline{x}$ ist hier das komplex Konjugierte gemeint.
- $(v,v) \geq 0$ für alle $v \in V$. Aus $(v,v) = 0$ folgt $v = 0$.

Sei $V = \mathbb{C}^n$. Dann nutzen wir üblicherweise das euklidische Skalarprodukt

$$(v,w) = \sum_{k=1}^{n} v_k \overline{w_k} \ .$$

Wir gewinnen aus einem Skalarprodukt eine Norm mit dem Satz von Cauchy-Schwartz:

$$|(u,v)| \leq \|u\| \cdot \|v\|$$

mit $\|u\| = \sqrt{(u,u)}$ für alle $u, v \in V$. Ein weiterer Satz besagt, dass die hier induzierte Norm $\|u\|$ tatsächlich eine Norm ist. Wir zeigen hier kurz die Dreiecksungleichung für den Fall, dass V ein Vektorraum über $\mathbb{R}$ ist:

$$\begin{aligned}
\|u+v\|^2 &= (u+v, u+v) \\
&= (u,u) + (v,u) + (u,v) + (v,v) \\
&= \|u\|^2 + \|v^2\| + 2 \cdot (u,v) \\
&\leq \|u\|^2 + \|v\|^2 + 2 \cdot \|v\| \cdot \|u\| \\
&= (\|u\| + \|v\|)^2
\end{aligned}$$

Wurzelziehen über positiven Zahlen liefert nun direkt die Dreiecksungleichung. Wenn wir künftig einen Vektorraum mit Skalarprodukt betrachten, dann wählen wir immer eine dazu passende Norm. Für $\mathbb{R}$ oder $\mathbb{C}$ nehmen wir immer die euklidische Norm.

Wir beachten, dass für endlichdimensionale Vektorräume alle Normen äquivalent sind.

Lineare Operatoren

Für uns bedeuten lineare Operatoren im Normalfall Matrizen. Seien U und V Vektorräume, sei $T : U \to V$. T heißt linearer Operator von U nach V genau dann, wenn $T(\alpha U + \beta V) = \alpha T(U) + \beta T(v)$ für alle $\alpha, \beta \in \mathbb{R} \vee \mathbb{C}$ und $u, v \in U$.

Ist $U = \mathbb{R}^n$ und $V = \mathbb{R}^m$, so wird T dargestellt durch eine $m \times n$-Matrix, d.h. m Zeilen und n Spalten. Ist nun $x \in \mathbb{R}^n$, so ist $T(x) = T \cdot x$ die Matrixmultiplikation. In der Numerik interessiert nun oft der Aufwand bei diversen Berechnungen. Wie aufwendig ist die Berechnung von $T \cdot x$? Wir wissen, dass

$$(Tx)_j = \sum_{k=1}^{n} T_{jk} x_k$$

für $j = 1, \ldots, m$. Der Aufwand für eine Komponente besteht aus n Multiplikationen und $n-1$ Additionen. Der Gesamtaufwand besteht somit in $m \cdot n$ Multiplikationen und $m(n-1) = mn - m$ Additionen. Für große m und n ist nur noch mn entscheidend für den Aufwand. Wir konzentrieren uns folglich auf die Leitterme. Dafür kennen wir die Landau-Symbole, mit denen wir sagen: Die Anzahl der Additionen ist $mn + O(m)$. Dabei bedeutet die Notation konkret (für $n \to \infty$):

$$f(n) = O(n^p)$$

genau dann, wenn eine Konstante c und ein Wert n_0 existieren, sodass

$$|f(n)| \leq c \cdot n^p$$

für alle $n \geq n_0$. In unserem Fall heißt das

$$-m = O(m)$$

für $c = 1$, da $|m| \leq m$ immer erfüllt ist. Ein anderes Beispiel: $n = O(n^2)$ ist wegen $n \leq c \cdot n^2$ richtig mit $n_0 = 0$ und $c = 1$. Geht auch $n^2 = O(n)$? Das ist natürlich Blödsinn, eine entsprechende Konstante c kann man nicht finden. Wir gehen nun zurück zur Ausgangsfrage nach dem Aufwand von Tx. Wir sehen, dass die Anzahl der Multiplikationen und der Additionen sich nur gering unterscheiden (bis auf einen vernachlässigbaren Term für große m und n). Dies ist überraschend oft der Fall, daher bezeichnen wir mit einer Rechenoperation fortan eine Multiplikation mit einer Addition. Damit bräuchten wir hier mn Rechenoperationen. Wir sehen, dass Normen auf Matrizen eine gewisse Relevanz haben.

