Diskretisierung und numerische Optimierung

Sommersemester 2023

Prof. Dr. Daniel Tenbrinck
Department of Data Science

Version vom **27. Oktober 2023**

FAU Friedrich-Alexander-Universität
Naturwissenschaftliche Fakultät

DDS Department of DATA SCIENCE

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

Das vorliegende Skript begleitet die Veranstaltung „Diskretisierung und numerische Optimierung“ im Sommersemster 2023 an der FAU Erlangen-Nürnberg. Das Skript basiert hauptsächlich auf einem Vorlesungsskript, welches ich im Sommersemester 2019 gemeinsam mit Prof. Dr. Martin Burger (Universität Hamburg) in einer ersten Version erstellt habe und in den darauffolgenden Jahren weiter angepasst und verbessert habe.

In dieser Vorlesung werden wir einige weiterführende Aspekte der numerischen Mathematik diskutieren, nämlich numerische Verfahren zur Lösung von Optimierungsproblemen und von (gewöhnlichen) Differentialgleichungen. Im ersten Teil der Vorlesung beschäftigen wir uns mit **Optimierungsproblemen**, welche in vielen mathematischen Anwendungsbereichen auftreten, von klassischen ökonomischen Problemen über Materialoptimierung bis hin zu modernen Problemen in der mathematischen Bildverarbeitung und im maschinellen Lernen. Hierbei konzentrieren wir uns auf hauptsächlich auf *unbeschränkte Minimierungsprobleme* der Form

$$\min_{x\in\Omega} F(x),$$

wobei F eine gegebene, zu minimierende Funktion ist und $\Omega \subset \mathbb{R}^n$ eine geeignete Menge von möglichen Eingabevektoren. Methodisch knüpfen wir im Teil zur Optimierung an die *iterativen Methoden zur Lösung von Gleichungssystemen* an, allerdings kommen hier noch einige Aspekte dazu: Mit einem Optimierungsproblem im Hintergrund können wir die Iterationsverfahren geeignet anpassen um tatsächlich mit jeder Iteration die Werte einer gegebenen Funktion zu verkleinern. Darüber hinaus werden wir geeignete *Wahlen der Schrittweite* der Iterationsverfahren kennenzulernen, um die Konvergenz gegen Minimierer oder zumindest stationäre Punkte gewährleisten zu können. Ein weiterer Aspekt ist die *Optimierung unter Nebenbedingung* und die *Minimierung konvexer nicht-differenzierbarer Probleme*, die in vielen modernen Anwendungen auftreten. Dazu werden wir exemplarisch ein spezielles Verfahren kennenlernen.

Der zweite Teil der Vorlesung beschäftigt sich mit **Lösungen von gewöhnlichen Differentialgleichungen**. Wir beginnen mit *einfachen Anfangswertproblemen* der Form

$$u'(t) \;=\; F(u(t),t), \qquad u(0) \;=\; u_0,$$

für eine unbekannte Funktion $u\colon [0,T] \to \mathbb{R}^m$ und die nur für spezielle Formen von $F\colon \mathbb{R}^m \times [0,T] \to \mathbb{R}^m$ gelöst werden können. In einer Vielzahl von Anwendungen treten jedoch *allgemeinere Funktionen* F auf, für die eine analytische Lösung nicht mehr möglich ist, wie zum Beispiel bei den Newtonschen Gesetzen für die Dynamik von mehr als $N > 2$ Teilchen (siehe das *N-Körper-Problem* [N-Körper]). Die entstehenden Gleichungssysteme können dann auch beliebig groß werden, z.B. in der Molekulardynamik, wo $u(t)$ die räumlichen Koordinaten $k \in \mathbb{N}$ verschiedener Teilchen im Zeitverlauf beschreibt und man somit Gleichungssysteme der Größe $n = 3k$ erhält. Andere klassische Anwendungsgebiete gewöhnlicher Differentialgleichungen sind die Modellierung von Populationsdynamiken oder auch von Aktienmärkten, wo meist noch eine zufällige Komponente hinzugefügt wird und man somit *stochastische Differentialgleichungen* erhält. Für solche Anwendungen lassen sich im Allgemeinen nur Lösungen mittels numerischer Verfahren approximieren, die wir in dieser Vorlesung herleiten wollen.

Die numerischen Verfahren zur Lösung von Anfangswertproblemen lassen sich unterschiedlich gestalten. Sie sind einerseits ähnlich zu bereits bekannten Iterationsverfahren, nämlich dann wenn die Ableitung durch *Differenzenquotienten* auf einem Gitter approximiert wird. Andererseits lässt sich ein Bezug zur numerischen Integration herstellen, wenn man die äquivalente Formulierung als *Volterra-Integralgleichung*

$$u(t) \;=\; u_0 + \int_0^t F(u(s), s)\,\mathrm{d}s$$

benutzt und anschließend numerische Quadraturformeln auf das Integral anwendet. Ein wichtiger Aspekt ist in beiden Fällen die *Diskretisierung*, d.h. wir approximieren das Problem in einem endlich-dimensionalen Lösungsraum, z.B. durch Werte auf einem Gitter. Mathematisch stellt sich dann natürlich die Frage ob und in welchem Sinne das diskretisierte Problem gegen das ursprüngliche Problem konvergiert.

Weiterhin werden wir auch *Randwertprobleme* betrachten, die in späteren Vorlesungen zu partiellen Differentialgleichungen führen werden. Ein einfaches Beispiel ist die numerische Lösung einer gewöhnlichen Differentialgleichung der Form

$$-(a(x)u'(x))' + c(x)u(x) \;=\; f(x), \quad x \in (0,1),$$

mit vorgegebenen Randwerten $u(0) = u(1) = 0$. Hier müssen wir die Diskretisierung für das gesamte Intervall $(0,1)$ auf einmal durchführen und nicht von einem Schritt zum Nächsten wie bei den oben beschriebenen Anfangswertproblemen. Diese Diskretisierung liefert uns ein lineares Gleichungssystem, das wir anschließend mit bekannten Methoden der Numerik lösen müssen. Die Abschätzung des Diskretisierungsfehlers erfordert weiterführende Methoden, welche wir im Laufe der Vorlesung kennenlernen werden.

Sollten Ihnen beim Studium dieses Skripts inhaltliche oder sprachliche Fehler auffallen, so würde ich mich über einen Hinweis per Email an `daniel.tenbrinck@fau.de` freuen. Ich wünsche Ihnen viel Erfolg und viel Spaß in der Vorlesung!

Prof. Dr. Daniel Tenbrinck — Erlangen, 27.10.2023

Kapitel 2

Numerische Optimierung

Optimierung ist ein omnipräsentes Phänomen, dass nicht nur in der abstrakten Welt der Mathematik existiert. Viel mehr stellt es ein naturgegebenes Prinzip dar, welches überall um uns herum zur Anwendung kommt. In der Physik beispielsweise spielt Optimierung eine wesentliche Rolle bei der Modellierung von Energiezuständen auf unterschiedlichen Skalen: Moleküle formieren sich in einer Art, die die Gesamtenergie des Teilchensystems unter Berücksichtigung aller wechselseitigen Kräfte minimiert. Gleichzeitig strebt das Universum mit all seinen Planeten, Sternen und Galaxien nach einem Zustand von maximaler Verteilung, beschrieben durch die thermodynamische Größe der Entropie. Auch hier folgt die Zunahme der Entropie dem Prinzip der Energieminimierung des Gesamtsystems. Menschen betreiben seit jeher Optimierung in den verschiedensten Anwendungen, oft mit unterschiedlichen Motivationen. Flugzeuge werden von Ingenieuren so entworfen und gebaut, dass sie möglichst stromlinienförmig aussehen, um damit den Reibungswiderstand in der Luft zu minimieren und gleichzeitig den nötigen Auftrieb für einen sicheren Flug zu erzeugen. Fondmanager streben danach Portfolios zu erstellen, deren Gewinn möglichst maximal ist und dennoch Spekulationsrisiken vermeiden. Die gesamte Automatisierung der Industrie, angefangen bei den ersten Manufakturen hin zu modernen vollautomatischen Roboterfabriken, dient lediglich dem Prinzip der Gewinnmaximierung durch Minimierung der Produktionskosten.

Es ist also nicht wirklich überraschend, dass sich ein gesamtes Teilgebiet der Angewandten Mathematik mit der Theorie der Optimierung befasst und somit dazu beiträgt viele Optimierungsprobleme besser zu verstehen und zu lösen. Im folgenden Abschnitt wollen wir uns hauptsächlich mit der unbeschränkten (oder: unrestringierten) Optimierung beschäftigen und uns nützliche Werkzeuge zum Lösen von numerischen Problemstellungen herleiten. Nach einer *allgemeinen mathematischen Einführung* in Abschnitt 2.1 beginnen wir in Kapitel Abschnitt 2.2 mit einer Klasse von Algorithmen, die einer einfachen Idee folgen: die *numerischen Abstiegsverfahren*. In Abschnitt 2.3 behandeln wir insbesondere das Verfahren der *konjugierten Gradienten*, welches zur iterativen Lösung von großen, linearen Gleichungssystemen mit besonderen Eigenschaften genutzt werden kann. Zum Schluss untersuchen wir in Abschnitt 2.5 zwei moderne Optimierungsalgorithmen zur Lösung von *konvexen, nicht-differenzierbaren Problemen*.

2.1 Mathematische Grundlagen

Im Folgenden wollen wir die mathematischen Grundlagen zur Untersuchung von allgemeinen Optimierungsproblemen einführen. Wir folgen hierbei zu großen Teilen der Notation von Nocedal und Wright in [NW99]. Wir beginnen mit der Definition des allgemeinen Optimierungsproblems, welches wir im Verlauf der Vorlesung noch weiter konkretisieren werden.

DEFINITION 2.1: Allgemeines Optimierungsproblem.

Sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Funktion, welche wir *Zielfunktion* nennen. Unser Ziel ist es einen unbekannten Vektor $x \in \Omega$, auch *Parametervektor* genannt, zu finden, welcher das folgende **allgemeine Optimierungsproblem** löst:

$$\min_{x\in\Omega} F(x) \qquad \text{mit} \qquad \begin{cases} c_i(x) = 0, & i \in \mathcal{E}, \\ c_i(x) \geq 0, & i \in \mathcal{I}. \end{cases} \tag{2.1}$$

Die reellwertigen Funktionen $c_i\colon \Omega \to \mathbb{R}, i = 1, \ldots, m$ bilden einen Vektor von *Nebenbedingungen*, welcher das Optimierungsproblem restringiert. Die Indexmengen $\mathcal{E}$ und $\mathcal{I}$ legen hierbei fest, ob es ich bei der jeweiligen Nebenbedingung um eine Gleichung oder eine Ungleichung handelt.

Zur Veranschaulichung betrachten wir ein zweidimensionales Beispiel für ein beschränktes, nichtlineares Optimierungsproblem.

BEISPIEL 2.2: Beschränktes Optimierungsproblem.

Wir betrachten das folgende mathematische Problem:

$$\min_{x\in\mathbb{R}^2} (x_1 - 2)^2 + (x_2 - 1)^2$$

unter den Nebenbedingungen $x_1^2 - x_2 \leq 0$ und $x_1 + x_2 \leq 2$. Wir können dieses Problem in die allgemeine Form des Optimierungsproblems (2.1) umschreiben als:

$$\min_{x\in\mathbb{R}^2} F(x) = \min_{x\in\mathbb{R}^2} (x_1 - 2)^2 + (x_2 - 1)^2, \qquad \text{mit} \qquad \begin{cases} c_1(x) = -x_1^2 + x_2 \geq 0, \\ c_2(x) = -x_1 - x_2 + 2 \geq 0. \end{cases}$$

Hierbei gilt für die Indexmengen $\mathcal{I} = \{1, 2\}$ und $\mathcal{E} = \emptyset$. Visualisiert man die Niveaulinien der Zielfunktion F zusammen mit den Nebenbedingungen, so erkennt man direkt, dass der globale Minimierer der quadratischen Zielfunktion F, nämlich $x = (x_1, x_2)^T = (2, 1)^T$, nicht in der erlaubten Menge der Parameter liegt, welche durch die Nebenbedingungen beschrieben ist. Trotzdem existiert ein eindeutiges globales Minimum des beschränkten Optimierungsproblems wie man in Abb. 2.1 sehen kann, nämlich $F(x^*) = 1$ für den Minimierer $x^* = (x_1^*, x_2^*)^T = (1, 1)^T$.

ToDo!

Abbildung 2.1: Visualisierung der Niveaumengen der Zielfunktion F und den Nebenbedingungen c_1 und c_2 des restringierten Optimierungsproblems in Beispiel 2.2.

Im Rahmen dieser Vorlesung wollen wir uns zunächst auf eine bestimmte Klasse von allgemeinen Optimierungsproblemen konzentrieren, den *unbeschränkten* oder *unrestringierten* Optimierungsproblemen.

DEFINITION 2.3: Unbeschränkte Optimierung.

Liegt ein allgemeines Optimierungsproblem der Form (2.1) ohne Nebenbedingungen vor, d.h., für die Indexmengen gilt $\mathcal{E} = \mathcal{I} = \emptyset$, so sprechen wir von einem **unbeschränkten** oder **unrestringierten Optimierungsproblem**.

BEMERKUNG 2.4 (Relaxation). Häufig lassen sich restringierte Optimierungsprobleme in unrestringierte Optimierungsprobleme überführen, indem man zusätzliche Strafterme zur Zielfunktion hinzufügt, die eine Verletzung der ursprünglichen Nebenbedingungen zwar mit Kosten belegt, diese jedoch grundsätzlich erlaubt. Hierbei spricht man auch von *relaxierten Optimierungsproblemen*.
Hat man beispielsweise das folgende restringierte Optimierungsproblem vorliegen

$$\min_{x\in\mathbb{R}} \{F(x) := e^x\} \qquad \text{mit} \qquad c(x) := x \geq 0,$$

so lässt sich stattdessen auch folgendes relaxiertes Optimierungsproblem ohne Nebenbedingungen betrachten:

$$\min_{x\in\mathbb{R}} \left\{G(x) := e^x + \lambda \cdot (\min(0, x))^2\right\},$$

wobei $\lambda \in \mathbb{R}^+$ den Strafterm zur Einhaltung der Nebenbedingung gewichtet. △

Daher gehen wir für den weiteren Verlauf der Vorlesung immer (wenn nicht anders beschrieben) von einem unrestringierten Optimierungsproblem aus.

Neben der Unterscheidung von Optimierungsproblemen in beschränkte und unbeschränkte Formulierungen, lassen sich noch weitere Kriterien zur Charakterisierung eines Optimierungsproblems heran ziehen:

- **Anzahl der unbekannten Parameter**, z.B. groß oder klein
- **Eigenschaften der Zielfunktion**, z.B. Linearität, Beschränktheit, Konvexität, Stetigkeit, Differenzierbarkeit
- **Charakteristik des Optimums**, z.B. Sattelpunkt, lokales oder globales Optimum
- **Modelleigenschaften**, z.B. stochastisch oder deterministisch

Da wir uns intensiv mit der Bestimmung und numerischen Approximation von Optima beschäftigen werden, macht es Sinn diese zuerst formal zu beschreiben.

DEFINITION 2.5: Lokales und globales Minimum.

Sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Zielfunktion. Wir nennen einen Punkt $x^* \in \Omega$ einen **lokalen Minimierer** der Zielfunktion F, falls es eine lokale Umgebung $U \subset \Omega$ von $x^* \in U$ gibt, so dass für alle $x \in U$ gilt:

$$F(x^*) \leq F(x), \quad \forall x \in U. \tag{2.2}$$

Den Funktionswert $F(x^*) \in \mathbb{R}$ nennen wir in diesem Fall ein **lokales Minimum** der Zielfunktion F.
Wir nennen $x^* \in \Omega$ einen **globalen Minimierer** von F, falls die Ungleichung (2.2) für jede beliebige Umgebung $U \subset \Omega$ gilt und somit insbesondere für $U = \Omega$. Den Funktionswert $F(x^*) \in \mathbb{R}$ nennen wir in diesem Fall ein **globales Minimum** der Zielfunktion F.

DEFINITION 2.6: Striktes Minimum.

Sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Zielfunktion. Sei $x^{\in}\Omega$ ein lokaler Minimierer der Zielfunktion F in einer offenen Umgebung $U \subset \Omega$.
Wir nennen $F(x^*)$ ein **striktes Minimum** von F, falls gilt

$$F(x^*) < F(x) \qquad \forall\, x \in U \setminus \{x^*\}.$$

BEMERKUNG 2.7 (Äquivalenz von Minimierung und Maximierung). In obiger Definition sprechen wir nur von Minima, jedoch ist klar, dass sich jedes Maximierungsproblem durch einen Vorzeichenwechsel leicht in ein Minimierungsproblem umformulieren lässt, d.h.,

$$\max_{x\in\Omega} F(x) \quad \Leftrightarrow \quad \min_{x\in\Omega} -F(x) \;=:\; \min_{x_\in\Omega} G(x).$$

Formal lassen sich folgende Gleichungen zeigen:

$$\begin{aligned} \max_{x\in\Omega} F(x) &= -\min_{x\in\Omega} -F(x), \\ \operatorname{argmax}_{x\in\Omega} F(x) &= \operatorname{argmin}_{x\in\Omega} -F(x). \end{aligned}$$

△

Da wir nun eine Charakterisierung von lokalen Minima haben, können wir mit folgendem Satz die notwendigen Bedingungen für solch ein lokales Minimum angeben.

THEOREM 2.8: Notwendige Optimalitätsbedingungen 1. Ordnung.

Sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Zielfunktion. Sei $x^* \in \Omega$ ein lokaler Minimierer von F in Ω und die Zielfunktion F sei stetig differenzierbar in einer lokalen, offenen Umgebung $U \subset \Omega$ des Punkts x^*.
Dann gilt $\nabla F(x^*) = 0$.

Beweis. Wir führen einen Beweis durch Widerspruch. Nehmen wir also an, dass $x^* \in \mathbb{R}$ ein lokaler Minimierer von F sei, jedoch aber $\nabla F(x^*) \neq 0$ gelte. Wir wählen den Richtungsvektor $\mathbf{p} \in \mathbb{R}^n$ mit $\mathbf{p} := -\nabla F(x^*) \neq 0$. Es ist somit klar, dass

$$\langle \mathbf{p}, \nabla F(x^*) \rangle \;=\; -\langle \nabla F(x^*), \nabla F(x^*) \rangle \;=\; -||\nabla F(x^*)||^2 \;<\; 0.$$

Da ∇F nach Vorraussetzung stetig in einer lokalen Umgebung $U \subset \Omega$ von x^* ist existiert ein $T > 0$, so dass auch gilt:

$$\langle \mathbf{p}, \nabla F(x^* + t\mathbf{p}) \rangle \;<\; 0, \qquad \text{für alle } t \in [0, T].$$

Nach dem Satz von Taylor gilt aber auch für jedes $\tilde{t} \in (0, T]$:

$$F(x^* + \tilde{t}\mathbf{p}) \;=\; F(x^*) + \underbrace{\tilde{t}\langle \mathbf{p}, \nabla F(x^* + t\mathbf{p}) \rangle}_{<\,0}, \qquad \text{für ein } t \in (0, \tilde{t}).$$

Somit gilt also $F(x^* + \tilde{t}\mathbf{p}) < F(x^*)$ für alle $\tilde{t} \in (0, T]$ und wir haben offenbar eine Richtung $\mathbf{p} \in \mathbb{R}^n/\{0\}$ gefunden in der die Funktionswerte von F abnehmen. Also ist $x^* \in \Omega$ kein lokaler Minimierer von F.

Das ist aber ein Widerspruch zur Annahme und somit ist die Behauptung bewiesen. □

Die für die Optimierung interessanten Punkte $x^* \in \Omega$, die die notwendigen Optimalitätsbedingungen aus Theorem 2.8 erfüllen, nennen wir stationäre Punkte.

DEFINITION 2.9: Stationärer Punkt.

Sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Zielfunktion. Wir nennen einen Punkt $x^* \in \Omega$ **stationären Punkt** von F, falls F in einer lokalen, offenen Umgebung $U \subset \Omega$ von x^* stetig differenzierbar ist und der Punkt x^* die Bedingung $\nabla F(x^*) = 0$ erfüllt.

Mit der Definition von stationären Punkten lässt sich folgendes Korollar direkt ableiten.

KOROLLAR 2.10.

Jeder lokale Minimierer $x^* \in \Omega$ einer Zielfunktion $F\colon \Omega \to \mathbb{R}$ ist ein stationärer Punkt.

BEMERKUNG 2.11 (Sattelpunkte). Die Umkehrung der Aussage in Theorem 2.8 gilt im Allgemeinen nicht. Man betrachte zum Beispiel die Zielfunktion $F(x) := -x^3$. Diese besitzt einen stationären Punkt in $x^* = 0$, d.h., es gilt $\nabla F(0) = 0$. Dennoch handelt es sich hierbei nicht um einen lokales Minimierer, sondern lediglich um einen **Sattelpunkt**. Aus diesem Grund handelt es sich nur um notwendige und nicht hinreichende Bedingungen. △

Bei der Suche nach lokalen Minima einer Zielfunktion F lässt sich ein weiteres Kriterium anwenden, welches die zweite Ableitung der Funktion verwendet.

THEOREM 2.12: Notwendige Optimalitätsbedingungen 2. Ordnung.

Sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Zielfunktion. Sei $x^* \in \Omega$ ein lokaler Minimierer von F in Ω und F sei zweimal stetig differenzierbar in einer lokalen Umgebung $U \subset \Omega$ von x^*, d.h., die Hessematrix $\nabla^2 F$ von F ist stetig in der offenen Umgebung $U \subset \Omega$ von x^*.
Dann gilt $\nabla F(x^*) = 0$ und $\nabla^2 F(x^*)$ ist positiv semidefinit, d.h., es gilt

$$\langle \mathbf{p}, \nabla^2 F(x^*)\mathbf{p}\rangle \geq 0 \qquad \forall\, \mathbf{p} \in \mathbb{R}^n.$$

Beweis. In den Übungsaufgaben zu zeigen. □

Schlussendlich wollen wir auch eine hinreichende Bedingung für das Vorliegen eines lokalen Minimums angeben.

THEOREM 2.13: Hinreichende Optimalitätsbedingungen 2. Ordnung.

Sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Zielfunktion. Sei $x^* \in \Omega$ ein lokaler Minimierer von F in Ω und F sei zweimal stetig differenzierbar in einer lokalen Umgebung $U \subset \Omega$ des Punkts x^*, d.h., die Hessematrix $\nabla^2 F$ von F sei stetig in einer offenen Umgebung $U \subset \Omega$ von x^*. Außerdem gelte

(i) $\nabla F(x^*) = 0$,

(ii) $\nabla^2 F(x^*)$ ist positiv definit.

Dann ist $F(x^*)$ ein striktes lokales Minimum von F.

Beweis. Da die Hessematrix $\nabla^2 F$ von F stetig und positiv definit in $x^* \in \Omega$ ist nach Voraussetzung können wir einen Radius $r > 0$ finden, so dass $\nabla^2 F(x)$ positiv definit ist für alle $x \in B_r(x^*)$. Für jeden Vektor $\mathbf{p} \in \mathbb{R}^n/\{0\}$ mit $||\mathbf{p}|| < r$ gilt dann nach dem Satz von Taylor:

$$F(x^* + \mathbf{p}) = F(x^*) + \underbrace{\langle \mathbf{p}, \nabla F(x^*)\rangle}_{=0} + \frac{1}{2}\langle \mathbf{p}, \nabla^2 F(x^* + t\mathbf{p})\mathbf{p}\rangle, \qquad \text{für ein } t \in (0,1).$$

Da $||t\mathbf{p}|| < r$ ist nach Konstruktion wissen wir, dass

$$\langle \mathbf{p}, \nabla^2 F(x^* + t\mathbf{p})\mathbf{p} \rangle \; > \; 0$$

gilt und somit schon $F(x^* + \mathbf{p}) > F(x^*)$ gelten muss. Da $\mathbf{p} \in \mathbb{R}^n/\{0\}$ mit $||\mathbf{p}|| < r$ beliebig gewählt war handelt es sich bei $F(x^*)$ um ein striktes lokales Minimum der Zielfunktion F. □

BEMERKUNG 2.14 (Definitheit der Hessematrix). Die in Theorem 2.13 genannten Bedingungen sind hinreichend, jedoch nicht notwendig für das Vorliegen eines strikten lokalen Minimums. Dies sieht man ein, wenn man beispielsweise die Zielfunktion $F(x) := x^4$ betrachtet. F besitzt ein striktes lokales Minimum in $x^* = 0$ und es gilt $\nabla F(0) = 0$, jedoch verschwindet die zweite Ableitung $\nabla^2 F(0) = 0$ und ist somit nicht positiv definit. △

Eine äußerst wertvolle Eigenschaft bei der Optimierung ist die Konvexität einer Zielfunktion, da jedes lokale Optimum einer konvexen Funktion bereits ein globales Optimum ist.

DEFINITION 2.15: Konvexität.

Sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Zielfunktion. Wir nennen F **konvex** wenn für beliebige Vektoren $x, y \in \Omega$ die folgende Ungleichung für alle $0 \leq \alpha \leq 1$ gilt:

$$F(\alpha x + (1-\alpha)y) \; \leq \; \alpha F(x) + (1-\alpha)F(y).$$

Anschaulich bedeutet Konvexität einer Funktion F, dass jede Verbindungsgerade zwischen zwei Punkten $F(x)$ und $F(y)$ in diesem Intervall oberhalb des Graphen der Funktion F verläuft.

2.2 Abstiegsverfahren

Zu Anfang dieser Vorlesung möchten wir eine Klasse von Algorithmen zur Optimierung von Funktionen besprechen, die einer simplen und anschaulichen Idee folgen: die sogenannten *Abstiegsverfahren* oder auch *Liniensuchverfahren* (im Englischen: *line search methods*). Diese wurden bereits kurz im Zusammenhang mit dem Gauss-Newton Verfahren in der Vorlesung „Einführung in die Numerik“ in [Numerik 1, Kapitel 4.5] erwähnt, jedoch nicht ausführlich diskutiert. Dies wollen wir im Folgenden nachholen.

Sei im Folgenden $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und sei $F\colon \Omega \to \mathbb{R}$ eine differenzierbare, reellwertige Zielfunktion. Die allgemeine Idee der Abstiegsverfahren ist es nun ausgehend von einem aktuellen Punkt $x_k \in \Omega$ einen Schritt in eine Richtung $p_k \in \mathbb{R}^n$ zu machen, so dass der Funktionswert von F in dem neuen Punkt x_{k+1} abnimmt. Das bedeutet wir versuchen ein **allgemeines Abstiegsverfahren** der Form

$$x_{k+1} \; = \; x_k + \alpha_k p_k, \quad \alpha_k \in \mathbb{R}^+, \tag{2.3}$$

zu konstruieren, das in jedem Schritt den Funktionswert $F(x_k)$ verringert bis keine Verbesserung mehr möglich ist. Die Schrittweite $\alpha_k > 0$ in Richtung $p_k \in \mathbb{R}^n$ wird häufig in jedem Schritt des Abstiegsverfahrens neu gewählt.

2.2.1 Gradientenabstiegsverfahren

Zuerst beschäftigen wir uns mit dem wohl bekanntesten Abstiegsverfahren, dem *Gradientenabstiegsverfahren* (im Englischen: *gradient descent (GD)*). Dieses wird wegen seiner Einfachheit für viele numerische Optimierungsprobleme verwendet.

Da F differenzierbar ist können wir dessen Gradienten in jedem Punkt $x \in \Omega$ betrachten. Wir gehen im Folgenden immer davon aus, dass wir den Gradienten ∇F der Zielfunktion F analytisch bestimmen können und eine Auswertung $\nabla F(x_k)$ für eine Folge von Punkten $(x_i)_{i\in\mathbb{N}} \subset \Omega$ numerisch durchführbar ist. Wir werden also nicht versuchen den Gradienten mit Hilfe von *finiten Differenzenverfahren* numerisch zu approximieren.

Der Gradient $\nabla F(x)$ beschreibt bekanntlich eine Richtung des stärksten Anstiegs der Funktion F im Punkt x und dementsprechend zeigt der negative Gradient $-\nabla F(x)$ in die Richtung des stärksten Abstiegs. Das lässt sich auch formal zeigen indem wir uns die Taylorapproximation erster Ordnung der Funktion F in allgemeine Richtung p mit Schrittweite $\alpha > 0$ für den k-ten Schritt des Abstiegsverfahren näher anschauen. Hier gilt nämlich

$$F(x_k + \alpha p) \;=\; F(x_k) + \alpha\langle \nabla F(x_k), p\rangle + \mathcal{O}(\nabla^2 F(x_k)),$$

wobei $\nabla^2 F$ die Hesse-Matrix der Zielfunktion F bezeichnet. Für kleine Schrittweiten α wird klar, dass die Änderung der Funktionswerte von F im Wesentlichen von der Größe des Terms $\langle \nabla F(x_k), p\rangle$ bestimmt wird. Wir können also für eine maximale Verringerung der Funktion F nach derjenigen Richtung mit Einheitslänge suchen, die das folgende Minimierungsproblem löst:

$$\min_{p\in\mathbb{R}^n} \langle \nabla F(x_k), p\rangle \quad \text{mit} \quad ||p|| = 1. \tag{2.4}$$

Da außerdem gilt

$$\langle \nabla F(x_k), p\rangle \;=\; ||\nabla F(x_k)|| \cdot ||p|| \cdot \cos(\theta), \tag{2.5}$$

wobei θ der Winkel zwischen den Vektoren $\nabla F(x_k)$ und p bildet, können wir das Problem (2.4) umschreiben zu

$$\min_{\theta\in[0,2\pi)} ||\nabla F(x_k)|| \cdot \cos(\theta).$$

Der Kosinus nimmt sein Minimum von $\cos(\theta) = -1$ in $\theta = \pi$ an. Eingesetzt in (2.5) erhalten wir damit, dass die optimale Richtung $p \in \mathbb{R}^n$ folgenden Zusammenhang erfüllen muss:

$$\langle p, \frac{\nabla F(x_k)}{||\nabla F(x_k)||}\rangle \;=\; -1.$$

Da die unbekannte Richtung $p \in \mathbb{R}^n$ aber normiert sein soll, folgt damit aber auch schon, dass gilt

$$p \;=\; -\frac{\nabla F(x_k)}{||\nabla F(x_k)||}. \tag{2.6}$$

Das bedeutet, dass der Funktionswert von F am stärksten in Richtung des negativen Gradienten abnimmt.

Da wir daran interessiert sind die Zielfunktion F schnellstmöglich zu minimieren, macht es Sinn in eben dieser Richtung nach einem lokalen Minimum zu suchen. Aus dieser Idee heraus lässt sich bereits ein sehr simpler Algorithmus zur Minimierung von F formulieren. Sei $x_0 \in \Omega$ ein beliebiger Startwert. Dann können wir iterativ eine Folge von Punkten $x_1, x_2, \ldots$ in Ω bestimmen, so dass die entsprechenden Funktionswerte von F monoton fallen sollten:

$$x_{k+1} \;=\; x_k - \nabla F(x_k). \tag{2.7}$$

Intuitiv stoppt man das Iterationsschema (2.7) sobald die Folge der Funktionswerte $F(x_k)$ nicht mehr kleiner wird. Man sieht leicht ein, dass das simple Iterationsschema (2.7) ein Spezialfall des allgemeinen Abstiegsverfahrens (2.3) mit einer festen Schrittweite $\alpha_k = 1$ und Richtungsvektor $p_k = -\nabla F(x_k)$ ist. Diese einfache Methode lässt sich durch folgenden Algorithmus implementieren.

ALGORITHMUS 2.16: Simples Gradientenabstiegsverfahren.

function $[x^*, F(x^*)]$ =`gradientDescentSimple`$(F, \nabla F, x_0)$

Initialisierung
$x_k = x_0$
$F(x_{k+1}) = -$`Inf`

while $F(x_{k+1}) - F(x_k) < 0$ **do**
  # Update in Richtung des größten Gradientenabstiegs
  $x_{k+1} = x_k - \nabla F(x_k)$
end while

Ausgabe des letzten Punktes
$x^* = x_k$
$F(x^*) = F(x_k)$

Unglücklicherweise ist Algorithmus 2.16 in dieser Form praktisch nicht anwendbar. Warum dies so ist sieht man leicht an folgendem Beispiel.

BEISPIEL 2.17: Simpler Gradientenabstieg.

Sei $\Omega = \mathbb{R}$ und sei $F(x) := |x|$. Man beachte, dass F überall differenzierbar ist, außer an der Stelle $x = 0$. Das eindeutig bestimmte, globale Minimum der konvexen, nichtlinearen Funktion F wird ebenfalls in $x^* = 0$ angenommen.
Definiert man $\nabla F(0) := 0$ in der Singularität, so erhält man das Minimum $x^* = 0$ durch Algorithmus 2.16 nur für Startwerte $x_0 \in \mathbb{Z}$. Wähle zum Beispiel den Startwert $x_0 = 0.5$, so terminiert das Gradientenabstiegsverfahren bereits nach dem ersten Schritt ohne eine gute Näherung an $x^* = 0$ zu liefern.

Wie das Beispiel 2.17 zeigt, besteht bei dem Gradientenabstiegsverfahren in Algorithmus 2.16 die Gefahr einen stationären Punkt und somit ein potentielles Minimum zu überspringen. Aus diesem Grund kommt man auf die Idee die Schrittweite $\alpha_k > 0$ in (2.3) **genügend klein** zu wählen. Damit diese feste Schrittweite unabhängig von der Magnitude des Gradienten ∇F ist, normiert man in der Regel die Richtung des steilsten Gradientenabstiegs durch die Norm des Gradienten, d.h., wir erhalten eine **steuerbare Version des Gradientenabstiegsverfahrens** in (2.7) durch:

$$x_{k+1} = x_k - \tau \frac{\nabla F(x_k)}{||\nabla F(x_k)||}, \quad \tau > 0. \tag{2.8}$$

Das Iterationsschema (2.9) ist wiederum ein Spezialfall des allgemeinen Abstiegsverfahrens in (2.3) für eine feste Schritteweite $\alpha_k = \tau > 0$ und den Richtungsvektor $p_k = -\nabla F(x_k)/||\nabla F(x_k)||$.

Es ist klar, dass durch eine kleinere Schrittweite $\tau > 0$ ein stationärer Punkt $x^* \in \Omega$ der Zielfunktion F immer besser angenähert werden kann. Leider erhöht sich aber gleichzeit die benötigte Iterationszahl zur Erreichung einer erwünschten Genauigkeit $|F(x^*) - F(x_k)| < \epsilon$ je kleiner man die Schrittweite τ wählt. Man muss also bei der Wahl der Schrittweite einen Kompromiss zwischen Genauigkeit der numerischen Approximation und der Laufzeit des Verfahrens eingehen.

Eine weiterführende Idee ist es die Schrittweiten **adaptiv** zu wählen, dass heißt man passt sie innerhalb des Iterationsschemas an die Funktionswerte von F geeignet an. Ein Iterationsschema, dass eine immer kleiner werdende Schrittweite $\alpha_k > 0$ verwendet, lässt sich für einen fest gewählten Reduktionsfaktor $0 < \sigma < 1$ wie folgt angeben:

$$\begin{aligned} \alpha_{k+1} &= \begin{cases} \alpha_k, & \text{falls } F\left(x_k - \alpha_k \frac{\nabla F(x_k)}{||\nabla F(x_k)||}\right) < F(x_k), \\ \sigma\alpha_k, & \text{sonst.} \end{cases}, \\ x_{k+1} &= x_k - \alpha_{k+1} \frac{\nabla F(x_k)}{||\nabla F(x_k)||}. \end{aligned} \tag{2.9}$$

Das adaptive Gradientenabstiegsverfahren in (2.9) lässt sich mit folgendem Algorithmus umsetzen.

ALGORITHMUS 2.18: Adaptives Gradientenabstiegsverfahren.

function $[x^*, F(x^*)]$ =`gradientDescentAdaptive`$(F, \nabla F, x_0, \alpha_0, \sigma, \epsilon)$

Initialisierung
$\alpha_k = \alpha_0$
$x_k = x_0$
$F(x_{k+1}) = +$`Inf`

Iteriere bis gewünschte Genauigkeit erreicht ist

while $|F(x_{k+1}) - F(x_k)| > \epsilon$ **do**
  while $F(x_k - \alpha_k \nabla F(x_k)/||\nabla F(x_k)||) > F(x_k)$ **do**
    # Verringere Schrittweite um Faktor σ
    $\alpha_k = \sigma \alpha_k$
  end while
  # Update in Richtung des größten Gradientenabstiegs
  $x_{k+1} = x_k - \alpha_k \nabla F(x_k)/||\nabla F(x_k)||$
end while

Ausgabe des letzten Punktes
$x^* = x_k$
$F(x^*) = F(x_k)$

BEMERKUNG 2.19 (Zurücksetzen der Schrittweite). In manchen Anwendungsfällen macht es Sinn die Schrittweite α_{k+1} in (2.9) in jedem Schritt wieder auf den initialen Wert $\alpha_0 > 0$ zurückzusetzen, um eine verbesserte Konvergenzgeschwindigkeit zu erhalten. Dies macht vor allem dann SInn, wenn eine Auswertung der Zielfunktion F und ihres Gradienten ∇F numerisch günstig zu realisieren ist. △

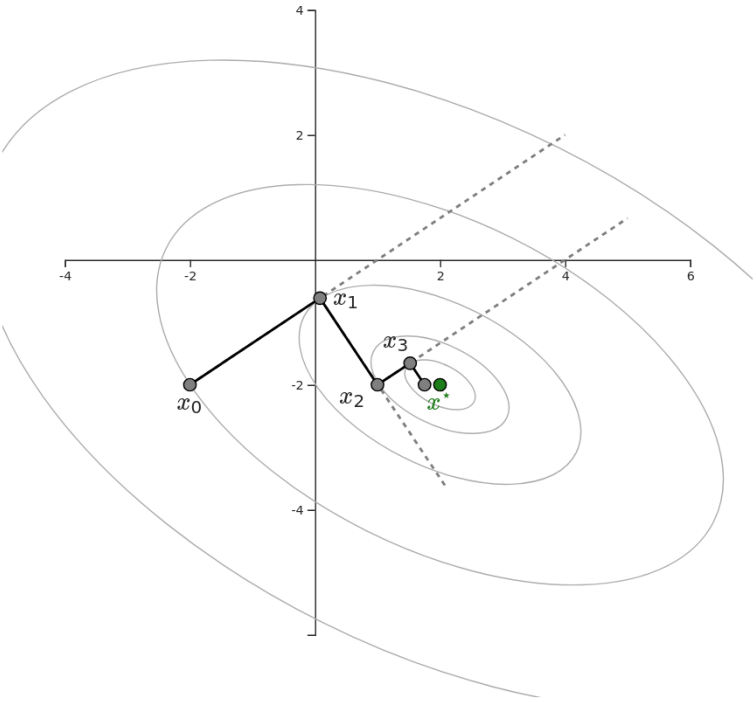

Abbildung 2.2: Approximation des Minimierers einer Funktion F in zwei Variablen mit Hilfe des adaptiven Gradientenverfahrens (2.9).

In Abb. 2.2 ist ein typischer Verlauf des adaptiven Gradientenverfahrens in Algorithmus 2.18 bei der Minimierung einer konvexen Zielfunktion $F\colon \mathbb{R}^2 \to \mathbb{R}$ zu sehen. Man erkennt, dass die Schrittweiten immer kleiner werden, je näher man sich dem lokalen Minimum x_* nähert. Außerdem sieht man, dass die Richtung des steilsten Gradientenabstiegs immer orthogonal zu den Niveaulinien der zu minimierenden Funktion steht.

BEMERKUNG 2.20. Das in diesem Abschnitt beschriebene Gradientenabstiegsverfahren mit adaptiver Schrittweite $\alpha_k > 0$ ist ein gängiger Algorithmus zur Minimierung einer Funktion F, wenn deren Ableitung ∇F bekannt und numerisch günstig zu berechnen ist. Dennoch gibt es Situationen in denen es ratsam ist alternative Optimierungsalgorithmen zu verwenden. Zum Beispiel ist ein häufiges Problem des Gradientenverfahrens die starke Verlangsamung in der Nähe eines Sattelpunktes, was zu sehr langen Laufzeiten des Algorithmus führt. Außerdem passiert die Minimierung einer Funktion F mit Hilfe des Gradientenabstiegsverfahrens in der Regel entlang eines Zickzack-Pfades (siehe Abb. 2.2), welcher in den meisten Fällen offensichtlich suboptimal ist. Aus diesen Gründen wollen wir uns in den nächsten Abschnitten mit alternativen Minimierungsmethoden beschäftigen. △

2.2.2 Koordinatenabstiegsverfahren

Eine weitere Variante des in Abschnitt 2.2.1 behandelten Gradientenabstiegsverfahrens ist das *Koordinatenabstiegsverfahren* (im Englischen: *coordinate descent* (CD)).