Induzierte Norm

Sei $A \in \mathbb{R}^{m \times n}$ eine Matrix. Dann ist die Form in etwa

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \vdots & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix} .$$

Nun könnte man

$$\|A\|^2 = \sum_{i,j} |a_{ij}|^2$$

vermuten (Frobenius-Norm), die jedoch nicht allen unseren Ansprüchen genügt. Wir fordern nämlich

$$\|Av\| \leq \|A\| \cdot \|v\|$$

für alle $v \in \mathbb{R}^n$. Dabei sollte $\|A\|$ die kleinstmögliche Zahl erzeugen, um möglichst gute Abschätzung in numerischen Anwendungen zu erhalten. Für $v \neq 0$ erhalten wir

$$\frac{\|Av\|}{\|v\|} \leq \|A\| \ .$$

Entsprechend definieren wir mit

$$\|A\| = \sup_{v \in \mathbb{R}^n \setminus \{0\}} \frac{\|Av\|}{\|v\|}$$

die Matrixnorm. Dies halten wir allgemein fest.

Definition A.1. Sei $A : U \to V$, wobei U und V normierte Vektorräume seien. Wir definieren

$$\|A\| = \sup_{v \in U \setminus \{0\}} \frac{\|Av\|_V}{\|v\|_U}$$

für $v \in U$. Damit ist $\|A\|$ tatsächlich eine Norm im Vektorraum der linearen Abbildung von U nach V und es gilt $\|Av\| \leq \|A\| \cdot \|v\|$. Wir nennen $\|A\|$ die induzierte Norm.

Man beachte, dass für nicht endlichdimensionale Vektorräume nicht unbedingt das Supremum existiert. Die jetzt verwendete Definition ist noch nicht so praktisch, da man nicht wirklich gut damit rechnen kann. Wir können auch schreiben

$$\begin{aligned}
\|A\| &= \sup \frac{\|Av\|}{\|v\|} \\
&= \sup \|\frac{1}{\|v\|} \cdot Av\| \\
&= \sup \|A \cdot \frac{v}{\|v\|}\| \\
&= \sup_{\|v\|=1} \|Av\|
\end{aligned}$$

Das bedeutet, dass wir im Endeffekt nur alle v mit der Norm 1 das Supremum bilden müssen. Praktisch betrachten wir jedoch $U = \mathbb{R}^n$, $V = \mathbb{R}^m$ und $A \in \mathbb{R}^{m \times n}$. Wir schreiben häufig $A = (a_1 \ \dots \ a_n)$, wobei a_i die einzelnen Spalten sind, oder

$$A = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix}$$

in Zeilenschreibweise. Daraus ergibt sich

$$Ax = (a_1 \ \dots \ a_n) \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = (x_1 a_1 \ \dots \ x_n a_n)$$

für $x \in \mathbb{R}^n$. Für $y \in \mathbb{R}^m$ erhalten wir analog

$$y^T A = (y_1\alpha_1 \ \dots \ y_m\alpha_m) \ .$$

Für $B \in \mathbb{R}^{n\times n}$ mit $B = (b_1 \ \dots \ b_n)$ erhalten wir

$$A \cdot B = (Ab_1 \ \dots \ Ab_n) \ .$$

Ein konkretes Beispiel: Seien $A, D \in \mathbb{R}^{n\times n}$ und

$$D = \begin{pmatrix} 5 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & \ddots & \dots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & \dots & \dots & \dots & 1 \end{pmatrix}$$