Die grundlegende Idee des Koordinatenabstiegsverfahrens ist es in jedem Schritt des Iterationsschemas eine *Koordinatenrichtung* auszuwählen und einen Abstieg in diese Richtung durchzuführen. Damit lässt sich ein möglicherweise kompliziertes multivariates Optimierungsproblem durch eine Reihe von einfachen univariaten Optimierungsproblemen behandeln. Die Auswahl der Koordinatenrichtung kann entweder mit Hilfe einer Auswahlregel, z.B. mit einem *Rundlaufverfahren*, oder aber *zufällig* geschehen.

Wir wollen im Folgenden den Fall einer **zufälligen Wahl der Koordinatenrichtung** diskutieren. Für einen zufälligen Index $j \in \{1, \dots, n\}$ und den entsprechenden zufälligen Einheitsvektor $e_j \in \mathbb{R}^n$ lässt sich das Koordinatenabstiegsverfahren schreiben als:

$$x_{k+1} \;=\; x_k - \alpha_k \langle \nabla F(x_k), e_j \rangle e_j \;=\; x_k - \alpha_k \frac{\partial F}{\partial x^i}(x_k) e_j. \tag{2.10}$$

Das bedeutet, dass man in jeder Iteration nur eine Koordinate des aktuellen Parametervektors $x_k \in \Omega$ verändern muss. Die Schrittweite $\alpha_k > 0$ in (2.10) kann hierbei ähnlich wie in Abschnitt 2.2.1 *fest* oder *adaptiv* gewählt werden. Das Koordinatenabstiegsverfahren in (2.10) mit adaptiver Schrittweite lässt sich mit folgendem Algorithmus umsetzen.

ALGORITHMUS 2.21: Koordinatenabstiegsverfahren.

function $[x^*, F(x^*)]$ =`coordinateDescentStochastic`$(F, \nabla F, x_0, \alpha_0, \sigma, \epsilon)$

```
# Initialisierung
α_k = α_0
x_k = x_0
F(x_{k+1}) = +Inf

# Iteriere bis gewünschte Genauigkeit erreicht ist
while |F(x_{k+1}) − F(x_k)| > ε do

  # Wähle zufällige Koordinatenrichtung
  i = randomDraw([1 : n])

  # Berechne Ableitung in Koordinatenrichtung
  p_k = ∂/∂x_k^i F(x_k) · e_i

  while F(x_k − α_k p_k) > F(x_k) do
    # Verringere Schrittweite um Faktor σ
    α_k = σ α_k
  end while

  # Update in Richtung des größten Gradientenabstiegs
  x_{k+1} = x_k − α_k p_k
end while

# Ausgabe des letzten Punktes
x* = x_k
F(x*) = F(x_k)
```

Das Koordinatenabstiegsverfahren benötigt in der Regel deutlich mehr Iterationen als das normale Gradientenabstiegsverfahren und beschreibt häufig noch mehr einen Zickzack-Pfad bei der Minimierung. Dennoch bietet das Verfahren Vorteile gegenüber dem Gradientenabstiegsverfahren gerade in Optimierungsproblemen mit vielen Variablen (z.B. für das Training eines künstlichen neuronalen Netzes), da jedes eindimensionale Optimierungsproblem wesentlich leichter zu lösen ist als die Berechnung des gesamten Gradienten in jedem Schritt.

BEMERKUNG 2.22 (Blockkoordinatenabstiegsverfahren). Um den Zufallseffekten und den damit verbundenen Zickzack Pfad bei der Minimierung der Funktion F durch das Koordinatenabstiegsverfahren entgegen zu wirken, kann man einen Kompromiss zwischen der Verwendung einer einzelnen Koordinatenrichtung und dem gesamten Gradienten eingehen. Hierbei spricht man von den sogenannten *Blockkoordi-*

natenabtiegsverfahren (im Englischen *block coordinate descent* (BCD)). Hierbei wählt man zuerst die Größe $s \in \{1, \dots, n\}$ der Koordinatenblöcke, d.h., die Größe der Teilmenge der verwendeten Richtungsableitungen des Gradienten ∇F. Anschließend wird in jedem Schritt ein Block an Koordinatenrichtungen deterministisch oder zufällig ausgewählt und in dessen Richtung minimiert. Für eine zufällige Wahl der Koordinatenblöcke ergibt sich somit:

$$x_{k+1} \;=\; x_k - \alpha_k \sum_{i=1}^{s} \langle \nabla F(x_k), e_{\sigma(i)} \rangle e_{\sigma(i)}. \tag{2.11}$$

Hierbei ist $\sigma\colon \{1, \dots, n\} \to \{1, \dots, n\}$ eine zufällige Permutation der Indizes $1, \dots, n$. Es ist klar, dass das Koordinatenabstiegsverfahren in (2.10) und das normale Gradientenabstiegsverfahren in (2.9) Spezialfälle des Blockkoordinatenabstiegsverfahrens in (2.11) für Blockgrößen $s = 1$ und $s = n$ sind. △

2.2.3 Stochastisches Gradientenabstiegsverfahren

Eine aktuell weit verbreitete Variante des Gradientenabstiegsverfahrens in (2.9) ist das *stochastische Gradientenabstiegsverfahren* (im Englischen: *stochastic gradient descent* (SGD)). Wie der Name schon verrät handelt es sich hierbei nicht um einen deterministischen Algorithmus. Das bedeutet, dass man bei mehrmaliger Anwendung des Verfahrens bei gleichbleibenden Startbedingungen in der Regel unterschiedliche Ergebnisse in unterschiedlichen Laufzeiten erhält. Was auf den ersten Blick wie ein Nachteil wirkt, kann in manchen Fällen jedoch praktische Eigenschaften mit sich bringen. So kann die Zufallsnatur des stochastischen Gradientenverfahrens dazu führen, dass Sattelpunkte und ungewollte, lokale Minima der Funktion durch die Folge der Punkte vermieden werden. Das Verfahren findet aktuell vor allem beim Training von neuronalen Netzen bei der sogenannten *Backpropagation* in verschiedenen Variationen Anwendung, da man hierdurch dem bekannten Problem des *Übertrainierens* des neuronalen Netzes entgegen wirken kann.

Beim stochastischen Gradientenverfahren geht man davon aus, dass sich die zu minimierende Zielfunktion $F\colon \Omega \to \mathbb{R}$ als eine Summe der folgenden Gestalt schreiben lässt:

$$F(x) \;=\; \sum_{i=1}^{m} F_i(x), \qquad \text{für alle } x \in \Omega. \tag{2.12}$$

Solche Zielfunktionen treten natürlicherweise in vielen Problemstellungen auf, zum Beispiel bei Maximimum-Likelihood Ansätzen oder der Methode der kleinsten Quadrate. Im Bereich des maschinellen Lernens lässt sich der Trainingsfehler über alle Trainingsdaten in der Regel als eine solche Summe schreiben. In diesem Fall lässt sich das normale Gradientenabstiegsverfahren in (2.9) umschreiben zu:

$$x_{k+1} \;=\; x_k - \alpha_k \frac{\nabla F(x_k)}{||\nabla F(x_k)||} \;=\; x_k - \alpha_k \frac{\sum_{i=1}^{m} \nabla F_i(x_k)}{||\sum_{i=1}^{m} \nabla F_i(x_k)||}, \quad \alpha_k > 0. \tag{2.13}$$

Die Idee des stochastischen Gradientenverfahrens ist es nun einen zufälligen Summanden aus (2.12) zu wählen und nur den Gradienten bezüglich dieses Summanden zu

betrachten. Durch diese starke Vereinfachung von (2.13) führt man mit einem zufällig ausgewählten Index $j \in \{1, \dots, m\}$ nun einen Gradientenabstieg der Form

$$x_{k+1} \;=\; x_k - \alpha_k \frac{\nabla F_j(x_k)}{||\nabla F_j(x_k)||}, \quad \alpha_k > 0 \tag{2.14}$$

durch.

BEMERKUNG 2.23. Ähnlich wie im Fall des Koordinatenabstiegsverfahrens in Kapitel 2.2.2, gibt es auch beim stochastischen Gradientenverfahren die Möglichkeit einen Kompromiss zwischen dem normalen Gradientenabstieg in (2.8) und dem auf einen Summanden beschränkten Gradientenabstieg (2.14) einzugehen. Indem man eine zufällige Untermenge von fester Größe $s \in \{1, \dots, m\}$ von Summanden von F auswählt, lässt sich das sogenannte *stochastische Minibatch-Gradientenabstiegsverfahren* formulieren:

$$x_{k+1} = x_k - \alpha_k \frac{\sum_{i=1}^{s} \nabla F_{\sigma(i)}(x_k)}{||\sum_{i=1}^{s} \nabla F_{\sigma(i)}(x_k)||}, \quad \alpha_k > 0.$$

Hierbei ist $\sigma\colon \{1, \dots, m\} \to \{1, \dots, m\}$ eine zufällige Permutation der Indizes $1, \dots, m$. △

2.2.4 Newton Verfahren

In diesem Abschnitt wollen wir uns das bereits bekannte Newton-Verfahren in Erinnerung rufen und dieses geeignet zur Optimierung von nichtlinearen Funktionen verallgemeinern. In [Numerik 1, Kapitel 4.2] haben wir das *Newton Verfahren* zur Approximation von Nullstellen nichtlinearer Gleichungssysteme hergeleitet. Wir haben zunächst die Taylorapproximation einer nichtlinearen Nullstellengleichung $F(x^*) = 0$ von der folgenden Form betrachtet

$$0 \;=\; F(x^*) \;\approx\; F(x) + F'(x)(x^* - x),$$

wobei F' die als regulär angenommene Jacobi-Matrix der differenzierbaren Funktion $F\colon \mathbb{R}^n \to \mathbb{R}^n$ bezeichnet. Hierauf basierend haben wir die folgende **Fixpunktfunktion** als Approximation erster Ordnung angegeben:

$$G(x) \;=\; x - (F'(x))^{-1} F(x), \qquad \text{für } F'(x) \text{ regulär.} \tag{2.15}$$

Hierbei haben wir die Fixpunktgleichung als erfüllt gesehen, wenn wir ein $x^* \in \Omega$ gefunden haben, so dass für die Fixpunktgleichung (2.15) gilt $x^* = G(x^*)$. Unter dieser Beobachtung haben wir das *Newton-Verfahren* als iteratives Schema zur Bestimmung eines solchen Fixpunktes $x^* \in \Omega$ hergeleitet:

$$x_{k+1} \;=\; x_k - (F'(x_k))^{-1} F(x_k),, \qquad \text{für } F'(x_k) \text{ regulär.} \tag{2.16}$$

Hierfür benötigten wir einen geeigneten Startwert $x_0 \in \Omega$ in einer lokalen Umgebung $U \subset \Omega$ des Fixpunktes $x^* \in U$.

Der folgende Satz formuliert Bedingungen für die lokale Konvergenz des Newton-Verfahrens.

THEOREM 2.24: Lokale Konvergenz des Newton Verfahrens.

Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ in einer Umgebung von $\overline{x} \in \mathbb{R}^n$ stetig differenzierbar und $\overline{x}$ sei eine Nullstelle von F mit $F(\overline{x}) = 0$. Sei außerdem die Jacobi-Matrix F' lokal Lipschitz-stetig und $F'(\overline{x})$ regulär in der Nullstelle.
Dann existiert eine lokale Umgebung $B_R(\overline{x})$, so dass das Newton-Verfahren für jeden Startwert $x_0 \in B_R(\overline{x})$ gegen die Nullstelle $\overline{x}$ konvergiert, d.h. es gilt $\lim_{x\to\infty} x_k = \overline{x}$.

Beweis. Siehe [Numerik 1, Satz 4.9]. □

Anstatt nun eine Nullstelle der Funktion F zu suchen, wollen wir das Newton-Verfahren nutzen, um eine Nullstelle des Gradienten ∇F (d.h einen stationären Punkt von F) zu approximieren und damit die notwendigen Optimalitätsbedingungen in Theorem 2.8 zu erfüllen. Im Folgenden sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und $F\colon \Omega \to \mathbb{R}$ eine differenzierbare, reellwertige Funktion. Wir betrachten wieder die Taylorapproximation der Funktion F in eine Abstiegsrichtung $x_k + p \in \Omega$ des allgemeinen Iterationsschemas (2.3), aber berücksichtigen diesmal auch Terme von zweiter Ordnung:

$$F(x_k + p) \;\approx\; F(x_k) + \langle p, \nabla F(x_k)\rangle + \frac{1}{2}\langle p, \nabla^2 F(x_k) p\rangle \;=:\; m_k(p). \tag{2.17}$$

Unter gewissen Bedingungen an die Hessematrix $\nabla^2 F(x_k)$, lässt sich ein eindeutiges Minimum der Modellfunktion $m_k(p)$ in (2.17) bestimmen, wie folgendes Theorem besagt.

THEOREM 2.25: Newton-Abstiegsrichtung.

Sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet. Sei außerdem $F\colon \Omega \to \mathbb{R}$ eine in einer lokalen Umgebung eines Punktes $x_k \in \Omega$ zweimal stetig differenzierbare Zielfunktion, deren Hessematrix $\nabla^2 F(x_k)$ im Punkt x_k positiv definit ist.
Dann ist der **Newton-Abstiegsrichtung** benannte Vektor $p_k^N \in \mathbb{R}^n$ mit

$$p_k^N \;=\; -(\nabla^2 F(x_k))^{-1} \nabla F(x_k) \tag{2.18}$$

das eindeutige Minimum der Modellfunktion $m_k(p)$ in (2.17).

Beweis. In den Übungsaufgaben zu zeigen. □

Mit der Newton-Abstiegsrichtung in (2.18) lässt sich ein iteratives Abstiegsverfahren für einen initialen Punkt $x_0 \in \Omega$, welcher geeignet in der Nähe des stationären Punktes $x^* \in \Omega$ gewählt wird, wie folgt konstruieren:

$$x_{k+1} \;=\; x_k + p_k^N \;=\; x_k - (\nabla^2 F(x_k))^{-1} \nabla F(x_k). \tag{2.19}$$

Damit das Newton-Abstiegsverfahren in (2.19) überhaupt sinnvoll ist, müssen wir fordern, dass die Hessematrix in jedem Punkt $x_k \in \Omega$ der Iterationsfolge *regulär* und somit invertierbar ist. Um sicher zu gehen, dass es sich tatsächlich um eine Abstiegsrichtung

handelt müssen wir fordern, dass die Hessematrix $\nabla^2 F(x_k)$ nicht nur invertierbar für alle $x_k \in \Omega$ der Iterationsfolge ist, sondern auch *positiv definit* in jedem Punkt x_k ist. Denn dann ergibt eine Taylorapproximation zweiter Ordnung die folgende Abschätzung:

$$\begin{aligned} F(x_{k+1}) &= F(x_k + p_k^N) \approx F(x_k) + \langle p_k^N, \nabla F(x_k)\rangle + \frac{1}{2}\langle p_k^N, \nabla^2 F(x_k) p_k^N\rangle \\ &= F(x_k) - \langle p_k^N, \nabla^2 F(x_k) p_k^N\rangle + \frac{1}{2}\langle p_k^N, \nabla^2 F(x_k) p_k^N\rangle \\ &= F(x_k) - \frac{1}{2}\underbrace{\langle p_k^N, \nabla^2 F(x_k) p_k^N\rangle}_{>0}. \end{aligned}$$

Wir sehen also, dass wir einen echten Abstieg der Funktionswerte erhalten, wenn die Hessematrix $\nabla^2 F(x_k)$ positiv definit ist für alle $x_k \in \Omega$ der Iterationsfolge. Sollte die Hessematrix nicht positiv definit in einem Punkt x_k der Iterationsfolge sein, so muss zumindest eine Abnahme der Funktionswerte vorliegen, d.h., es muss für die Newton-Abstiegsrichtung gelten:

$$\langle (\nabla^2 F(x_k))^{-1} \nabla F(x_k), \nabla F(x_k)\rangle > 0.$$

Sollte dies nicht der Fall sein, so existieren Methoden um dennoch einen Abstieg zu erzwingen, siehe zum Beispiel [NW99, Kapitel 6]. Auf diese werden wir jedoch im weiteren Verlauf der Vorlesung nicht näher eingehen.

BEMERKUNG 2.26 (Schrittweite und Konvergenz). Das Newton-Abstiegsverfahren in (2.19) ist ein Abstiegsverfahren der Art (2.3) dessen Schrittweite $\alpha_k > 0$ implizit durch die lokale Krümmung und die Ableitung der Funktion F bestimmt ist. In diesem Fall können wir $\alpha_k \equiv 1$ für alle $k \in \mathbb{N}$ setzen. Das Newton-Abstiegsverfahren konvergiert in der Regel *quadratisch* gegen einen stationären Punkt $x^* \in \Omega$ mit $\nabla F(x^*) = 0$, d.h. man erreicht sehr schnell eine hohe Genauigkeit bei der Approximation von x^*. △

2.2.5 Quasi-Newton Verfahren

Im Abschnitt 2.2.4 haben wir das Newton Verfahren zur iterativen Approximation eines stationären Punktes $x^* \in \Omega$ einer Funktion F mit $\nabla F(x^*) = 0$ hergeleitet. Hierbei haben wir im Gegensatz zu den vorherigen numerischen Verfahren auch Ableitungen höherer Ordnung hinzugezogen. Dies führt in der Regel zu einem verbesserten Konvergenzverhalten im Vergleich zu den Verfahren, die nur die lokale Ableitung ∇F der Zielfunktion F verwenden.

Dennoch ist das Newton Verfahren aus numerischer Sicht noch nicht ideal, da es einige Probleme mit sich bringt. Zuerst mussten wir fordern, dass die Hessematrix $\nabla^2 F(x_k)$ in jedem Punkt des Iterationsverfahrens positiv definit ist, da ansonsten kein Abstieg der Funktionswerte garantiert werden kann. Zweitens muss für die Berechnung der Newton-Richtung in (2.18) zuerst die Hessematrix bestimmt und anschließend invertiert werden.

Dies ist aus Effizienzgründen unerwünscht, da die Inversion einer $n \times n$-Matrix bekanntlich in $\mathcal{O}(n^3)$ Rechenoperationen liegt (siehe [Numerik 1, Kapitel 1.1]. Da die Bestimmung und die Inversion der Hessematrix in jedem Iterationsschritt passieren müssen, ist das Newton Verfahren nur eingeschränkt empfehlenswert für die numerische Optimierung.

Eine naheliegende Idee ist es nun die echte Hessematrix in jedem Iterationsschritt durch eine geeignete Matrix zu approximieren, so dass der numerische Aufwand geringer wird, d.h., wir suchen nach einer Matrix $B_k \in \mathbb{R}^{n\times n}$

$$B_k \approx \nabla^2 F(x_k). \tag{2.20}$$

Damit können wir die Modellfunktion $m_k(p)$ in (2.17) schreiben als:

$$m_k(p) \;=\; F(x_k) + \langle p, \nabla F(x_k)\rangle + \frac{1}{2}\langle p, B_k p\rangle,$$

das heißt, wir approximieren die Zielfunktion F im k-ten Iterationsschritt entlang der Richtung $p \in \mathbb{R}^n$ lokal durch eine quadratische Funktion. Für sehr kleine Schrittweiten können wir davon ausgehen, dass der Fehler dieser Approximation gering ist, da wir davon ausgehen, dass F stetig differenzierbar in einer lokalen Umgebung $U \subset \Omega$ des stationären Punktes $x^* \in \Omega$ ist und für $p = 0$ die Approximation exakt ist, da gilt

$$m_k(0) \;=\; F(x_k).$$

Wenn wir fordern, dass B_k in (2.20) eine positiv definite Matrix ist, so lässt sich ein Abstiegsschritt des Iterationsverfahrens (2.3) analog zur Herleitung des Newton Abstiegsverfahrens in Abschnitt 2.2.4 angeben als:

$$x_{k+1} \;=\; x_k + \alpha_k p_k, \qquad p_k \;=\; -B_k^{-1}\nabla F(x_k). \tag{2.21}$$

Die sogenannten *Quasi-Newton Verfahren* verfolgen diesen Ansatz.

BEMERKUNG 2.27 (Konvergenzgeschwindigkeit Quasi-Newton Verfahren). Durch die Approximation der echten Hessematrix verlieren Quasi-Newton Verfahren an Genauigkeit, wodurch ihre Konvergenzgeschwindigkeit *superlinear* anstatt *quadratisch* ist. Dafür gewinnen sie zusätzliche Geschwindigkeit durch die Vermeidung der Bestimmung und Inversion von $\nabla^2 F(x_k)$. Der Vorteil der Quasi-Newton Methoden ist es, dass man nur den Gradienten ∇F für einen Schritt des numerischen Optimierungsverfahrens benötigt und keine expliziten Informationen über die zweiten Ableitungen. Dadurch werden sie in bestimmten Problemen sogar effizienter bei der Approximation eines stationären Punktes als das Newton Abstiegsverfahren in Abschnitt 2.2.4. △

Sekantengleichung und Krümmungsbedingung

Die entscheidende Frage bei der Konstruktion eines Quasi-Newton Abstiegsverfahrens der Form (2.21) ist es, wie die positiv definite Matrix B_k in jedem Schritt möglichst

effizient bestimmt werden kann. Anstatt die Näherung B_k der Hessematrix $\nabla^2 F(x_k)$ in jedem Schritt von Grund auf neu zu berechnen, wäre es wünschenswert ein initiales B_0 zu bestimmen, das in jedem Schritt des Iterationsverfahrens nur aktualisiert werden muss. Hierbei ist es möglich die durch den Iterationsschritt erhaltenen Informationen über den Gradienten ∇F zu Hilfe zu nehmen.

Wir nehmen an, wir haben bereits einen Abstiegsschritt durchgeführt und so einen neuen Punkt $x_{k+1} = x_k + \alpha_k p$ erhalten. Unsere quadratische Approximation in diesem neuen Punkt für eine neue Richtung $p \in \mathbb{R}^n$ sieht dementsprechend wie folgt aus:

$$m_{k+1}(p) \;=\; F(x_{k+1}) + \langle \nabla F(x_{k+1}), p\rangle + \frac{1}{2}\langle p, B_{k+1}p\rangle. \tag{2.22}$$

Es ist leicht einzusehen, dass die Modellfunktion m_{k+1} im Punkt $x_{k+1} \in \mathbb{R}^n$ zentriert ist und für $p = 0$ mit dem Funktionswert der Zielfunktion im Punkt x_{k+1} übereinstimmt, d.h., es gilt $m_{k+1}(0) = F(x_{k+1})$.

Da wir uns bei der Wahl der positiv definiten Matrix B_{k+1} noch nicht festgelegt haben, wird durch (2.22) eine Funktionenschar beschrieben. Eine Forderung, die man nun die Modellfunktion m_{k+1} stellen kann, um eine sinnvolle Matrix B_{k+1} zu bestimmen, ist, dass ihre Ableitung ∇m_{k+1} mit der Ableitung der Zielfunktion F in den letzten beiden Punkten x_k und x_{k+1} übereinstimmt. Dies bedeutet, dass man die Matrix B_{k+1} versucht so zu bestimmen, dass die Modellfunktion m_{k+1} die Krümmung der Zielfunktion F gut approximiert. Da bereits gilt

$$\nabla m_{k+1}(0) \;=\; \nabla F(x_{k+1}),$$

ist eine der beiden Forderungen automatisch erfüllt. Für die zweite Forderung können wir nutzen, dass $x_k = x_{k+1} - \alpha_k p_k$ gilt und wir erhalten somit:

$$\nabla F(x_k) \;\overset{!}{=}\; \nabla m_{k+1}(-\alpha_k p_k) \;=\; \nabla F(x_{k+1}) - B_{k+1}\alpha_k p_k. \tag{2.23}$$

Durch Umstellen von (2.23) erhalten wir die Bedingung

$$\nabla F(x_{k+1}) - \nabla F(x_k) \;\overset{!}{=}\; B_{k+1}\alpha_k p_k \;=\; B_{k+1}(x_{k+1} - x_k).$$

Eine vernünftige Wahl der Matrix B_{k+1} in (2.20) sollte diese Eigenschaft, auch bekannt als **Sekantengleichung**, versuchen zu imitieren. Im eindimensionalen Fall mit $F\colon \Omega \subset \mathbb{R} \to \mathbb{R}$ bedeutet die Sekantengleichung nichts anderes, als dass der Faktor B_{k+1} eine Approximation des zweiten Ableitung von F im Sinne eines Differenzenquotienten ist, d.h., im Fall $n = 1$ soll gelten:

$$B_{k+1} \;\overset{!}{=}\; \frac{F'(x_{k+1}) - F'(x_k)}{x_{k+1} - x_k}.$$

Für unser allgemeines Quasi-Newton Verfahren in (2.21) suchen wir also einen Weg die bereits bekannte Approximation der Hessematrix $B_k \approx \nabla^2 F(x_k)$ zu einer Matrix B_{k+1} zu aktualisieren, so dass der folgende Zusammenhang für den nächsten Punkt $x_{k+1} \in \Omega$ erfüllt wird:

$$B_{k+1}s_k \;=\; y_k, \tag{2.24}$$

wobei

$$s_k \;=\; x_{k+1} - x_k, \qquad y_k = \nabla F(x_{k+1}) - \nabla F(x_k).$$

Es wird klar, dass diese Forderung alleine nicht genügt für die Konstruktion eines Abstiegsverfahrens, da die Sekantengleichung in (2.24) für $n > 1$ unterbestimmt ist, d.h., dass es mehr unbekannte Einträge der Matrix $B_{k+1} \in \mathbb{R}^{n\times n}$ gibt als durch die n Gleichungen festgelegt werden. Daher versuchen wir im Folgenden weitere Forderungen an die Matrix B_{k+1} zu stellen.

Um die positive Definitheit der Matrix B_{k+1} in Schrittrichtung $x_{k+1} - x_k = \alpha p_k \in \mathbb{R}^n$ zu gewährleisten müssen wir fordern, dass die Vektoren y_k und s_k die sogenannte **Krümmungsbedingung** erfüllen:

$$\langle s_k, y_k\rangle \;>\; 0. \tag{2.25}$$

Dies ist eine hinreichende Bedingung für die positive Definitheit von B_{k+1} bezüglich der Richtung $\alpha_k p_k$, da wir einfach die Sekantengleichung (2.24) von links mit dem Vektor s_k^T multiplizieren können und so erhalten wir mit der Forderung (2.25) schon:

$$\langle s_k, B_{k+1}s_k\rangle \;=\; \langle s_k, y_k\rangle \;>\; 0.$$

BEMERKUNG 2.28 (Krümmungsbedingung und Konvexität). Falls die Zielunktion F strikt konvex ist, so ist die Krümmungsbedingung (2.25) für alle Punktepaare $x_k, x_{k+1} \in \Omega$ erfüllt und die Matrix B_{k+1} wird damit positiv definit. Für nichtkonvexe Funktionen hingegen muss man die Krümmungsbedingung explizit forcieren, um ein Abstiegsverfahren zu erhalten. △

Falls die Krümmungsbedingung (2.25) erfüllt ist, so existiert mindestens eine Lösung B_{k+1} der Sekantengleichung (2.24). Man sieht ein, dass es in der Tat sogar unendlich viele Lösungen B_{k+1} gibt, da eine symmetrische $n \times n$ Matrix $n(n+1)/2$ Freiheitsgrade besitzt und die Sekantengleichung (2.24) nur n Bedingungen an B_{k+1} stellt. Zusätzlich erhält man n Bedingungen an B_{k+1} durch die Forderung von positiver Definitheit, da alle n Hauptminoren von B_{k+1} positiv sein müssen. Dies reicht jedoch nicht für die eindeutige Bestimmung der Matrix B_{k+1}. Hierfür müssen wir zusätzlich fordern, dass die Matrix B_{k+1} diejenige Matrix unter allen möglichen Lösungen ist, die der vorherigen Matrix B_k am nächsten bezüglich eines geeigneten Maßes ist. Das heißt wir suchen eine Lösung des folgenden Optimierungsproblems:

$$\begin{aligned} &\min_{B\in\mathbb{R}^{n\times n}} ||B - B_k||, \qquad \text{unter den Nebenbedingungen:} \\ B \;&=\; B^T, \qquad Bs_k \;=\; y_k, \qquad \langle p, Bp\rangle > 0, \forall p \in \mathbb{R}^n/\{0\}, \end{aligned} \tag{2.26}$$

wobei s_k und y_k definiert sind wie in der Sekantengleichung (2.24). Man beachte, dass man eine unterschiedliche Lösung B_{k+1} des Optimierungsproblems (2.26) in Abhängigkeit der gewählten Matrixnorm erhält und somit auch ein unterschiedliches Quasi-Newton Verfahren herleiten kann.

Das Davidon-Fletcher-Powell Verfahren

Im ursprünglich im Jahr 1959 von Davidon vorgeschlagenen Verfahren [Dav59] (das im Übrigen bei der Erstbegutachtung abgelehnt wurde) wählt man für die Norm im Optimierungsproblem (2.26) eine gewichtete Frobeniusnorm der Form

$$||A||_W \;:=\; ||W^{\frac{1}{2}} A W^{-\frac{1}{2}}||_F.$$

Die Gewichtungsmatrix W dient dazu, dass das implizierte Quasi-Newton Verfahren zur Approximation eines stationären Punktes $x^* \in \Omega$ skalierungs-invariant wird. Hierzu wählt man eine beliebige Matrix für die die Relation $W y_k = s_k$ gilt, d.h., eine Matrix W, die sich wie die Inverse der Matrix B in (2.26) verhält. Ein konkretes Beispiel für solch eine Gewichtungsmatrix wäre $W := G_k^{-1}$, wobei G_k die *durchschnittliche Hessematrix* von F entlang des letzten Abstiegsschritts von $x_k \to x_{k+1}$ ist mit

$$G_k \;:=\; \int_0^1 \nabla^2 F(x_k + t\alpha_k p_k)\, \mathrm{d}t.$$

Mit der konkreten Wahl dieser Gewichtungsmatrix $W = G_k^{-1}$ wird die gewichtete Frobeniusnorm dimensionslos und man erhält eine eindeutige Lösung des Optimierungsproblems (2.26) wie folgt:

$$B_{k+1} \;=\; (I - \gamma_k y_k s_k^T) B_k (I - \gamma_k s_k y_k^T) + \gamma_k y_k y_k^T, \qquad \text{mit } \gamma_k \;:=\; \frac{1}{\langle y_k, s_k \rangle}. \tag{2.27}$$

Die Gleichung (2.27) wird auch **DFP-Schritt** genannt, da sie zuerst von Davidon vorgeschlagen und später von Fletcher und Powell untersucht und verbreitet wurde.

Obwohl wir die explizite Berechnung der Hessematrix $\nabla^2 F(x_k)$ vermieden haben und die Aktualisierung der Matrix B_k zu B_{k+1} lediglich auf den Gradienteninformationen von F basiert ist der numerische Aufwand bei direkter Verwendung von B_{k+1} in (2.27) noch zu hoch. Das liegt daran, dass wir für einen Schritt des Quasi-Newton Verfahrens in (2.21) die Inverse der Matrix B_k benötigen und die Inversion einen numerischen Aufwand von $\mathcal{O}(n^3)$ besitzt. Glücklicherweise gibt es einen Trick, wie wir die Inverse von B_k in jedem Schritt des Iterationsverfahrens numerisch günstig erhalten können. Sei $H_k := B_k^{-1}$, dann können wir die sogenannte Sherman-Morrison-Woodbury Formel (siehe [SMW Formel]) auf Gleichung (2.27) anwenden um die neue Inverse H_{k+1} durch eine Aktualisierung der Matrix H_k zu berechnen:

$$H_{k+1} \;=\; H_k - \frac{H_k y_k y_k^T H_k}{\langle y_k, H_k y_k \rangle} + \frac{s_k s_k^T}{\langle y_k, s_k \rangle}. \tag{2.28}$$

Wie man einssieht liegt der numerische Rechenaufwand für das Update von H_{k+1} in (2.28) in $\mathcal{O}(n^2)$. Es fällt außerdem auf, dass H_k nur durch die Addition zweier Matrizen mit Rang 1 verändert wird, also insgesamt eine Änderung von höchstes Rang 2 erfährt. Das passt gut zu der Forderung, dass wir erwarten, dass sich die Approximation der Hessematrix $\nabla^2 F$ in einer lokalen Umgebung nur wenig ändert.

Das Broyden–Fletcher–Goldfarb–Shanno Verfahren

Das Davidon–Fletcher–Powell Verfahren wurde trotz seiner Effektivität bald schon durch ein Verfahren abgelöst, das noch besser war und bis heute zu den effizientesten Quasi-Newton Verfahren gehört: das Broyden–Fletcher–Goldfard–Shanno (BFGS) Verfahren in [BFGS70]. Die Idee des BFGS Verfahrens leitet sich unmittelbar aus der Idee des DFP Verfahren ab. Anstatt das Optimierungsproblem (2.26) mit bestimmten Bedingungen an die Approximation B_{k+1} der Hessematrix $\nabla^2 F(x_k)$ zu stellen, versucht man direkt die Inverse der Hessematrix $(\nabla^2 F(x_k))^{-1}$ geeignet zu approximieren. Hierfür nehmen wir an, dass wir eine Matrix H_{k+1} als geringfügige Aktualisierung einer bereits vorher bestimmten Matrix H_k suchen, die gleichzeitig symmetrisch und positiv definit ist und zusätzlich die Sekantenbedingung in umgeschriebener Form erfüllt:

$$H_{k+1} y_k \;=\; s_k.$$

Hierzu formuliert man ein analoges Optimierungsproblem zu (2.26) von der Form:

$$\begin{aligned} &\min_H ||H - H_k||, \qquad \text{unter den Nebenbedingungen:} \\ H \;=\; H^T, \qquad &H y_k \;=\; s_k, \qquad \langle p, Hp\rangle > 0, \forall p \in \mathbb{R}^n/\{0\}. \end{aligned} \tag{2.29}$$

Unter der Verwendung der gewichteten Frobeniusnorm und einer beliebigen Gewichtsfunktion, die die Sekantengleichung $W s_k = y_k$ erfüllt, erhält man wiederum die eindeutige Lösung des Minimierungsproblems (2.29) als:

$$H_{k+1} \;=\; (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T, \qquad \text{mit } \rho_k := \frac{1}{\langle y_k, s_k\rangle}. \tag{2.30}$$

Das Update der Matrix H_k in (2.30) kann numerisch in $\mathcal{O}(n^2)$ durchgeführt werden, was man schnell einsieht, wenn man das Produkt ausschreibt:

$$H_{k+1} \;=\; H_k - H_k \rho_k y_k s_k^T - \rho_k s_k y_k^T H_k + \rho_k s_k y_k^T H_k \rho_k y_k s_k^T + \rho_k s_k s_k^T.$$

In dieser Schreibweise sieht man gut, dass man lediglich Skalarprodukte in $\mathcal{O}(n)$, Matrix-Vektor Multiplikationen in $\mathcal{O}(n^2)$ und dyadische Produkte in $\mathcal{O}(n^2)$ berechnen muss. Im Gegensatz hierzu würde eine naive Implementierung des BFGS-Updates in (2.30) zu einem numerischen Aufwand von $\mathcal{O}(n^3)$ führen.

Abschließend bleibt die Frage was eine gute Initialisierung der Matrix H_0 ist. Idealerweise hat man bereits Informationen über die Inverse der Hessematrix $(\nabla^2 F(x_0))^{-1}$ im Initialisierungspunkt $x_0 \in \Omega$, zum Beispiel durch eine numerische Approximation mittels finiter Differenzen (später in der Vorlesung!). Andererseits erwarten wir, dass die Aktualisierung von H_k im k-ten Schritt des Iterationsverfahrens (2.30) zu H_{k+1} die aktuellen Informationen über den Verlauf der Gradienten $\nabla F(x_k)$ und $\nabla F(x_{k+1})$ berücksichtigt. Darum ist eine häufige Wahl von H_0 die Initialisierung als Einheitsmatrix I_n oder ein Vielfaches der Einheitsmatrix, wobei die Vorfaktoren der Diagonaleinträge entsprechend der Skalierung der Variablen gewählt werden.

2.3 Verfahren der konjugierten Gradienten

Im Folgenden wollen wir uns mit einem besonders eleganten Verfahren der Optimierung beschäftigen: dem Verfahren der konjugierten Gradienten. Ursprünglich wurde das Verfahren von Hestenes und Stiefel in [HS52] im Jahr 1952 vorgeschlagen. Obwohl das Verfahren im Allgemeinen für die nichtlineare Optimierung eingesetzt werden kann, wird es insbesondere zur Lösung von großen linearen Gleichungssystemen $Ax = b$ mit symmetrischer, dünn besetzter, positiv definiter Matrix $A \in \mathbb{R}^{n\times n}$ eingesetzt. Solche Gleichungssysteme treten zum Beispiel bei der numerischen Modellierung und Lösung partieller Differentialgleichungen auf. Das Verfahren lässt sich in diesem Fall besonders anschaulich motivieren und herleiten. Darum wollen wir uns im Folgenden zunächst auf das Lösen von großen linearen Gleichungssystemen $Ax = b$ konzentrieren. Wir folgen bei der Herleitung des Verfahren der konjugierten Gradienten der didaktisch sehr gelungenen Arbeit von Jonathan Shewchuk in [She94]. Für eine ansprechende, interaktive Visualisierung des Verfahren der konjugierten Gradienten empfehlen wir den Mathematik Blog von Philipp Wacker [Wacker].

2.3.1 Problemstellung

Sei im Folgenden also $A \in \mathbb{R}^{n\times n}$ eine sehr große Matrix und $b \in \mathbb{R}^n$ ein reeller Vektor. Wir suchen einen unbekannten Vektor $x \in \mathbb{R}^n$, der das lineare Gleichungssystem

$$Ax \;=\; b \tag{2.31}$$

löst. Wir suchen also nach denjenigen Koeffizienten, mit denen sich der Vektor b als Linearkombination aus Spaltenvektoren der Matrix A darstellen lässt. Diese Koeffizienten entsprechen den Einträgen des unbekannten Vektors x.

Aus der Vorlesung „Einführung in die Numerik“ in [Numerik 1, Kapitel 2] ist bekannt, dass es genau dann eine eindeutige Lösung $x \in \mathbb{R}^n$ für die Gleichung (2.31) gibt, falls die Determinante $\det(A) \neq 0$ ist. Eine hinreichende Bedingung für die Eindeutigkeit des Lösungsvektors $x \in \mathbb{R}^n$ ist es also zu fordern, dass die Matrix A symmetrisch und positiv definit ist.

Wir gehen aus diesem Grund im Folgenden immer davon aus, dass $A \in \mathbb{R}^{n\times n}$ eine *symmetrische* und *positiv definite* Matrix ist. In diesem Fall ist das Bestimmen einer Lösung von (2.31) ein gut-gestelltes Problem und die Lösung lässt sich direkt angeben als:

$$x \;=\; A^{-1}b.$$

Wie wir jedoch ebenfalls aus [Numerik 1, Kapitel 2] wissen ist die Inversion einer Matrix $A \in \mathbb{R}^{n\times n}$ numerisch sehr aufwänding und selbst unter Ausnutzung der Symmetrie lässt sich höchstens ein Verfahren mit Rechenaufwand $\mathcal{O}(\frac{1}{6}n^3)$ angeben. Sollte die Dimension des Problems jedoch sehr groß sein (d.h. wir nehmen $n >> 1$ an), so ist eine direkte Lösung von (2.31) mittels Inversion nicht durchführbar. Glücklicherweise liefert uns das Verfahren der konjugierten Gradienten (neben anderen iterativen Lösungsalgorithmen) eine Möglichkeit das lineare Gleichungssystem numerisch zu lösen. Man beachte, dass

wir explizit darauf verzichten zu fordern, dass die Matrix A *dünnbesetzt* ist. In diesem Fall könnten wir nämlich ebenfalls die numerischen Iterationsverfahren aus [Numerik 1, Kapitel 4.6] anwenden.

Wir betrachten zunächst das folgende konvexe, quadratische Optimierungsproblem der Form

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle + c, \tag{2.32}$$

wobei A und b wie im Fall des linearen Gleichungssystems in (2.31) gewählt sind und $c \in \mathbb{R}$ eine beliebige, reelle Konstante ist. Der folgende Satz liefert uns eine hilfreiche Aussage zur Lösung des ursprünglichen Problems.

THEOREM 2.29: Äquivalenzaussage für lineares Gleichungssystem.

Das konvexe, quadratische Minimierungsproblem in (2.32) ist äquivalent zum ursprünglichen linearen Gleichungssystem in (2.31), d.h., jede Lösung von (2.32) ist schon Lösung von (2.31) und anders herum.