Dann haben wir $AD = (5a_1 \ a_2 \ \dots \ a_n)$. Wir kehren nun zurück der Betrachtung der Matrixnorm. Sei dazu fortan $A \in \mathbb{R}^{m\times n}$ mit der Norm $\|\cdot\|$, sodass

$$\begin{aligned} \|A\| &:= \sup_{x\in\mathbb{R}^n, \ x\neq 0} \frac{\|Ax\|}{\|x\|} \\ &= \sup_{x\in\mathbb{R}^n, \ \|x\|=1} \|Ax\| \end{aligned}$$

gilt. Achtung: Diese Definition hängt von der Norm in $\mathbb{R}^n$ ab! Sei nun $x \in \mathbb{R}^n$ mit $x \neq 0$. Dann haben wir

$$\|Ax\| = \|x\| \cdot \|A \cdot \frac{x}{\|x\|}\| \leq \|x\| \cdot \|A\|$$

wegen $\|x/\|x\|\| = 1$ und wegen der Definition von $\|A\|$ als Supremum. Sei nun $B \in \mathbb{R}^{n\times p}$. Dann ist

$$\begin{aligned} \|A \cdot B\| &= \sup_{x\in\mathbb{R}^p, \ \|x\|=1} \|ABx\| \\ &\leq \sup_{\|x\|=1} \|A\| \cdot \|Bx\| \\ &\leq \sup_{\|x\|=1} \|A\| \cdot \|B\| \cdot \|x\| = \|A\| \cdot \|B\| \end{aligned}$$

und das bedeutet, dass die Norm ähnlich wie der Betrag sehr angenehme Eigenschaften besitzt.

Beispiel A.2. Wir betrachten $\|x\|_\infty = \max |x_k|$ für $x \in \mathbb{R}^n$, also die sogenannte ∞-Norm. Wir berechnen nun $\|A\|_\infty$. Wir behaupten zunächst, dass

$$\|A\|_\infty = \max_{i=1}^{m} \sum_{k=1}^{n} |A_{ik}|$$

Wir müssen nun beide Anforderungen an eine Matrixnorm nachweisen. eine Matrixnorm definiert. Praktisch gesehen: Für

$$C = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

ist das Maximum über alle Zeilensummen 24, somit $\|C\| = 24$. Sei nun allgemein $x \in \mathbb{R}^n$ mit $\|x\|_\infty = 1$. Daraus folgt direkt, dass $|x_k| \leq 1$. Betrachten wir nun die i-te Komponente

$$\begin{aligned} |(Ax)_i| &= \left| \sum_{k=1}^{n} A_{ik} x_k \right| \\ &\leq \sum_{k=1}^{n} |A_{ik}| \cdot |x_k| \\ &\leq \sum_{k=1}^{n} |A_{ik}| \leq \max_{i=1}^{n} \sum_{k=1}^{n} |A_{ik}| \end{aligned}$$

Wir folgern, dass

$$\|Ax\|_\infty \leq \max_{i=1}^{n} \sum_{k=1}^{n} |A_{ik}| \ .$$

Zu zeigen ist noch, dass dies tatsächlich die beste obere Schranke ist – man sagt auch, dass „die Schranke scharf ist“. Dies ist durch die Angabe eines Elementes erreicht, für dass wir hier die beste obere Schranke haben – dafür haben wir oben ein Beispiel. Den Rest zeigen wir nicht mehr.

Eigenwerte und Eigenvektoren

Sei $A \in \mathbb{R}^{n \times n}$. Ein $\lambda \in \mathbb{C}$ heißt Eigenwert von A zum Eigenvektor $v \in \mathbb{R}^n$, falls

$$Av = \lambda v$$

und $v \neq 0$. Das bedeutet, $\lambda \in \mathbb{C}$ ist Eigenwert von A genau dann, wenn es einen Vektor $v \neq 0$ gibt, zu dem λ Eigenwert gibt.

(Selbst)Adjungiertheit und hermitesche Matrizen

Definition A.3. Seien U und V Vektorräume mit Skalarprodukt. Sei $A : U \to V$ eine lineare Abbildung. Sei $A^* : V \to U$. Dann heißt A^* adjungierte Abbildung zu A, falls

$$(Au, v) = (u, A^* v)$$

für alle $u \in U$, $v \in V$.