Beweis. Für die erste Richtung des Beweises nehmen wir an, dass $x^* \in \mathbb{R}^n$ eine Lösung des linearen Gleichungssystems $Ax = b$ sei. Wir betrachten die hinreichenden Optimalitätsbedingungen zweiter Ordnung aus Theorem 2.13 für das Minimierungsproblem (2.32). Hierzu suchen wir zunächst die stationären Punkte der Funktion $F\colon \mathbb{R}^n \to \mathbb{R}$ mit

$$F(x) \;:=\; \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle + c.$$

Der Gradient von F lässt sich wegen der Symmetrie von A bestimmen als

$$\nabla F(x) \;=\; \frac{1}{2}(A + A^T)x - b \;=\; Ax - b \;\stackrel{!}{=}\; 0.$$

Alle stationären Punkte $x \in \mathbb{R}^n$ von F mit $\nabla F(x) = 0$ sind also gerade die Lösungen des linearen Gleichungssystems $Ax = b$. Damit ist x^* nach Vorraussetzung also einziger stationärer Punkt von F. Um zu zeigen, dass $x^* \in \mathbb{R}^n$ auch schon ein lokales Minimum von F ist müssen wir noch die Hessematrix von F betrachten, welche gegeben ist durch:

$$\nabla^2 F(x) \;=\; A.$$

Da A nach Vorraussetzung positiv definit ist, ist auch die Hessematrix $\nabla^2 F(x) = A$ positiv definit und somit sind die hinreichenden Kriterien für das Vorliegen eines lokalen Minimums von F im Punkt $x^* \in \mathbb{R}^n$ erfüllt.

Für die Rückrichtung des Beweises nehmen wir, dass $x^* \in \mathbb{R}^n$ ein lokales Minimum der Zielfunktion F ist. Damit folgt direkt, dass x^* ein stationärer Punkt von F ist und somit muss gelten:

$$\nabla F(x^*) \;=\; Ax^* - b \;=\; 0.$$

Das bedeutet aber schon, dass $x^* \in \mathbb{R}^n$ Lösung des linearen Gleichungssystems $Ax = b$ ist. □

Der Theorem 2.29 erlaubt es uns also ein quadratisches Optimierungsproblem der Form (2.32) numerisch zu lösen anstatt einen unbekannten Lösungsvektor für ursprüngliche lineare Gleichungssystem (2.31) zu finden.

Wir interessieren uns nun also für ein iteratives Verfahren, welches eine Folge von Punkten $x_0, x_1, \ldots \in \mathbb{R}^n$ konstruiert, die gegen ein Minimum von (2.32) und somit gegen die eindeutige Lösung des linearen Gleichungssystems (2.31) konvergiert. Hierfür benötigen wir noch zusätzliche Notation, um das angestrebte Verfahren vernünftig zu beschreiben.

DEFINITION 2.30: Fehler und Residuum.

Sei $x_{k+1} = G(x_k)$ ein Iterationsverfahren, dass gegen ein lokales Minimum $x^* \in \mathbb{R}^n$ der quadratischen Funktion F in (2.32) konvergiert, d.h., $x_k \to x^*$ für $k \to \infty$. Dann können wir die beiden folgenden Begriffe definieren:

(i) Wir bezeichnen den Vektor $e_k \in \mathbb{R}^n$ mit

$$e_k \;:=\; x_k - x^*$$

als den aktuellen **Fehler**, den man durch den aktuellen Punkt $x_k \in \mathbb{R}^n$ macht.

(ii) Wir bezeichnen den Vektor $r_k \in \mathbb{R}^n$ mit

$$r_k \;:=\; b - Ax_k$$

als das aktuelle **Residuum**, das man durch den aktuellen Punkt $x_k \in \mathbb{R}^n$ erhält.

BEMERKUNG 2.31 (Fehler und Residuum). In Bezug auf Definition 2.30 lassen sich folgende Aussagen festhalten:

(i) Der Fehler $e_k \in \mathbb{R}^n$ ist eher abstrakter Natur und dient zur besseren Analyse des Verfahrens der konjugierten Gradienten. Explizit werden wir diesen Vektor jedoch nie bestimmen können innerhalb des Iterationsverfahrens, da wir dann schon fertig wären mit einem einfach Update der Form $x^* = x_k - e_k$.

(ii) Wie wir bereits im Beweis von Theorem 2.29 gesehen haben, lässt sich das Residuum $r_k \in \mathbb{R}^n$ außerdem wie folgt umschreiben:

$$r_k \;=\; \underbrace{b - Ax_k}_{=-\nabla F(x_k)} \;=\; Ax^* - Ax_k \;=\; A(x^* - x_k) \;=\; -Ae_k.$$

Daher lässt sich das Residuum r_k auch als die Richtung des stärksten Abstiegs interpretieren und es ist klar, dass r_k immer orthogonal zu den Niveaulinien der Funktion F steht.

△

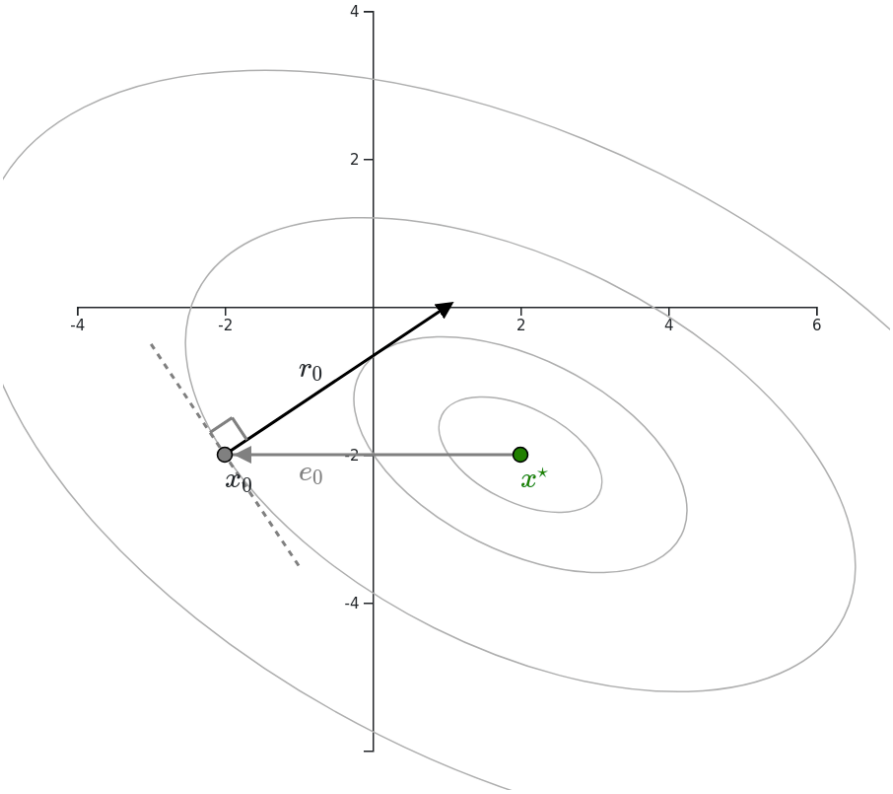

Abbildung 2.3: Visualisierung des Fehlers $e_0 \in \mathbb{R}^n$ und des Residuums $r_0 \in \mathbb{R}^n$ für einen Startpunkt $x_0 \in \mathbb{R}^n$.

Abb. 2.3 illustriert anschaulich die geometrische Bedeutung der beiden in Definition 2.30 eingeführten Vektoren. Wie man unschwer erkennt zeigen Fehler und Residuum im Allgemeinen nicht in die selbe Richtung. Das erklärt auch warum das Gradientenabstiegsverfahren in Abschnitt 2.2.1 selbst bei optimaler Schrittweite $\alpha_k > 0$ nicht in einem Schritt die gesuchte Lösung $x^* \in \mathbb{R}^n$ erreicht.

2.3.2 Motivation

Um das Vorgehen beim Verfahren der konjugierten Gradienten zu motivieren rufen wir uns noch einmal das Gradientenabstiegsverfahren aus Abschnitt 2.2.1 in Erinnerung. Nehmen wir an wir befinden uns im k-Schritt des Gradientenabstiegsverfahrens in Algorithmus 2.18 in einem Punkt $x_k \in \mathbb{R}^n$ und es sei eine Schrittweite $\alpha_k > 0$ gegeben. Dann erhalten wir den nächsten Punkt $x_{k+1} \in \mathbb{R}^n$ der Iterationsfolge durch folgendes Update:

$$x_{k+1} \;=\; x_k - \alpha_k \nabla F(x_k) \;=\; x_k + \alpha_k r_k,$$

wobei $r_k \in \mathbb{R}^n$ das aktuelle Residuum des Punktes x_k bezeichnet. Wir machen also in Richtung des steilsten Gradientenabstiegs einen Schritt der Länge $\alpha_k > 0$. Da die Abstiegsrichtung in jedem Schritt $x_k \to x_{k+1}$ orthogonal zu den Niveaulinien von F steht,

erhält man typischerweise einen Zickzack-Pfad durch das Gradientenabstiegsverfahren (vgl. Abb. 2.2).

Um dieses typische Verhalten besser zu verstehen können wir eine Vorüberlegung zur Schrittweitenwahl für das quadratische Optimierungsproblem in (2.32) machen. Hierzu gehen wir analog zur Bestimmung der optimalen Schrittrichtung in Abschnitt 2.2.1 vor, nur dass wir diesmal die Schrittrichtung $p_k := -\nabla F(x_k)$ festhalten und bezüglich der unbekannten Schrittweite optimieren.

Wir gehen davon aus, dass wir das lokale Minimum von F noch nicht erreicht haben, denn dann wäre $\alpha_k = 0$. Wir suchen also eine Schrittweite $\alpha > 0$, so dass der Funktionswert $F(x_{k+1})$ entlang der Linie $x_k - \alpha\nabla F(x_k)$ minimal wird. Da F eine quadratische Funktion ist, wissen wir, dass ein eindeutiges Minimum α_k entlang dieser Linie existieren muss. Wir nutzen also die notwendigen Optimalitätsbedingungen aus Theorem 2.8 für das totale Differential, um folgenden Zusammenhang herzustellen:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\alpha} F(x_{k+1}) &= \left\langle \nabla F(x_{k+1}), \frac{\mathrm{d}x_{k+1}}{\mathrm{d}\alpha} \right\rangle = \left\langle \nabla F(x_{k+1}), \frac{\mathrm{d}(x_k - \alpha\nabla F(x_k))}{\mathrm{d}\alpha} \right\rangle \\ &= \langle \nabla F(x_{k+1}), -\nabla F(x_k) \rangle = \langle \nabla F(x_{k+1}), p_k \rangle \overset{!}{=} 0. \end{aligned} \tag{2.33}$$

Das Ergebnis ist durchaus interessant. Die optimale Schrittweite $\alpha > 0$ muss so gewählt werden, dass der nächste Punkt $x_{k+1} \in \mathbb{R}^n$ der Iterationsfolge an der Stelle liegt an der unsere Abstiegsrichtung orthognal auf den Gradienten der Funktion $\nabla F(x_{k+1})$ trifft. Das bedeutet, dass die optimale Abfolge der Abstiegsrichtungen im quadratischen Fall eine Menge von 90 Grad Zickzack-Linien ergibt, was zu unseren Beobachtungen in Abb. 2.2 passt. Da jedoch der Punkt $x_{k+1} \in \mathbb{R}^n$ bislang noch unbekannt ist, können wir das optimale α_k nicht in dieser Form angeben. Das folgende Lemma bestimmt die optimale Schrittweite im Fall der quadratischen Optimierung in (2.32).

LEMMA 2.32: Optimale Schrittweite.

Sei $F\colon \mathbb{R}^n \to \mathbb{R}$ die quadratische Funktion aus (2.32). Wir betrachten das Gradientenabstiegsverfahren im k-ten Iterationsschritt mit einer unbekannten Schrittweite $\alpha_k > 0$, die jedoch so gewählt werden muss, dass

$$\langle \nabla F(x_{k+1}), \nabla F(x_k) \rangle \overset{!}{=} 0.$$

Sei außerdem $r_k = b - Ax_k$ das Residuum im aktuellen Iterationsschritt. Dann lässt sich die optimale Schrittweite α_k berechnen als:

$$\alpha_k = \frac{\langle r_k, r_k \rangle}{\langle r_k, Ar_k \rangle}. \tag{2.34}$$

Beweis. Wir erinnern uns daran, dass $r_k = -\nabla F(x_k) = b - Ax_k$ ist und somit können wir folgern:

$$\begin{aligned} 0 \overset{!}{=} \langle \nabla F(x_{k+1}), \nabla F(x_k) \rangle &= \langle r_{k+1}, r_k \rangle = \langle b - Ax_{k+1}, r_k \rangle \\ &= \langle b - A(x_k + \alpha_k r_k), r_k \rangle = \langle b - Ax_k, r_k \rangle - \alpha_k \langle Ar_k, r_k \rangle \\ &= \langle r_k, r_k \rangle - \alpha_k \langle r_k, Ar_k \rangle \end{aligned}$$

Da wir A als positiv definit angenommen haben, können wir die Gleichung umstellen und erhalten so die behauptete Berechnungsformel für α_k in (2.34). □

Obwohl wir die optimale Schrittweite α_k in (2.6) für das quadratische Optimierungsproblem (2.32) bestimmen konnten ist das Gradientenabstiegsverfahren weit davon entfernt optimal zu sein. Trotz optimaler Schrittweiten und optimaler Abstiegsrichtungen erhalten wir eine Folge von Richtungsvektoren, die immer wieder in die gleiche Richtung zeigen (siehe Abb. 2.4). Das ist numerisch gesehen äußerst ineffizient. Man könnte sich also fragen, warum man nicht einfach nur zwei orthogonale Schritte macht und die Schrittweiten als Summe der optimalen Schrittweiten der geraden bzw. ungeraden Iterationsschritte $k \in \mathbb{N}$ wählt. In der Tat würde man für $N \in \mathbb{N}$ Schritte des Gradientenabstiegsverfahren im selben Punkt $x_N \in \mathbb{R}^n$ mit nur zwei Iterationen landen, wie in Abb. 2.4 illustriert ist.

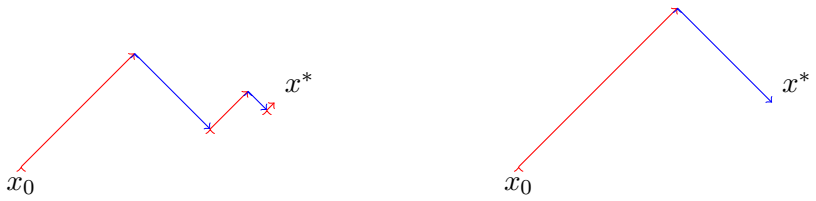

Abbildung 2.4: Vergleich des Gradientenabstiegsverfahrens mit optimaler Schrittweite $\alpha_k > 0$ aus Lemma 2.32 (links) mit einem idealen Abstiegsverfahren (rechts), bei dem alle orthogonalen Teilschritte zusammengefasst sind.

Leider können wir nicht alle Schrittweiten aufaddieren, da wir zur Berechnung der optimalen Schrittlänge $\alpha_k > 0$ bereits alle vorangegangen Schritte $k = 0, \dots, k-1$ kennen müssten. Außerdem würde ein großer, zusammengefasster Schritt in die erste der Richtungen eventuell dazu führen, dass man keinen Abstieg der Funktionswerte von F mehr realisiert, sondern einen Aufstieg. Diese Beobachtung ist in Abb. 2.5 illustriert.

Die Ideallösung wäre natürlich von einem Startpunkt $x_0 \in \mathbb{R}^n$ in nur einem Schritt zum lokalen Optimum $x^* \in \mathbb{R}^n$ zu gelangen. Da wir aber den Punkt x^* a-priori nicht kennen ist das eine unrealistische Forderung. Dennoch lässt sich zeigen, dass das Gradientenabstiegsverfahren mit der optimalen Schrittweite α_k in (2.34) im Fall des quadratischen Optimierungsproblems (2.32) in genau einem Schritt zum lokalen Minimum $x^* \in \mathbb{R}^n$ führt, wenn der Fehler $e_0 = x_0 - x^*$ ein Eigenvektor von A ist. Wir wissen nämlich aus Bemerkung 2.31, dass $r_k = -Ae_k$ gilt und somit erhalten wir für das Gradientenabstiegsverfahren im ersten Schritt:

$$x_1 = x_0 - \alpha_0 \nabla F(x_0) = x_0 + \alpha_0 r_0 = x_0 - \alpha_0 A e_0 = x_0 - \alpha_0 \lambda e_0.$$

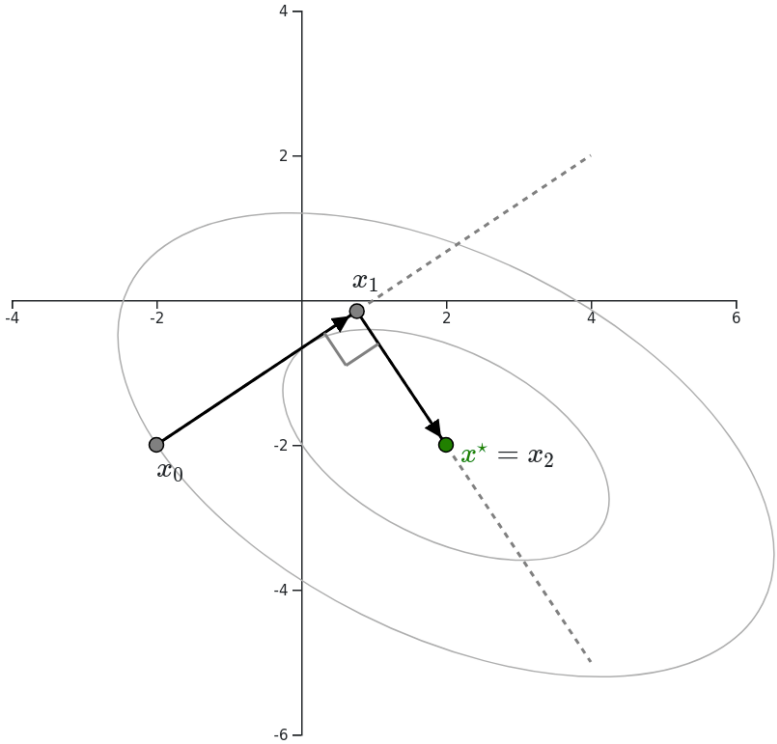

Abbildung 2.5: Illustration eines idealen Abstiegsverfahrens mit zwei orthogonalen Richtungen. Man beachte, dass die Schrittweite $\alpha_0 > 0$ so gewählt werden muss, dass man im ersten Schritt nicht in einem Punkt $x_1 \in \mathbb{R}^2$ mit minimalen Funktionswert $F(x_1)$ entlang der Richtung $x_0 - \alpha_0 \nabla F(x_0)$ endet.

Man müsste bei der Wahl des Startpunktes $x_0 \in \mathbb{R}^n$ jedoch viel Glück haben, um diese Forderung zu erfüllen. Darum wollen wir uns mit alternativen Ideen beschäftigen.

2.3.3 Orthogonale Abstiegsrichtungen

Wir wünschen uns einen Algorithmus, der ähnlich dem Gradientenabstiegsverfahren nur orthogonale Richtungen $\{d_0, \ldots, d_{n-1}\}$ mit $d_i \in \mathbb{R}^n$ für $0 \leq i \leq n-1$ verwendet, jedoch mit der Einschränkung, dass diese nur ein einziges Mal genutzt werden können. Ziel dieses Verfahrens soll es außerdem sein durch n Schritte in die jeweils n orthogonalen Richtungen $\{d_0, \ldots, d_{n-1}\}$ das lokale Minimum der Funktion zu erreichen. Damit hätten wir ein Itrerationsverfahren der Form

$$x_{k+1} \;=\; x_k + \alpha_k d_k, \quad \alpha_k > 0,\ k = 0, \ldots, n-1 \tag{2.35}$$

gewonnen. Wir könnten dies erzwingen indem wir im k-ten Schritt des Iterationsverfahrens fordern, dass ein Schritt in Richtung $d_k \in \mathbb{R}^n$ dazu führt, dass der Fehler $e_{k+1} \in \mathbb{R}^n$ keinerlei Komponenten dieser Richtung mehr enthält, d.h. wir fordern

$$\langle e_{k+1}, d_k \rangle \;\stackrel{!}{=}\; 0. \tag{2.36}$$

Da wir den zu erwartenden Fehler e_{k+1} in Bezug auf den aktuellen Punkt $x_k \in \mathbb{R}^n$ folgendermaßen umschreiben können:

$$e_{k+1} \;=\; x_{k+1} - x^* \;=\; x_k + \alpha_k d_k - x^* \;=\; e_k + \alpha_k d_k,$$

können wir die Forderung (2.36) umformulieren zu:

$$\langle e_k + \alpha_k d_k, d_k \rangle \;\overset{!}{=}\; 0.$$

Hieraus können wir die optimale Schrittweite $\alpha_k > 0$ in Richtung $d_k \in \mathbb{R}^n$ ableiten als

$$\alpha_k \;=\; -\frac{\langle e_k, d_k \rangle}{\langle d_k, d_k \rangle}. \tag{2.37}$$

Obwohl wir in (2.37) eine optimale Schrittweite α_k für das Verfahren mit orthogonalen Abstiegsrichtungen in (2.35) bestimmen konnten, hilft und diese nicht in der praktischen Anwendung des Verfahrens, da sie von dem unbekannten Fehlervektor $e_k \in \mathbb{R}^n$. Dieser hängt natürlich von der unbekannten Lösung $x^* \in \mathbb{R}^n$ ab und wenn wir diese kennen würden, so müssten wir kein iteratives Verfahren konstruieren. Selbst wenn man den Fehler e_k weiter rekursiv umschreibst, so würde man schlussendlich doch bei einer Abhängigkeit des initialen Fehlers e_0 landen. Wir müssen uns also vorerst von dieser Idee verabschieden und nach einer alternativen Möglichkeit suchen.

2.3.4 Konjugierte Abstiegsrichtungen

Obwohl unsere Idee von orthogonalen Abstiegsrichtungen in Abschnitt 2.3.3 nicht zum Ziel geführt hat, so war die Idee gar nicht schlecht. Das Hauptproblem an der ursprünglichen Idee liegt in der Forderung (2.36), nämlich dass der Fehlervektor e_{k+1} orthogonal zur aktuellen Richtung d_k stehen soll. Diese Forderung führt nämlich dazu, dass man orthogonale Vektoren erhält, die nicht an die Geometrie des quadratischen Minimierungsproblems (2.32) angepasst sind.

Wenn man sich die Niveaulinien der Funktion F genauer anschaut (siehe zum Beispiel Abbildung 2.5), so erkennt man, dass es Richtungen gibt entlang derer die Abstiegsrichtung zum lokalen Minimum $x^* \in \mathbb{R}^n$ steiler verläuft als entlang der anderen Richtungen. Die geometrischen Eigenschaften des Graphen von F sind maßgeblich durch die Gestalt der Matrix A, genauer gesagt durch deren Eigenvektoren bestimmt. Daher wollen wir diese Eigenschaften bei der Konstruktion eines iterativen Abstiegsverfahren berücksichtigen. Hierzu führen wir folgendes hilfreiche Konzept ein.

DEFINITION 2.33: Konjugierte Vektoren.

Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische, positiv definite Matrix und $u, v \in \mathbb{R}^n/\{0\}$ zwei Vektoren. Wir nennen v und w **konjugiert bezüglich A** oder auch **A-orthogonal** falls gilt

$$\langle v, Aw \rangle \;=\; \langle w, Av \rangle \;=\; 0.$$

Anstatt nun also die Orthogonalität unserer Richtungsvektoren $\{d_0, \ldots, d_{n-1}\}$ zu erzwingen wie in Abschnitt 2.3.3, fordern wir nun, dass diese Vektoren konjugiert bezüglich der Matrix A und damit besser an das Problem angepasst sind.

BEMERKUNG 2.34. Es ist leicht einzusehen, dass A-orthogonal und orthogonal die selbe Eigenschaft beschreiben, falls die Matrix A ein Vielfaches der Einheitsmatrix $I_n \in \mathbb{R}^{n\times n}$ ist. In diesem Fall ist das quadratische Optimierungsproblem (2.32) symmetrisch in alle Richtungen. △

Anschaulich lässt sich die Forderung nach A-Orthogonalität auch so deuten, dass wir ein Paar von Vektoren $v, w \in \mathbb{R}^n/\{0\}$ suchen, welche in einem Winkel so zueinander stehen, dass wenn man die Niveaulinien der Funktion F symmetrisch reskaliert, diese Vektoren anschließend orthogonal zueinander stehen. Diese Idee ist in Abbildung 2.6 dargestellt.

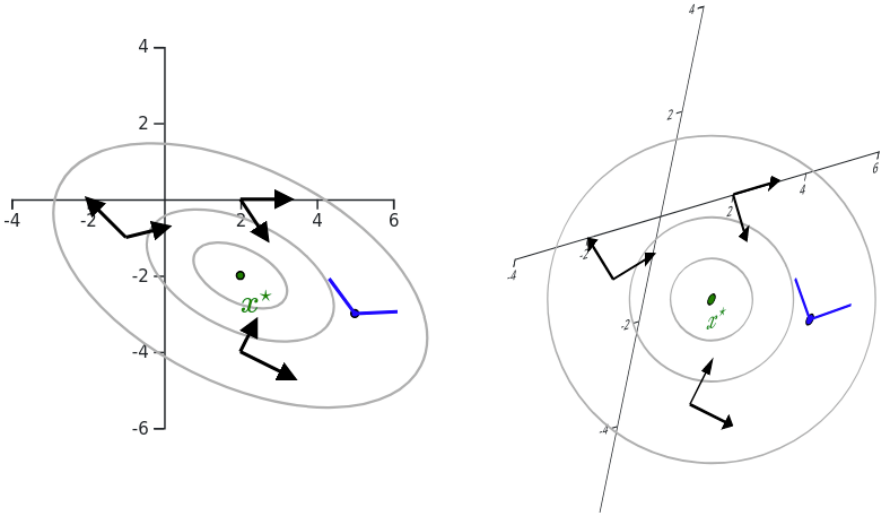

Abbildung 2.6: Illustration der Geometrie von konjugierten Vektoren im Referenzsystem $\mathbb{R}^2$ (links) und der selben Vektoren in einem symmetrisierten System bezüglich der Matrix A (rechts).

Anstatt also ein Abstiegsverfahren der Form (2.35) mit orthogonalen Vektoren zu verwenden, wollen wir ein Abstiegsverfahren mit A-orthogonalen Vektoren $\{d_0, \ldots, d_{n-1}\}$ konstruieren, d.h., wir verwenden das Iterationsschema

$$x_{k+1} \;=\; x_k + \alpha_k d_k, \quad \alpha_k > 0,\ k = 0, \ldots, n-1 \tag{2.38}$$

wobei für die Abstiegsrichtungen $d_k \in \mathbb{R}^n$ gelten soll:

$$\langle d_i, Ad_j \rangle = 0 \qquad \text{für alle } i \neq j.$$

Wir nehmen für den Moment an, dass wir einen numerischen Algorithmus kennen mit dem wir eine Menge von A-orthogonalen Vektoren $\{d_0, \ldots, d_{n-1}\}$ konstruieren können.

Wie man diese Menge konkret erhält werden wir uns im Anschluss erschließen. Sei also nun im Folgenden $\{d_0, \dots, d_{n-1}\}$ eine gegebene Menge von A-orthogonalen Vektoren. Dann stellen wir uns die Frage, wie die optimalen Schrittweiten $\alpha_k > 0$ in (2.38) gewählt werden müssen, um in n Schritten das lokale Minimum $x^* \in \mathbb{R}^n$ der Funktion F zu erhalten. Man beachte hierbei, dass wir nicht nur daran interessiert sind den Punkt x^* genügend gut zu approximieren, sondern wir fordern die eindeutige Lösung des linearen Gleichungssystems $Ax = b$ in n Schritten zu finden, d.h., wir nehmen explizit $x_n = x^*$ an.

Um das lokale Minimum wirklich in n Schritten zu erreichen müssen wir fordern, dass wir in jede Richtung d_k nur einmal gehen und der entstehende Fehler e_{k+1} A-orthogonal hierzu ist. Das entspricht der Forderung, dass man im entzerrten Problem auf der rechten Seite von Abb. 2.6 nur orthogonale Richtungen verwendet. Wir wollen also folgende Eigenschaft erzwingen:

$$\langle Ae_{k+1}, d_k \rangle \;=\; 0. \tag{2.39}$$

Analog zur Idee der orthogonalen Richtungen in Abschnitt 2.3.3 können wir den Fehler e_{k+1} in (2.39) wieder entwickeln, um die optimale Schrittweitenlänge $\alpha_k > 0$ zu bestimmen

$$\begin{aligned} 0 \;\overset{!}{=}\; \langle d_k, Ae_{k+1} \rangle \;&=\; \langle d_k, A(x_{k+1} - x^*) \rangle \;=\; \langle d_k, A(x_k + \alpha_k d_k - x^*) \rangle \\ &=\; \langle d_k, A(e_k + \alpha_k d_k) \rangle \;=\; \langle d_k, -r_k + \alpha_k A d_k \rangle \;=\; \alpha_k \langle d_k, Ad_k \rangle - \langle d_k, r_k \rangle. \end{aligned}$$

Da wir A als positiv definit vorausgesetzt haben, können wir die folgende Gleichung umstellen zu

$$\alpha_k \;=\; \frac{\langle d_k, r_k \rangle}{\langle d_k, Ad_k \rangle}. \tag{2.40}$$

Im Gegensatz zur Idee der orthogonalen Richtungen in (2.37) lässt sich der Ausdruck in (2.40) explizit berechnen und hängt nicht von dem unbekannten lokalen Minimum $x^* \in \mathbb{R}^n$ ab. Zusammenfassend heißt das, dass wir aus der Bedingung, dass die Abstiegsrichtung $d_k \in \mathbb{R}^n$ A-orthogonal zum Fehlervektor $e_{k+1} \in \mathbb{R}^n$ sein soll, eine Schrittlänge $\alpha_k > 0$ finden konnten, welche diese Bedingung erfüllt.

Andersherum könnte man fragen, welche Bedingung man aus der Optimalität einer unbekannten Schrittlänge $\alpha > 0$ folgern könnte. Dazu betrachen wir wieder die notwendigen Optimalitätsbedingungen im totalen Differential

$$\begin{aligned} 0 \;\overset{!}{=}\; \frac{\mathrm{d}}{\mathrm{d}\alpha} F(x_{k+1}) \;&=\; \langle \nabla F(x_{k+1}), \frac{\mathrm{d}}{\mathrm{d}\alpha} x_{k+1} \rangle \;=\; \langle -r_{k+1}, \frac{\mathrm{d}}{\mathrm{d}\alpha}(x_k + \alpha d_k) \rangle \\ &=\; \langle -r_{k+1}, d_k \rangle \;=\; \langle Ae_{k+1}, d_k \rangle. \end{aligned}$$

Wir erhalten also für die Optimalität der unbekannten Schrittlänge $\alpha > 0$, dass die Abstiegsrichtung $d_k \in \mathbb{R}^n$ und der Fehlervektor e_{k+1} konjugiert bezüglich der Matrix A sein müssen. Das ist aber genau die Eigenschaft, die wir bereits in (2.39) gefordert hatten. Wegen der positiven Definitheit der Matrix A können wir ebenfalls folgern, dass die Forderung, dass e_{k+1} und d_k konjugiert bezüglich A sind, bei optimaler Schrittweite

α_k aus (2.40) in jedem Schritt zu einem Abstieg in Richtung d_k führt. Wir erhalten also für eine gegebene Menge von A-orthogonalen Vektoren $\{d_0, \dots, d_{n-1}\}$ ein Iterationsverfahren mit optimalen Schrittlangen $\alpha_k > 0$, die wir in (2.40) angeben können und die uns einen Abstieg garantieren.

Zunächst benötigen wir die Einsicht aus folgendem Lemma, die es uns ermöglicht eine Basis aus A-orthogonalen Vektoren des $\mathbb{R}^n$ zu betrachten.

LEMMA 2.35: Basis von A-orthogonalen Vektoren.

Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische, positiv definite Matrix und sei $\{d_0, \dots, d_{n-1}\}$ eine Menge von A-orthogonalen Vektoren mit $d_k \in \mathbb{R}^n \setminus \{0\}$, d.h., es gilt $\langle d_i, Ad_j\rangle = \langle Ad_i, d_j\rangle = 0$ für alle $i \neq j$.
Dann bilden die Vektoren $d_0, \dots, d_{n-1} \in \mathbb{R}^n \setminus \{0\}$ eine Basis des $\mathbb{R}^n$.

Beweis. In den Übungsaufgaben zu zeigen. □

Folgender Satz zeigt uns, dass das Verfahren für eine gegebene Menge von A-konjugierten Vektoren in der Tat in n Schritten das lokalen Minimum $x^* \in \mathbb{R}^n$ von F erreicht.

THEOREM 2.36: Konvergenz des CG-Verfahrens.

Sei eine Menge von A-konjugierten Vektoren $\{d_0, \dots, d_{n-1}\}$ mit $d_k \in \mathbb{R}^n/\{0\}$ gegeben. Dann konvergiert das Abstiegsverfahren in konjugierten Richtungen

$$x_{k+1} = x_k + \alpha_k d_k, \qquad \alpha_k = \frac{\langle r_k, d_k\rangle}{\langle d_k, Ad_k\rangle} \tag{2.41}$$

in genau n Schritten gegen die Lösung $x^* \in \mathbb{R}^n$ des quadratischen Optimierungsproblems (2.32).

Beweis. Für den Beweis der Konvergenz des Iterationsverfahrens (2.41) betrachten wir zunächst den initialen Fehler $e_0 \in \mathbb{R}^n$ durch die Wahl eines beliebigen Startpunktes $x_0 \in \mathbb{R}^n$. Da wir A als positiv definit angenommen haben folgt mit Lemma 2.35, dass die Menge $\{d_k\}_{k=0,\dots,n-1}$ eine Basis des $\mathbb{R}^n$ bildet. Daher können wir den initialen Fehler $e_0 \in \mathbb{R}^n$ als Linearkombination in dieser Basis darstellen als:

$$e_0 = \sum_{k=0}^{n-1} \delta_k d_k. \tag{2.42}$$

Um die unbekannten Koeffizienten $\delta_k \in \mathbb{R}^n$ zu bestimmen können wir obige Gleichung nun jeweils von links mit einem Vektor $d_i^T A, i = 0, \dots, n-1$ multiplizieren und erhalten so für jeden Index eine Gleichung

$$\langle d_i^T A, e_0\rangle = \langle d_i^T A, \sum_{k=0}^{n-1} \delta_k d_k\rangle = \sum_{k=0}^{n-1} \delta_k \langle d_i^T A, d_k\rangle = \delta_i \langle d_i^T A, d_i\rangle.$$

Hierbei haben wir die Linearität des Skalarproduktes in $\mathbb{R}^n$ ausgenutzt und verwendet, dass die Vektoren $\{d_k\}_{k=0,\dots,n-1}$ konjugiert bezüglich der Matrix A sind. Damit können wir nach den unbekannten Koeffizienten $\delta_i \in \mathbb{R}$ in jeder Gleichung auflösen und erhalten so einen Ausdruck für die unbekannten Koeffizienten:

$$\delta_i \;=\; \frac{\langle d_i^T A, e_0\rangle}{\langle d_i^T A, d_i\rangle}.$$

Man beachte, dass dieser Ausdruck wohldefiniert ist, da wir angenommen haben, dass die Matrix A positiv definit ist. Wir addieren eine Null hinzu, indem wir Terme hinzufügen, die A-konjugiert zur Richtung $d_i \in \mathbb{R}^n$ sind:

$$\delta_i \;=\; \frac{\langle d_i^T A, e_0\rangle}{\langle d_i^T A, d_i\rangle} + \underbrace{\frac{\langle d_i^T A, \sum_{k=0}^{i-1}\alpha_k d_k\rangle}{\langle d_i^T A, d_i\rangle}}_{=0} \;=\; \frac{\langle d_i^T A, e_0 + \sum_{k=0}^{i-1}\alpha_k d_k\rangle}{\langle d_i^T A, d_i\rangle}. \tag{2.43}$$

Wir verwenden wieder den Trick, dass sich der Fehlervektor $e_{i+1} \in \mathbb{R}^n$ entwickeln lässt zu $e_{i+1} = e_i + \alpha_i d_i$ und somit können wir rekursiv herleiten, dass

$$e_i \;=\; e_0 + \sum_{k=0}^{i-1}\alpha_k d_k. \tag{2.44}$$

Nun können wir die Gleichung (2.44) in die Darstellung der Koeffizienten δ_i in (2.43) einsetzen und erhalten:

$$\delta_i \;=\; \frac{\langle d_i^T A, e_0 + \sum_{k=0}^{i-1}\alpha_k d_k\rangle}{\langle d_i^T A, d_i\rangle} \;=\; \frac{\langle d_i^T A, e_i\rangle}{\langle d_i^T A, d_i\rangle} \;=\; \frac{\langle d_i, Ae_i\rangle}{\langle d_i^T A, d_i\rangle} \;=\; -\frac{\langle d_i, r_i\rangle}{\langle d_i^T A, d_i\rangle} \;=\; -\alpha_i.$$

Das bedeutet, dass die Koeffizienten δ_i in (2.43) gerade den negativen optimalen Schrittweiten α_i in (2.40) entsprechen, d.h., $\delta_i = -\alpha_i$. Aus der Basisdarstellung des initialen Fehlers $e_0 = x_0 - x^*$ in (2.42) können wir somit die Behauptung des Satzes folgern:

$$x^* \;=\; x_0 - e_0 \;=\; x_0 - \sum_{k=0}^{n-1}\delta_k d_k \;=\; x_0 + \sum_{k=0}^{n-1}\alpha_k d_k \;=\; x_n.$$

□

BEMERKUNG 2.37 (Veränderung des Fehlers im Iterationsverfahren).
Anstatt im Beweis von Theorem 2.36 zu zeigen, dass sich das lokale Minimum $x^* \in \mathbb{R}^n$ durch das Iterationsverfahren zerlegen lässt, hätte man auch zeigen können, dass der Fehlervektor $e_i \in \mathbb{R}^n$ in jedem Schritt des Iterationsverfahren kleiner wird. Es gilt nämlich nach (2.44):

$$e_i \;=\; e_0 + \sum_{k=0}^{i-1}\alpha_k d_k \;=\; \sum_{k=0}^{n-1}\delta_k d_k + \sum_{k=0}^{i-1} -\delta_k d_k \;=\; \sum_{k=i}^{n-1}\delta_k d_k.$$

Man sieht also, dass für eine wachsende Anzahl an Iterationen $i = 0, \dots, n-1$ der Fehlerterm $e_i \in \mathbb{R}^n$ immer weniger Terme hat, bis er schlussendlich ganz verschwindet. Außerdem sagt es uns, dass der Abstieg mit konjugierten Richtungen in dem Sinne optimal ist, als dass der Fehlerterm $e_i = \sum_{k=i}^{n-1} \delta_k d_k$ keine Anteile der Richtungen $\{d_j\}_{j=0,\dots,k-1}$ mehr besitzt. Wir müssen also nicht mehr entlang dieser Richtungen gehen, um zum lokalen Minimum $x^* \in \mathbb{R}^n$ von F zu gelangen. Aus Sicht der Numerik ist das eine sehr schöne Eigenschaft, da wir nicht gezwungenermaßen n Iterationen des Abstiegsverfahrens (2.41) durchführen müssen, sondern bereits nach $k < n$ abbrechen können, um eventuell eine gute Approximation des lokalen Minimums $x_k \approx x^* \in \mathbb{R}^n$ zu erhalten. Dies spielt insbesondere bei sehr großen Dimensionen $n \gg 1$ eine wichtige Rolle. △

BEISPIEL 2.38.

Wir wollen im Folgenden ein Beispiel zur Durchführung eines Abstiegsverfahrens mit gegebenen konjugierten Richtungen angeben. Seien folgende Werte für das lineare Gleichungssystem $Ax = b$ gegeben:

$$A = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}, \quad b = \begin{pmatrix} 2 \\ -8 \end{pmatrix}.$$

Wir nehmen eine Menge von zwei A-orthogonalen Vektoren $d_0, d_1 \in \mathbb{R}^2/\{0\}$ als gegeben an mit:

$$d_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad d_1 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}.$$

Wir sehen ein, dass die Vektoren d_0 und d_1 konjugiert bezüglich der Matrix A sind, denn es gilt:

$$\langle d_0, Ad_1 \rangle = (0,1) \cdot \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 3 \\ -1 \end{pmatrix} = (0,1) \cdot \begin{pmatrix} 7 \\ 0 \end{pmatrix} = 0.$$

Als Startwert für unser Iterationsverfahren wählen wir $x_0 = (-2, 2)^T$. Für den ersten Schritt des Iterationsverfahren berechnen wir zuerst das aktuelle Residuum

$$r_0 = b - Ax_0 = \begin{pmatrix} 2 \\ -8 \end{pmatrix} - \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} -2 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ -8 \end{pmatrix} - \begin{pmatrix} -2 \\ 8 \end{pmatrix} = \begin{pmatrix} 4 \\ -16 \end{pmatrix}.$$

Nun können wir die optimale Schrittweite $\alpha_0 > 0$ für den ersten Schritt durch den Ausdruck (2.40) bestimmen mit:

$$\alpha_0 = \frac{\langle d_0, r_0 \rangle}{\langle d_0, Ad_0 \rangle} = \frac{4}{3}.$$

Hiermit können wir den ersten Abstieg durchführen und erhalten so den nächsten Iterationspunkt

$$x_1 = x_0 + \alpha_0 d_0 = \begin{pmatrix} -2 \\ -2/3 \end{pmatrix}.$$

Wir wollen nur den zweiten Schritt des Verfahrens angehen und benötigen wiederum das aktuelle Residuum

$$r_1 \;=\; b - Ax_1 \;=\; \begin{pmatrix} 28/3 \\ 0 \end{pmatrix}.$$

Wir berechnen wieder die neue optimale Schrittweite mittels (2.40):

$$\alpha_1 \;=\; \frac{\langle d_1, r_1 \rangle}{\langle d_1, Ad_1 \rangle} \;=\; \frac{28}{21}.$$

Mit dieser können wir den letzten Abstiegsschritt für $n = 2$ berechnen und erhalten somit:

$$x_2 \;=\; x_1 + \alpha_1 d_1 \;=\; \begin{pmatrix} 2 \\ -2 \end{pmatrix} \;=\; x^*.$$

Der folgende Satz hilft uns zu verstehen, warum ein Abstiegsverfahren mit konjugierten Richtungen besser funktioniert als das Gradientenabstiegsverfahren in Abschnitt 2.2.1.