Beispiel A.4. Sei $U = \mathbb{C}^n$ und $V = \mathbb{C}^m$, sei $A \in \mathbb{C}^{m\times n}$. Sei $u \in U$, $v \in V$. Dann berechnen wir

$$\begin{aligned}(Au, v) &= (Au)^T \cdot \overline{v} \\ &= u^T \cdot A^T \cdot \overline{v} = u^T \cdot \overline{\overline{A^T} \cdot v} \\ &= (u, \overline{A^T} \cdot v)\end{aligned}$$

unter Berücksichtigung des komplexen Skalarproduktes $(x, y) = x^T \cdot \overline{y}$. Wir haben somit, dass $\overline{A^T} = A^*$. Für reelle Matrizen erhalten wir sogar $A^* = A^T$.

Wir betrachten nun weiter

$$(AB)^* = B^* \cdot A^*$$

und überlegen dazu, dass

$$\begin{aligned}(u, (AB)^* v) &= (ABu, v) \\ &= (Bu, A^* v) \\ &= (u, B^* A^* v)\end{aligned}$$

für alle u und v. Daraus folgt $(AB)^* = B^* A^*$. Weiter betrachten wir $(A^*)^*$ mittels

$$\begin{aligned}(u, (A^*)^* v) &= (A^* u, v) \\ &= \overline{(v, A^* u)} \\ &= \overline{(Av, u)} = (u, Av)\ .\end{aligned}$$

Definition A.5. Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt selbstadjungiert genau dann, wenn $A = A^*$.

Für reelle Matrizen bedeutet dies wegen $A = A^* = A^T$, dass symmetrische Matrizen über $\mathbb{R}$ selbstadjungiert sind. Für komplexe Matrizen müsste gelten, dass $A = A^* = \overline{A^T}$. Diese Matrizen sind also symmetrisch, nachdem komplex konjugiert wurde. Solche Matrizen nennen wir hermitesch.

Lemma A.6. *Sei $A \in \mathbb{C}^{m\times n}$. Dann ist $A^*A \in \mathbb{C}^{n\times n}$ hermitesch. Ebenso ist $AA^* \in \mathbb{C}^{m\times m}$ hermitesch.*

Beweis. Das ist einfaches Nachrechnen: $(A^*A)^* = A^*A^{**} = A^*A$, den zweiten Fall sparen wir uns. □

Satz A.7. *Sei $A \in \mathbb{C}^{n\times n}$ hermitesch. Dann gibt es eine Orthonormalbasis $(v_1, \ldots, v_n)$ aus Eigenvektoren von A zu den reellen Eigenwerten $\lambda_1, \ldots, \lambda_n$.*

Sei v_i Eigenvektor zum Eigenwert $\lambda_i \neq \lambda_k$ mit v_k Eigenvektor zum Eigenwert λ_k. Dann

$$\begin{aligned}\lambda_i(v_i, v_k) &= (\lambda_i v_i, v_k) \\ &= (Av_i, v_k) \\ &= (v_i, Av_k) \\ &= (v_i, \lambda_k v_k) = \lambda_k(v_i, v_k)\end{aligned}$$

Das bedeutet, dass $(\lambda_i - \lambda_k)(v_i, v_k) = 0$ und somit sind v_i und v_k orthogonal.

Satz A.8. *Sei $A \in \mathbb{R}^{m\times n}$ und $\mathbb{R}^m$, $\mathbb{R}^n$ versehen mit der euklidischen Norm. Dann ist*

$$\|A\|_2 = \sqrt{\lambda_1}$$

mit λ_1 als maximalem Eigenwert von A^TA.

Beweis. Sei λ Eigenwert von A^TA zum Eigenvektor v mit $\|v\| = 1$. Dann ist (unter Anwendung der oben bereits genutzten Rechenregeln für Skalarprodukt und Adjunktion)

$$\begin{aligned}\|Av\|_2^2 &= (Av, Av)\\ &= (v, A^TAv)\\ &= (v, \lambda v) = \lambda(v, v)\end{aligned}$$