THEOREM 2.39: Orthogonalität des Residuums.

Das Residuum $r_{i+1} = b - Ax_{i+1}$ des Abstiegsverfahren mit konjugierten Richtungen in (2.41) ist orthogonal zu allen bisherigen Abstiegsrichtungen $d_j, j = 0, \ldots, i$, d.h.

$$\langle r_{i+1}, d_j \rangle \;=\; 0, \qquad \text{für alle } j = 0, \ldots, i.$$

Beweis. Aus Bemerkung 2.37 wissen wir, dass wir den Fehler e_{i+1} nach i Iterationen des Abstiegsverfahrens angeben können als

$$e_{i+1} \;=\; \sum_{k=i+1}^{n-1} \delta_k d_k.$$

Wir können beide Seiten der Gleichung mit einem Zeilenvektor $-d_j^T A \in \mathbb{R}^n$ für einen Index $0 \leq j \leq i$ multiplizieren und erhalten damit:

$$-\langle d_j, Ae_{i+1} \rangle \;=\; -\sum_{k=i+1}^{n-1} \delta_k \underbrace{d_j A d_k}_{=\,0} \quad \Rightarrow \; \langle d_j, r_{i+1} \rangle \;=\; 0 \quad \text{für alle } \; 0 \leq j \leq i.$$

□

Man beachte, dass die Eigenschaft optimal bezüglich **aller vorherigen Abstiegsrichtungen** nur für den Fall von konjugierten Richtungen funktioniert und nicht im Fall des Gradientenabstiegsverfahren, wie wir in Abschnitt 2.3.2 gesehen haben. Hier war man nur optimal bezüglich der **letzten Abstiegsrichtung** und nicht bezüglich aller vorherigen Richtungen. Das resultiert in dem typischen Zickzack-Pfad beim Abstieg, wie wir ihn in Abb. 2.4 gesehen haben.

2.3.5 Konjugierte Gradienten

Wir haben in Abschnitt 2.3.4 gesehen, dass wir ein iteratives Abstiegsverfahren mit konjugierten Abstiegsrichtungen $\{d_j\}_{j=0,\dots,n-1}$ verwenden können, um in n Iterationen die eindeutige Lösung des quadratischen Minimierungsproblems (2.32) und somit die Lösung des linearen Gleichungssystems $Ax = b$ zu erhalten. Bisher sind wir jedoch davon ausgegangen, dass wir die Menge der konjugierten Vektoren $\{d_0, \dots, d_{n-1}\}$ bereits kennen. Um einen Algorithmus angeben zu können müssen wir also noch ergründen, wie sich diese Menge mit möglichst geringen numerischen Aufwand finden lässt.

Eine naheliegende Idee wäre es das **Gram-Schmidtsche Orthogonalisierungsverfahren** so umzugestalten, dass wir eine Menge von linear unabhängigen Vektoren $\{u_0, \dots, u_{n-1}\}$ mit $u_k \in \mathbb{R}^n, k = 0, \dots, n-1$ konjugieren bezüglich der Matrix A. Hierzu würde man die erste Abstiegsrichtung $d_0 \in \mathbb{R}^n$ des Abstiegsverfahrens mit konjugierten Richtungen als den ersten Vektor der Menge wählen, d.h., wir setzen $d_0 = u_0$. Anschließend konstruieren wir die nächste Abstiegsrichtung d_1 indem wir alle Komponenten von u_1 entfernen, die nicht A-orthogonal zu d_0 sind. Für die nächste Abstiegsrichtung d_2 gehen wir analog vor, nur müssen wir darauf achten alle Komponenten von u_2 zu entfernen, die nicht A-orthogonal zu d_0 und d_1 sind. Dieses Vorgehen lässt sich iterativ bis zum Vektor d_{n-1} fortführen und man erhält eine Menge von konjugierten Vektoren $\{d_0, \dots, d_{n-1}\}$. Diese lassen sich in geschlossener Form angeben als:

$$d_i \;=\; u_i + \sum_{k=0}^{i-1} \beta_{i,k} d_k, \quad i = 1, \dots, n-1. \tag{2.45}$$

Wir müssen jedoch die Koeffizienten $\beta_{i,k}$ so bestimmen, dass die Vektoren d_i konjugiert zu allen vorherigen Richtungsvektoren $d_j \in \mathbb{R}^n, 0 \le j < i$ sind. Um diese Koeffizienten zu bestimmen multiplizieren wir (2.45) wieder von links mit einem Zeilenvektor $d_j^T A \in \mathbb{R}^n$ für ein $j \in \{0, \dots, i-1\}$ und erhalten

$$\begin{aligned}
& \langle d_j, Ad_i \rangle \;=\; \langle d_j, Au_i \rangle + \sum_{k=0}^{i-1} \beta_{i,k} \langle d_j, Ad_k \rangle \\
\Rightarrow \quad & 0 \;=\; \langle d_j, Au_i \rangle + \beta_{i,j} \langle d_j, Ad_j \rangle \\
\Rightarrow \quad & \beta_{i,j} \;=\; -\frac{\langle d_j, Au_i \rangle}{\langle d_j, Ad_j \rangle}.
\end{aligned}$$

Der Ausdruck für die Koeffizienten $\beta_{i,j}$ ist wohldefiniert, da wir angenommen haben, dass A eine symmetrische, positiv definite Matrix ist. Eigentlich könnten wir jetzt zufrieden sein, da wir ein Verfahren angeben können mit dem sich ein Abstiegsverfahren mit konjugierten Richtungen konstruieren lässt.

Leider haben wir durch die Verwendung des Gram-Schmidtschen Orthogonalisierungsverfahrens nichts gewonnen, da der numerische Aufwand zur Berechnung der unbekannten Koeffizienten im besten Fall in $\mathcal{O}(n^3)$ liegt, was genau so teuer ist wie eine Invertierung der Matrix A, zum Beispiel mit dem Eliminationsverfahren von Gauss [Numerik 1, Kapitel 1.1].

Glücklicherweise gibt es eine Möglichkeit eine Menge von konjugierten Abstiegsrichtungen im Laufe des Iterationsverfahren (2.41) zu konstruieren ohne den numerischen Rechenaufwand des modifizierten Gram-Schmidtschen Orthogonalisierungsverfahren zu benötigen. Hierzu werden wir ähnlich mit einer Menge von initialen Richtungsvektoren $\{u_0, \dots, u_{n-1}\}$ beginnen und diese geeignet anpassen. Es stellt sich nämlich heraus, dass in jeder Iteration ein Vektor existiert, der bereits A-orthogonal zu allen vorherigen Abstiegsrichtungen ist mit Ausnahme der letzten. Diese Aussage wird durch folgendes Lemma präzisiert.

LEMMA 2.40: Eigenschaften des Residuums.

Für $i \in \{1, \dots, n-1\}$ befinden wir uns im $(i+1)$-ten Schritt des Abstiegsverfahrens mit konjugierten Richtungen in (2.41) und $\{d_0, \dots, d_{i-1}\}$ ist eine Menge von A-orthogonalen Vektoren und $u_i = r_i$ sei ein initialer Richtungsvektor für den aktuellen Iterationsschritt. Dann gilt für $r_{i+1} = b - Ax_{i+1} = b - A(x_i + \alpha_i r_i)$:

$$\langle r_{i+1}, r_j \rangle \;=\; 0, \qquad \text{für alle} \;\; j = 0, \dots, i,$$

und außerdem auch

$$\langle r_{i+1}, Ad_j \rangle \;=\; 0, \qquad \text{für alle} \;\; j = 0, \dots, i-1.$$

Beweis. Wir definieren zuerst die lineare Hülle, die durch die A-orthogonalen Vektoren aufgespannt wird durch:

$$\mathcal{D}_i \;:=\; \operatorname{span}\{d_0, \dots, d_{i-1}\} \subset \mathbb{R}^n.$$

Wir gehen in diesem Beweis konstruktiv vor und wollen in jedem Schritt den unbekannten A-orthogonalen Vektor $d_i \in \mathbb{R}^n$ aus dem aktuellen Residuum $r_i \in \mathbb{R}^n$ und den vorigen Richtungsvektoren $\{d_0, \dots, d_{i-1}\}$ konstruieren.

Wir wählen als erste Abstiegsrichtung $d_0 = r_0$ und erhalten damit:

$$\operatorname{span}\{r_0\} \;=\; \operatorname{span}\{d_0\} \;=\; \mathcal{D}_1.$$

Aus Theorem 2.39 wissen wir, dass r_1 orthogonal zu d_0 steht und somit folgt auch schon:

$$\langle r_0, r_1 \rangle \;=\; \langle d_0, r_1 \rangle \;=\; 0.$$

Da wir den nächsten A-orthogonalen Richtungsvektor $d_1 \in \mathbb{R}^n$ aus dem aktuellen Residuum r_1 und dem Unterraum $\mathcal{D}_1$ konstruieren wollen, können wir damit folgern:

$$\mathcal{D}_2 \;=\; \operatorname{span}\{\mathcal{D}_1, d_1\} \;=\; \operatorname{span}\{\mathcal{D}_1, r_1\} \;=\; \operatorname{span}\{r_0, r_1\}.$$

Analog können wir nun für beliebiges $i \in \{1, \dots, n-1\}$ folgern, dass

$$\mathcal{D}_i \;=\; \operatorname{span}\{r_0, \dots, r_{i-1}\}.$$

Aus Theorem 2.39 wissen wir, dass $r_{i+1} \perp \mathcal{D}_{i+1}$ und damit wissen wir schon, dass die erste Aussage des Satzes gilt:

$$\langle r_{i+1}, r_j \rangle \;=\; 0, \qquad \text{für alle } j = 0, \dots, i.$$

Für die zweite Aussage des Lemmas drücken wir nun das Residuum r_i durch den Fehler e_i aus und erhalten:

$$r_i \;=\; -Ae_i \;=\; -A(e_{i-1} + \alpha_{i-1} d_{i-1}) \;=\; \underbrace{r_{i-1}}_{\in \mathcal{D}_i} - \underbrace{\alpha_{i-1} A d_{i-1}}_{\in A\mathcal{D}_i}.$$

Wir sehen also, dass $r_i \in \operatorname{span}\{r_{i-1}, Ad_i\}$. Außerdem wissen wir durch unsere Folgerungen oben, dass $r_{i-1} \in \mathcal{D}_i$ und $Ad_{i-1} \in A\mathcal{D}_i$ gilt. Damit gilt aber schon

$$\mathcal{D}_{i+1} \;=\; \operatorname{span}\{\mathcal{D}_i, r_i\} \;=\; \operatorname{span}\{\mathcal{D}_i, A\mathcal{D}_i\}.$$

Wenn wir dies rekursiv entwickeln sehen wir interessanterweise ein, dass

$$\mathcal{D}_i \;=\; \operatorname{span}\{d_0, Ad_0, \dots, A^{i-1} d_0\}$$

Nach Theorem 2.39 wissen wir jedoch auch, dass $r_{i+1} \perp \mathcal{D}_{i+1}$ und somit muss gelten auch für den Unterraum gelten $r_{i+1} \perp A\mathcal{D}_i$. Und damit haben wir die zweite Aussage des Satzes gezeigt, nämlich dass $r_{i+1} \perp_A \mathcal{D}_i$ oder

$$\langle r_{i+1}, Ad_j \rangle, \qquad \text{für alle } j = 0, \dots, i-1.$$

□

Lemma 2.40 sagt aus, dass das Residuum r_i ein guter Ausgangspunkt für einen weiteren A-orthogonalen Richtungsvektor $d_i \in \mathbb{R}^n$ im Punkt x_i ist, da es bereits zu allen bisherigen A-orthogonalen Richtungen $d_0, \dots, d_{i-2}$ konjugiert bezüglich der Matrix A ist. Wir müssen also nur noch dafür sorgen, dass der initiale Richtungsvektor $r_i \in \mathbb{R}^n$ A-orthogonal zur letzten Suchrichtung d_{i-1} wird. Dies ist numerisch wesentlich günstiger als einen beliebig gewählten Richtungsvektor $u_i \in \mathbb{R}^n / \{0\}$ A-orthogonal zu machen bezüglich aller vorigen Richtungsvektoren $d_0, \dots, d_{i-1}$.

Wir wollen also im Folgenden das vollständige Abstiegsverfahren mit konjugierten Richtungen angeben. Da die initialen Richtungen nun als $r_i = -\nabla F(x_i)$ für alle Iterationen $i = 0, \dots, n-1$ gewählt werden, nennt man dieses Verfahren auch das **Abstiegsverfahren der konjugierten Gradienten**.

THEOREM 2.41: Konjugation der Residuen bezüglich A.

Wir befinden uns im $(i+1)$-ten Schritt des Verfahrens der konjugierten Gradienten für $i = 0, \dots, n-1$ in einem Punkt $x_{i+1} \in \mathbb{R}^n$ und wir wählen als initialen Richtungsvektor das aktuelle Residuum $r_{i+1} \in \mathbb{R}^n$, welches nach Lemma 2.40 bereits A-orthogonal zu fast allen vorherigen Richtungsvektoren $\{d_0, \dots, d_{i-1}\}$

ist. Indem wir den neuen Richtungsvektor $d_{i+1} \in \mathbb{R}^n$ definieren als

$$d_{i+1} \;:=\; r_{i+1} + \beta_{i+1} d_i, \qquad \beta_{i+1} \;:=\; \frac{\langle r_{i+1}, r_{i+1}\rangle}{\langle r_i, r_i\rangle}, \tag{2.46}$$

erhalten wir die Eigenschaft, dass dieser Richtungsvektor nun A-orthogonal zu allen bisherigen Richtungsvektoren des Verfahrens ist, d.h.,

$$d_{i+1} \;\perp_A\; d_j, \quad j = 0, \ldots, i.$$

Beweis. Wir müssen den initialen Richtungsvektor $u_{i+1} \in \mathbb{R}^n/\{0\}$ so modifizieren, dass der resultierende Vektor d_{i+1} konjugiert zu d_i bezüglich der Matrix A ist. Mit dem modifizierten Gram-Schmidtschen Orthogonalisierungsverfahren erhalten wir die Form

$$d_{i+1} \;=\; r_{i+1} + \beta_{i+1} d_i, \qquad \beta_{i+1} \;=\; -\frac{\langle r_{i+1}, Ad_i\rangle}{\langle d_i, Ad_i\rangle}.$$

Die neue Abstiegsrichtung $d_{i+1} \in \mathbb{R}^n/\{0\}$ ist nach Konstruktion A-orthogonal zu allen vorherigen Abstiegsrichtungen $\{d_0, \ldots, d_i\}$, jedoch wollen wir den Koeffizienten β_{i+1} noch näher charakterisieren im Folgenden.

Aus dem Beweis von Lemma 2.40 wissen wir, dass wir das aktuelle Residuum ausdrücken können als

$$r_{i+1} \;=\; r_i - \alpha_i A d_i.$$

Wir multiplizieren diese Gleichung von links mit dem Zeilenvektor $r_{i+1}^T \in \mathbb{R}^n$ und erhalten

$$\langle r_{i+1}, r_{i+1}\rangle \;=\; \langle r_{i+1}, r_i\rangle - \alpha_i \langle r_{i+1}, Ad_i\rangle.$$

Wir wissen aus Lemma 2.40 jedoch auch, dass $\langle r_{i+1}, r_i\rangle = 0$ gilt und damit erhalten wir den Ausdruck

$$-\frac{1}{\alpha_i}\langle r_{i+1}, r_{i+1}\rangle \;=\; \langle r_{i+1}, Ad_i\rangle.$$

Wenn wir nun die optimale Schrittweite $\alpha_i = \frac{\langle r_i, d_i\rangle}{\langle d_i, Ad_i\rangle}$ aus (2.41) einsetzen erhalten wir für den Koeffizienten β_{i+1}:

$$\begin{aligned} \langle r_{i+1}, Ad_i\rangle \;&=\; -\frac{1}{\alpha_i}\langle r_{i+1}, r_{i+1}\rangle \;=\; -\frac{\langle d_i, Ad_i\rangle}{\langle r_i, d_i\rangle}\langle r_{i+1}, r_{i+1}\rangle \\ \Rightarrow \frac{\langle r_{i+1}, r_{i+1}\rangle}{\langle r_i, d_i\rangle} \;&=\; -\frac{\langle r_{i+1}, Ad_i\rangle}{\langle d_i, Ad_i\rangle} \;=\; \beta_{i+1}. \end{aligned}$$

Schlussendlich können wir den Nenner in diesem Ausdruck noch weiter umschreiben, da der letzte Richtungsvektor $d_i \in \mathbb{R}^n$ mit dem modifizierten Gram-Schmidt Orthogonalisierungsverfahren auch ausgedrückt werden kann als

$$\langle r_i, d_i\rangle \;=\; \langle r_i, r_i + \beta_i d_{i-1}\rangle \;=\; \langle r_i, r_i\rangle + \beta_i \underbrace{\langle r_i, d_{i-1}\rangle}_{=\,0} \;=\; \langle r_i, r_i\rangle.$$

Das Skalarprodukt in obiger Gleichung verschwindet auf Grund von der Aussage von Theorem 2.39 und somit erhalten wir schlussendlich für den Koeffizienten β_{i+1} den simplen Ausdruck:

$$\beta_{i+1} = \frac{\langle r_{i+1}, r_{i+1} \rangle}{\langle r_i, r_i \rangle}.$$

□

Mit der Herleitung von (2.46) können wir nun einen Algorithmus für das Abstiegsverfahren mit konjugierten Gradienten zum Lösen eines Gleichungssystems $Ax = b$ angeben.

ALGORITHMUS 2.42: Lineares konjugierte Gradientenverfahren.

function x^* =`conjugateGradient`(A, b, x_0)

Initialisierung
$d_0 = r_0 = b - Ax_0$

# Führe genau n Schritte durch
for $k = 0, \ldots, n-1$ **do**

  # Berechne Schrittweite
  $\alpha_k = \frac{\langle r_k, d_k \rangle}{\langle d_k, Ad_k \rangle}$

  # Führe Abstiegsschritt durch
  $x_{k+1} = x_k + \alpha_k d_k$

  if $k < n - 1$ **then**

    # Berechne effizient neues Residuum
    $r_{k+1} = r_k - \alpha_k A d_k$

    # Berechne Koeffizienten
    $\beta_{k+1} = \frac{\langle r_{k+1}, r_{k+1} \rangle}{\langle r_k, r_k \rangle}$

    # Berechne neue Abstiegsrichung mit Gram-Schmidt
    $d_{k+1} = r_{k+1} + \beta_{k+1} d_k$

  end if
end for

Ausgabe des letzten Punktes
$x^* = x_{k+1}$

2.3.6 Verallgemeinerung für nichtlineare Optimierung

Wir haben festgestellt, dass das konjugierte Gradientenverfahren in Algorithmus 2.42 als Minimierung der konvexen quadratischen Funktion $F(x) := \langle x, Ax\rangle - \langle x, b\rangle + c$ konzipiert ist, um das äquivalente lineare Gleichungssystem $Ax = b$ zu lösen. Es ist natürlich zu fragen, ob wir diesen Ansatz anpassen können, um allgemeine konvexe Zielfunktionen oder sogar allgemeine nichtlineare Zielfunktionen F zu minimieren.

Um eine nichtlineare Variante des konjugierte Gradientenverfahrens zu erhalten, müssen wir einige Anpassungen vornehmen. Fletcher und Reeves [**Fletcher**] zeigten, dass man anstelle der explizit berechenbaren Schrittweite α_k für das quadratische Probleme in (2.40) zunächst einmal eine Schrittweite wählen muss, die ein approximatives Minimum der nichtlinearen Zielfunktion F entlang der Suchrichtung $d_k \in \mathbb{R}^n$ identifiziert. Darüber hinaus muss das bisher verwendete Residuum $r_k = b - Ax_x \in \mathbb{R}^n$ für den Einsatz in der nichtlinearen Optimierung durch den Gradienten der Zielfunktion F ersetzt werden.