und es folgt

$$\lambda = \frac{\|Av\|_2^2}{\|v\|_2^2} \geq 0 \ .$$

Wir wissen nun, dass $A^TA \in \mathbb{R}^{n\times n}$ hermitesch und im reellen gerade symmetrisch ist. Sei damit $(v_1, \ldots, v_n)$ eine Orthonormalbasis aus Eigenvektoren von A^TA zu den Eigenwerten $\lambda_1, \ldots, \lambda_n$ (die Existenz verrät der obige Satz). Sei nun $v \in \mathbb{R}^n$ beliebig. Es gibt nun $\mu_1, \ldots, \mu_n$ mit

$$v = \sum_{k=1}^{n} \mu_k v_k \ .$$

Damit ist

$$\begin{aligned}\|v\|_2^2 &= \|\sum_{k=1}^{n} \mu_k v_k\|_2^2\\ &= \left(\sum_{k=1}^{n} \mu_k v_k \ , \ \sum_{j=1}^{n} \mu_j v_j\right)\\ &= \sum_{k=1}^{n}\sum_{j=1}^{n} (\mu_j v_k, \mu_k v_k)\\ &= \sum_{k=1}^{n} \mu_k^2 (v_k, v_k) = \sum_{k=1}^{n} \mu_k^2 \ ,\end{aligned}$$

wobei die dritte Gleichheit dadurch erklärt ist, dass das Skalarprodukt für orthogonale Vek-

toren Null wird. Wir betrachten nun

$$\begin{aligned}
\|Av\|_2^2 &= \left(A\sum_{k=1}^{n} \mu_k v_k \ , \ A\sum_{j=1}^{n} \mu_j v_j \right) \\
&= \left(\sum_{k=1}^{n} \mu_k v_k \ , \ A^T A \left(\sum_{j=1}^{n} \mu_j v_j \right) \right) \\
&= \left(\sum_{k=1}^{n} \mu_k v_k \ , \ \sum_{j=1}^{n} \mu_j A^T A v_j \right) \\
&= \left(\sum_{k=1}^{n} \mu_k v_k \ , \ \sum_{j=1}^{n} \mu_j \lambda_j v_j \right) \\
&= \sum_{k=1}^{n} \mu_k \mu_k \lambda_k = \sum_{k=1}^{n} \lambda_k \mu_k^2
\end{aligned}$$

Die λ_k seien der Größe nach geordnet, wobei λ_1 der größte Wert sei (wir haben oben bereits gezeigt, dass λ nichtnegativ ist). Daher können wir weiter abschätzen

$$\sum_{k=1}^{n} \lambda_k \mu_k^2 \leq \sum_{k=1}^{n} \lambda_1 \mu_k^2 = \lambda_1 \cdot \|v\|_2^2 \ .$$

Das bedeutet: Für alle $v \in \mathbb{R}^n \setminus \{0\}$ gilt

$$\frac{\|Av\|_2^2}{\|v\|_2^2} \leq \lambda_1 \ .$$

Somit

$$\|A\|_2 = \sup_{v \in \mathbb{R}^n \setminus \{0\}} \frac{\|Av\|}{\|v\|} \leq \sqrt{\lambda_1}$$

Wir betrachten nun

$$\begin{aligned}
\|Av_1\|_2^2 &= (Av_1, Av_1) = (v_1, A^T A v_1) \\
&= \lambda_1 (v_1, v_1) = \lambda_1 \|v_1\|^2
\end{aligned}$$

und wissen nun

$$\frac{\|Av_1\|^2}{\|v_1\|^2} = \lambda_1$$

und somit auch, dass

$$\|A\|_2 = \sup_{v \in \mathbb{R}^n \setminus \{0\}} \frac{\|Av\|}{\|v\|} \geq \frac{\|Av_1\|}{\|v_1\|} = \sqrt{\lambda_1}$$

und somit ist der Beweis erbracht. □

Sei fortan $A \in \mathbb{R}^{m \times n}$. Wir arbeiten auf $(\mathbb{R}^n, \|\cdot\|)$ und $(\mathbb{R}^m, \|\cdot\|)$. Wir setzen die bekannte Matrixnorm mit den beiden äquivalenten Definition voraus. Insbesondere sei nochmal an die fundamentale Eigenschaft $\|Ax\| \leq \|A\| \cdot \|x\|$ erinnert. Dies ist keine triviale Identität, sondern nur nach entsprechender Definition richtig.

Anhang B

Grundlegendes aus der Analysis

Im Folgenden wiederholen wir kurz eine wichtige Grundlagen aus der Analysis, die wir häufiger in dieser Vorlesung benutzen werden.

B.1 Taylor Entwicklung

Ist $f : \mathbb{R} \to R$ eine k-mal stetig differenzierbare Funktion, so können wir die Werte von f in einem Intervall durch den Wert an einem einzelnen Punkt und dessen Ableitungen annähern. Es gilt

$$f(x) = f(x_0) + \sum_{j=1}^{k-1} \frac{f^{(j)}(x_0)}{j!}(x - x_0)^j + \frac{f^{(k)}(\tilde{x})}{k!}(x - x_0)^k,$$

für ein $\tilde{x} \in [x_0, x]$. Ist $x - x_0$ klein, so können wir f durch ein Polynom vom Grad $k - 1$ annähern, der Fehler dabei ist von der Ordnung $|x - x_0|^k$.

B.2 Fundamentalsatz der Integralrechnung

Wir können die Stammfunktion einer stetigen Funktion f aus dem unbestimmten Integral bestimmen und umgekehrt:

$$F(t) = F(t_0) + \int_{t_0}^{t} f(s)\ ds \Leftrightarrow F'(t) = f(t).$$

B.3 Banach'scher Fixpunktsatz

Wir betrachten einen normierten linearen Raum X. Ist X vollständig, d.h. der Grenzwert jeder konvergenten Folge in X liegt wieder in X, so heisst X Banachraum. Dann gilt der folgende Fixpunktsatz:

Satz B.1. *Sei $\Phi : X \to X$ eine kontraktive Abbildung, d.h. es gibt ein $\eta < 1$ sodass für alle $x_1, x_2 \in X$ gilt*

$$\|Phi(x_1) - \Phi(x_2)\| \leq \eta \|x_1 - x_2\|.$$

Dann existiert ein eindeutiger Fixpunkt von Φ, d.h. ein $\overline{x} \in X$ mit

$$\Phi(\overline{x}) = \overline{x}.$$

Anhang C

Gewöhnliche Differentialgleichungen

Im Folgenden wiederholen wir einige grundlegende Eigenschaften linearer Differentialgleichungen. Allgemein nennen wir eine Gleichung gewöhnliche Differentialgleichung, wenn wir als Unbekannte eine Funktion u einer Variablen suchen, von der auch Ableitungen in der Gleichung vorkommen. Wir nennen die Gleichung von der Ordnung k, wenn sowohl die Funktion als auch die k-te Ableitung in der Gleichung auftreten, jedoch keine höheren Ableitungen. Das kanonische Beispiel ist die Differentialgleichung erster Ordnung

$$u'(t) = F(t, u(t)) \tag{C.1}$$

mit einer Funktion $F : \mathbb{R}^2 \to \mathbb{R}$. Analog nennen wir mehrere Gleichungen ein System von gewöhnlichen Differentialgleichungen, wenn mehrere Funktionen und deren Ableitungen auftreten. Die Ordnung wird nach der höchsten auftretenden Ableitung definiert. Wir bemerken, dass wir jede Gleichung in ein System von Gleichungen erster Ordnung umschreiben können. Tritt etwa die zweite Ableitung u'' auf, so können wir statt u die beiden Funktionen u und v suchen mit der zusätzlichen Gleichung $u' = v$, statt u'' können wir dann v' schreiben.

Wir sprechen von einem Anfangswertproblem, wenn wir neben der Gleichung (C.1) auch noch den Wert $u(t_0)$ gegeben haben. Die Situation aus dem Fundamentalsatz der Integralrechnung ist also die einfachste Version

C.1 Separierbare Differentialgleichungen

Eine der einfachsten Klassen von Differentialgleichungen, die aber häufig auftreten, sind Anfangswertprobleme für separierbare Gleichungen, diese sind von der Form

$$u'(t) = f(t)g(u(t)), \qquad t \in [t_0, t_1]$$

mit $u(t_0) = u_0$. Diese kann man durch einen einfachen Trick lösen. Ist G eine Stammfunktion von $\frac{1}{g}$, so gilt wegen der Kettenregel

$$G(u)' = G'(u)u' = \frac{1}{g(u)}u' = f.$$

Mit dem Fundamentalsatz der Integralrechnung angewandt auf $\tilde{u} = G(u)$ folgt dann

$$G(u(t)) - G(u(t_0)) = \int_{t_0}^{t} f(s)\ ds.$$

Ist G invertierbar, so folgt

$$u(t) = G^{-1}(G(u(t_0)) + \int_{t_0}^{t} f(s)\ ds).$$

Das einfachste Beispiel ist $g(u) = u$, d.h. die lineare Differentialgleichung

$$u'(t) = u(t)f(t).$$

Hier gilt $G(u) = \log u$ und $G^{-1}(v) = e^v$. Damit erhalten wir die Lösung als

$$u(t) = u(t_0)e^{\int_{t_0}^{t} f(s)\ ds}.$$

C.2 Der Satz von Picard-Lindelöf

Der Satz von Peano existiert die Existenz einer Lösung des Anfangswertproblems für die Differentialgleichung (C.1) wenn F eine stetige Funktion ist. Eine verbesserte Version ist der Satz von Picard-Lindelöf:

Satz C.1. *Sei F stetig und genüge einer Lipschitzbedingung bezüglich der zweiten Variable, d.h. es gibt ein $L > 0$ sodass für alle $t \in \mathbb{R}$ und $u_1, u_2 \in \mathbb{R}$ gilt*

$$|F(t, u_1) - F(t, u_2)| \leq L|u_1 - u_2|.$$

Dann hat das Anfangswertproblem für die Differentialgleichung (C.1) *genau eine Lösung.*

C.3 Variation der Konstanten

Zur Lösung einer inhomogenen Differentialgleichung der Form

$$u'(t) = u(t)f(t) + h(t)$$

können wir die sogenannte Methode der Variation der Konstanten anwenden. Diese basiert auf dem Ansatz

$$u(t) = c(t)e^{\int_{t_0}^{t} f(s)\ ds},$$

d.h. wir suchen u als Vielfaches der homogenen Lösung. Wir sehen leicht per Produkt- und Kettenregel

$$u'(t) = c(t)e^{\int_{t_0}^{t} f(s)\ ds} f(t) + c'(t)e^{\int_{t_0}^{t} f(s)\ ds} = u(t)f(t) + c'(t)e^{\int_{t_0}^{t} f(s)\ ds}.$$

Damit erhalten wir für $c(t)$ die Gleichung

$$c'(t) = e^{-\int_{t_0}^{t} f(s)\ ds} h(t),$$

und damit

$$c(t) = c(t_0) + \int_{t_0}^{t} e^{-\int_{t_0}^{\tau} f(s)\ ds} h(\tau)\ d\tau.$$

Berechnen wir nun wieder u durch Multiplikation mit der homogenen Lösung, dann sehen wir

$$u(t) = u(t_0)e^{\int_{t_0}^{t} f(s)\ ds} + \int_{t_0}^{t} e^{\int_{\tau}^{t} f(s)\ ds} h(\tau)\ d\tau.$$

C.4 Das Lemma von Gronwall

Die Argumentation von oben lässt sich auch für eine Ungleichung durchführen, da die Multiplikation mit den positiven Exponentialfunktionen die Positivität ebenso erhalt wie die Integration. Damit folgt aus

$$u'(t) \leq u(t)f(t) + h(t)$$

auch

$$u(t) \leq u(t_0)e^{\int_{t_0}^{t} f(s)\ ds} + \int_{t_0}^{t} e^{\int_{\tau}^{t} f(s)\ ds} h(\tau)\ d\tau.$$

Dies ist die allgemeinste Version des Lemma von Gronwall.

Einen wichtigen Spezialfall erhält man mit konstanten Funktionen $f(t) = a$ und $h(t) = b$, es gilt dann

$$u(t) \leq u(t_0)e^{a(t-t_0)} + b\int_{t_0}^{t} e^{a(t-\tau)}\ d\tau = u(t_0)e^{a(t-t_0)} - \frac{b}{a}(1 - e^{a(t-t_0)}).$$