Diese beiden Änderungen führen zu dem folgenden Algorithmus für die nichtlineare Optimierung.

ALGORITHMUS 2.43: Nichtlineares konjugierte Gradientenverfahren.

function x^* =`nonlinearConjugateGradient`$(F, \nabla F, x_0, \alpha_0, \sigma)$

Initialisierung
$d_0 = -\nabla F(x_0)$

while $||\nabla F(x_k)||^2 > \epsilon$ **do**

  while $F(x_k + \alpha_k d_k) > F(x_k)$ **do**
    # Verringere Schrittweite um Faktor σ
    $\alpha_k = \sigma\alpha_k$
  end while

  # Führe Abstiegsschritt durch
  $x_{k+1} = x_k + \alpha_k d_k$

  # Berechne Koeffizienten
  $\beta_{k+1} = \frac{\langle \nabla F(x_{k+1}), \nabla F(x_{k+1})\rangle}{\langle \nabla F(x_k), \nabla F(x_k)\rangle}$

  # Berechne neue Abstiegsrichung mit Gram-Schmidt
  $d_{k+1} = -\nabla F(x_{k+1}) + \beta_{k+1} d_k$

end while

Ausgabe des letzten Punktes
$x^* = x_{k+1}$

Man beachte, dass man für die Implementierung von Algorithmus 2.43 keinerlei Matrix-Vektor Multiplikationen benötigt und man lediglich Auswertungen der Zielfunktion F und ihres Gradienten ∇F braucht. Für den Fall, dass es sich bei der Zielfunktion F um eine strikt konvexe, quadratische Funktion handelt und wir in jedem Schritt die optimale Schrittweite $\alpha_k > 0$ bestimmten können, so entspricht Algorithmus 2.43 dem linearen konjugierte Gradientenverfahren in Algorithmus 2.42.

In der Literatur hat sich unter Anderem eine Modikation des nichtlinearen konjugierte Gradientenverfahrens von Fletcher und Reeves etabliert, die sich in numerischen Experimenten häufig als robuster und effizienter herausgestellt hat.

BEMERKUNG 2.44 (Polak-Ribière Variante). Bei der sogenannten Polak-Ribière Variante des Algorithmus 2.43 unterscheidet sich hauptsächlich die Berechnung des Parameters $\beta_{k+1} \in \mathbb{R}$ in (2.46) für die Anpassung der nächsten Abstiegsrichtung. Hierbei wird der Faktor β_{k+1} nämlich wie folgt berechnet

$$\beta_{k+1}^{PR} \;:=\; \frac{\langle \nabla F(x_{k+1}), \nabla F(x_{k+1}) - \nabla F(x_k) \rangle}{\langle \nabla F(x_k), \nabla F(x_k) \rangle}.$$

Man sieht ein, dass im Falle einer strikt konvexen, quadratischen Zielfunktion F mit optimaler Schrittweitenwahl für die $\alpha_k > 0$ die Orthogonalitätsbedingung für sukzessive Gradienten aus Lemma 2.32 gilt mit

$$\langle \nabla F(x_{k+1}), \nabla F(x_k) \rangle \;=\; 0.$$

In diesem Fall stimmt der Faktor β_{k+1}^{PR} mit dem Faktor β_{k+1} aus Algorithmus 2.43 überein. In allen anderen Fällen hingegen unterscheiden sich die Faktoren im Allgemeinen und führen so zu euben signifikant unterschiedlichen Konvergenzverhalten der jeweiligen Abstiegsverfahren. △

2.4 Wahl der Schrittweite

Wir haben in den vorausgegangenen Abschnitten bereits verschiedene Abstiegsverfahren der Form

$$x_{k+1} \;=\; x_k + \alpha_k p_k, \qquad \alpha_k > 0, \quad p_k \in \mathbb{R}^n,$$

wie in (2.3) kennen gelernt, die basierend auf verschiedenen Annahmen unterschiedliche Abstiegsrichtungen $p_k \in \mathbb{R}^n$ realisiert haben. Bis auf eine heuristische Wahl von adaptiven Schrittweiten in (2.9) und der Wahl einer optimalen Schrittweite im Fall von strikt konvexen, quadratischen Zielfunktionen in (2.40) haben wir uns bisher noch nicht weiter mit der Frage einer geeigneten Wahl der Schrittweiten $\alpha_k > 0$ beschäftigt. Dies wollen wir im Folgenden nachholen.

Wir werden ab jetzt immer annehmen, dass $p_k \in \mathbb{R}^n$ eine fest gewählte Abstiegsrichtung ist, d.h. für die Richtungsableitung in Richtung p_k in jedem Iterationsschritt $k = 0, 1, \ldots$ gilt

$$\langle \nabla F(x_k), p_k \rangle \;<\; 0.$$

Nehmen wir an, dass die Zielfunktion F stetig differenzierbar ist und Ist die gewählte $\alpha_k > 0$ sehr klein, so bleiben wir sicher in einer lokalen Umgebung von x_k in der die folgende Taylor-Approximation erster Ordnung gilt:

$$F(x_k + \alpha_k p_k) \;\approx\; F(x_k) + \alpha_k \langle \nabla F(x_k), p_k \rangle \;<\; F(x_k).$$

Allerdings würde das Iterationsverfahren dann in der Regel sehr langsam konvergieren. Ist andererseits α_k zu groß, dann ist die Abstiegsbedingung nicht mehr garantiert. Es könnte zum Beispiel passieren, dass man beim Iterationsschritt $x_{k+1} = x_k + \alpha_k p_k$ zu weit über ein lokales Minimum springt. Deshalb benötigen wir intuitiv zwei Bedingungen an die Schrittweite $\alpha_k > 0$, die zu kleine und zu große Schritte verhindern sollen.

Zunächst wollen wir analog zu (2.40) eine theoretische Möglichkeit der optimalen Wahl der Schrittweite $\alpha_k > 0$ untersuchen, nämlich jene Schrittweite, die zum größtmöglichen Abstieg führt:

$$\alpha_k \;:=\; \arg\min_{\alpha \in \mathbb{R}^+} F(x_k + \alpha p_k).$$

Um eine optimale Schrittweite α_k ausrechnen zu können, müssen wir beliebige eindimensionale Probleme analytisch lösen können, was jedoch im Allgemeinen schwierig ist. Deshalb fordern wir nicht die analytische Optimalität der Schrittweite $\alpha_k > 0$, sondern versuchen lediglich Bedingungen zu finden, die wir numerisch leicht überprüfen können und für die wir Konvergenz des Abstiegverfahrens garantieren können.

Die Idee hierbei ist es für eine vorgegebene Schrittweite $\alpha > 0$ eine Linearisierung des Problems zu betrachten und den linearisierten Abstieg mit dem echten Abstieg zu vergleichen. Hierzu machen wir zunächst folgende Definitionen.

DEFINITION 2.45: Erwarteter und tatsächlicher Abstieg.

Sei $F\colon \Omega \to \mathbb{R}$ eine stetig differenzierbare Funktion für die wir das Abstiegsverfahren $x_{k+1} = x_k + \alpha p_k$ für $x_k \in \Omega$, $p_k \in \mathbb{R}^n \setminus \{0\}$ und $\alpha > 0$ betrachten. Wir definieren basierend auf der Taylorapproximation erster Ordnung den **erwarteten Abstieg** im Punkt x_k in Richtung p_k mit Schrittweite α als

$$E_k(\alpha) \;:=\; F(x_k) + \alpha \langle \nabla F(x_k), p_k \rangle - F(x_k) \;=\; \alpha \langle \nabla F(x_k), p_k \rangle. \tag{2.47}$$

Darüber hinaus definieren wir für die gleichen Größen den **tatsächlichen Abstieg** als

$$D_k(\alpha) \;:=\; F(x_k + \alpha p_k) - F(x_k). \tag{2.48}$$

Unsere beiden Bedingungen an eine geeignete Schrittweite $\alpha > 0$ können wir nun über die Abweichung des erwarteten und tatsächlichen Abstiegs $D_k(\alpha)$ und $E_k(\alpha)$ in Definition 2.45 formulieren.

DEFINITION 2.46: Armijo-Goldstein Bedingungen.

Sei ein allgemeines Abstiegsverfahren der Form $x_{k+1} = x_k + \alpha_k p_k$ für eine vorgegebene Abstiegsrichtung $p_k \in \mathbb{R}^n$ gegeben und seien $c_1, c_2 \in \mathbb{R}^+$ Konstanten mit $0 < c_1 < c_2 < 1$.
Dann formuliert man die sogenannten **Armijo-Goldstein Bedingungen** für die Wahl einer geeigneten Schrittweite $\alpha_k > 0$ des Abstiegsverfahrens als

$$c_1 E_k(\alpha_k) \; > \; D_k(\alpha_k) \; > \; c_2 E_k(\alpha_k), \tag{2.49}$$

wobei E_k und D_k den erwarteten und tatsächlichen Abstieg aus (2.47) und (2.48) definieren. Da wir den erwarteten Abstieg für kleine Schrittweiten $\alpha_k > 0$ als negativ annehmen, d.h., es gilt $E_k(\alpha_k) < 0$, ist (2.49) sinnvoll definiert.

Die erste Bedingung auf der linken Seite der Armijo-Goldstein Bedingungen in (2.49) garantiert, dass zumindest ein gewisser Teil des Abstiegs erreicht wird. Die zweite Bedingung auf der rechten Seite verhindert, dass wir uns zu stark dem Fall $\alpha_k = 0$ annähern, in dem die rechte Seite zu einer Gleichheit mit Konstante gleich eins wird. Eine typische Wahl der Parameter in Definition 2.46 ist $c_1 = 0.1$ und $c_2 = 0.9$.

In der Praxis lassen sich die Armijo-Goldstein Bedingungen wie folgt einsetzen. Man beginnt mit einer Schrittweite von $\alpha_k = \alpha_{k-1} > 0$, die im letzten Iterationsschritt $k-1$ zu einem Abstieg geführt hat und testet mit dieser die Armijo-Goldstein Bedingungen aus Definition 2.46. Ist die erste Ungleichung nicht erfüllt, d.h., für den tatsächlichen Abstieg gilt $c_1 E_k(\alpha_k) \leq D_k(\alpha_k)$, so verkleinert man die Schrittweite (zum Beispiel durch Halbierung). Ist andererseits die zweite Ungleichung nicht erfüllt, d.h., für den tatsächlichen Abstieg gilt $D_k(\alpha_k) \leq c_2 E_k(\alpha_k)$, so vergrößert man die Schrittweite entsprechend. Um nicht in einen periodischen Zyklus zu geraten, sollte man zur Vergrößerung der Schrittweite einen anderen Faktor als zur Verkleinerung wählen, etwa $\sigma = 1.5$. Die Wahl der Schrittweite nach den Armijo-Goldstein Regeln ist also relativ einfach durchführbar und führt zu einer beweisbaren Konvergenz eines Abstiegsverfahrens, wie das folgende Theorem zeigt.

THEOREM 2.47: Konvergenz von Abstiegsverfahren.

Sei ein Abstiegsverfahren der Form $x_{k+1} = x_k + \alpha_k p_k$ gegeben, mit einer Menge von Vektoren $p_k \in \mathbb{R}^n$, die für jeden Punkt $x_k \in \Omega$, der kein stationärer Punkt ist, eine uniforme Abstiegsrichung liefern, d.h., es existieren fixe Konstanten $\beta, \gamma > 0$, so dass gilt

$$\langle \nabla F(x_k), p_k \rangle \; < \; -\gamma \cdot \|\nabla F(x_k)\|^{\beta+1}.$$

Die Folge der Schrittweiten $(\alpha_k)_{k \in \mathbb{N}}$ erfülle die Armijo-Goldstein Bedingungen. Außerdem sei $F : \mathbb{R}^n \to \mathbb{R}$ eine nach unten beschränkte, stetig differenzierbare Zielfunktion, für die somit die Niveaumenge $K := \{x \in \mathbb{R}^n \mid F(x) \leq F(x_0)\}$ beschränkt ist.

Ist darüber hinaus die Folge der Abstiegsrichtungen $(p_k)_{k\in\mathbb{N}}$ beschränkt, dann bestitzt die Folge der Iterationsschritte $(x_k)_{k\in\mathbb{N}}$ eine konvergente Teilfolge und jeder Häufungspunkt der Folge ist ein stationärer Punkt der Zielfunktion F.

Beweis. Falls für ein $k \in \mathbb{N}$ gilt, dass der Vektor $p_k = \mathbf{0}$ ist, so haben wir bereits einen stationären Punkt erreicht und die Aussage des Theorems ist trivialerweise erfüllt.

Nehmen wir also im Folgenden an, dass $p_k \neq \mathbf{0}$ für alle $k \in \mathbb{N}$ gilt und wir damit einen echten Abstieg vorliegen haben. In diesem Fall impliziert die erste Armijo-Goldstein Bedingung $c_1 E_k(\alpha_k) > D_k(\alpha_k)$, dass gilt

$$F(x_{k+1}) - F(x_k) \;<\; c_1 \cdot \alpha_k \langle \nabla F(x_k), p_k \rangle \;<\; 0.$$

Induktiv gilt somit ebenfalls

$$F(x_{k+1}) < F(x_k) < \ldots < F(x_0). \tag{2.50}$$

Also liegt die gesamte Folge $(x_k)_{k\in\mathbb{N}}$ in der nach Voraussetzung beschränkten Menge $K := \{x \in \mathbb{R}^n \mid F(x) \leq F(x_0)\}$ und hat nach dem *Satz von Bolzano-Weierstrass* somit eine konvergente Teilfolge $(x_{k_\ell})_{\ell\in\mathbb{N}}$. Darüber hinaus sehen wir ein, dass die Menge K ebenfalls abgeschlossen und somit nach dem *Satz von Heine-Borell* sogar kompakt ist. Daher liegt der Grenzwert $x^* \in K$ ebenfalls in der Menge K.

Durch Einsetzen in die folgende Teleskopsumme

$$F(x_k) - F(x_0) \;=\; \sum_{j=0}^{k-1} F(x_{j+1}) - F(x_j)$$

erhalten wir die stärkere Bedingung

$$F(x_k) + c_1 \sum_{j=0}^{k-1} -\alpha_j \langle \nabla F(x_j), p_j \rangle \;<\; F(x_0). \tag{2.51}$$

Mit der nach Voraussetzung geltenden uniformen Schranke folgt dann für fixe Konstanten $\beta, \gamma > 0$ und für alle $k \in \mathbb{N}$ die Ungleichung

$$||\nabla F(x_k)||^{\beta+1} \;<\; -\frac{1}{\gamma} \langle \nabla F(x_k), p_k \rangle.$$

Da die Niveaumenge K kompakt ist wissen wir, dass ein Punkt $x^* \in K$ existiert in der die Zielfunktion F ihr Minimum auf K annimmt. Dann gilt wegen (2.50) offensichtlich für alle $k \in \mathbb{N}$

$$0 \;<\; F(x_0) - F(x_k) \;<\; F(x_0) - F(x^*) \;=:\; M.$$

Zusammen mit der Ungleichung (2.51) können wir somit folgern, dass gilt

$$\sum_{j=0}^{k-1} \alpha_j \|\nabla F(x_j)\|^{\beta+1} \;\leq\; \frac{1}{\gamma} \sum_{j=0}^{k-1} -\alpha_j \langle \nabla F(x_j), p_j \rangle \;\leq\; \frac{1}{\gamma c_1} (F(x_0) - F(x_k)) \;<\; \frac{M}{\gamma c_1}.$$

Damit gilt offensichtlich $\alpha_k ||\nabla F(x_k)||^{\beta+1} \to 0$ und somit gilt ebenfalls $\alpha_k \nabla F(x_k) \to \mathbf{0}$ für $k \to \infty$.

Nun müssen wir abschließend noch zeigen, dass die Folge der Schrittweiten $(\alpha_k)_{k\in\mathbb{N}}$ selbst nicht gegen Null konvergiert, damit $\nabla F(x_k) \to \mathbf{0}$ gilt und somit die Folge der Iterationsschritte $(x_k)_{k\in\mathbb{N}}$ gegen einen stationären Punkt der Zielfunktion F konvergiert. Nehmen wir also das Gegenteil für eine Teilfolge $(\alpha_{k_\ell})_{\ell\in\mathbb{N}}$ an, die gegen Null konvergiert, so gilt auch $\alpha_{k_\ell} p_{k_\ell} \to \mathbf{0}$. Damit gilt aber schon für beliebiges $\epsilon > 0$ mit $(1-\epsilon) > c_2$, dass für hinreichend große $k_\ell \in \mathbb{N}$ die folgende Ungleichung erfüllt ist:

$$F(x_{k_\ell} + \alpha_{k_\ell} p_{k_\ell}) - F(x_{k_\ell}) \;\leq\; (1-\epsilon) \cdot \alpha_{k_\ell} \langle \nabla F(x_{k_\ell}), p_{k_\ell} \rangle \;<\; c_2 \cdot \alpha_{k_\ell} \langle \nabla F(x_{k_\ell}), p_{k_\ell} \rangle.$$

Dies ist jedoch nicht möglich, da die Armijo-Goldstein Bedingungen nach Voraussetzung erfüllt sind. □

Für das Gradientenabstiegsverfahren bzw. dessen Varianten in Abschnitt 2.2.1 können wir im folgenden Korollar noch mehr zeigen„ da die Abstiegsvektoren $p_k \in \mathbb{R}^n$ in direkter Verbindung zum Gradienten der Zielfunktion $-\nabla F(x_k)$ stehen.

KOROLLAR 2.48: Konvergenz des Gradientenabstiegverfahrens.

Sei $F : \mathbb{R}^n \to \mathbb{R}$ eine nach unten beschränkte, stetig differenzierbare Funktion, so dass die Niveaumenge $K := \{x \in \mathbb{R}^n \mid F(x) \leq F(x_0)\}$ beschränkt ist. Gegeben sei außerdem eine Wahl an Abstiegsrichtungen der Form

$$p_k \;=\; -A_k \nabla F(x_k),$$

wobei für jedes $k \in \mathbb{N}$ die Matrix $A_k \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit ist. Der kleinste und größte Eigenwert der Matrizen A_k seien darüber hinaus für jedes $k \in \mathbb{N}$ uniform durch $\lambda_{max} \geq \lambda_{min} > 0$ nach unten bzw. nach oben beschränkt. Dann hat die Folge $(x_k)_{k\in\mathbb{N}}$ eine konvergente Teilfolge und jeder Häufungspunkt ist ein stationärer Punkt von F.

Beweis. In den Übungsaufgaben zu zeigen. □

2.5 Nicht-differenzierbare Optimierung

Im Folgenden widmen wir uns abschließend der Optimierung von nicht-differenzierbaren Zielfunktionen, die man häufig in der Datenanalyse und der mathematischen Bildverarbeitung findet. Hierzu betrachen wir zunächst das folgende motivierende Beispiel.

BEISPIEL 2.49: LASSO-Problem.

Wir betrachten das sogenannte LASSO-Problem (engl. für „*Least Absolute Shrinkage and Selection Operator*“), das man als Variante eines linearen Ausgleichsprob-

lem interpretieren kann (siehe [Numerik 1, Kapitel 3.1.1]) mit

$$F(x) \;=\; \frac{1}{2}\|b - Ax\|^2 + \alpha\|x\|_{\ell^1}.$$

Hierbei ist $A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^n$ und $b \in \mathbb{R}^m$, wobei typischerweise $m < n$ gilt.
Mit der Lösung dieses Problems für $\alpha \to 0$ approximiert man eine Lösung des linearen Gleichungssystems $Ax = b$ mit minimaler ℓ^1-Norm.
Solche Lösungen besitzen in der Regel viele Nulleinträge und sind daher besonders interessant. Das Problem bei der Minimierung der Zielfunktion F ist, dass die ℓ^1-Norm

$$\|x\|_{\ell^1} \;=\; \sum_{j=1}^{n} |x_j|$$

nicht differenzierbar in allen Punkten $x \in \mathbb{R}^n$ ist, für die mindestens eine Koordinate Null ist, d.h., für die $x_j = 0$ gilt für ein $1 \le j \le n$.

Eine interessante Klasse von Zielfunktion, die wir im Folgenden betrachten wollen, ist von der Form

$$F(x) = G(x) + H(x),$$

mit einer stetig differenzierbaren Funktion $G\colon \mathbb{R}^n \to \mathbb{R}$ und einer konvexen, nicht notwendigerweise differenzierbaren Funktion $H\colon \mathbb{R}^n \to \mathbb{R}$. In diesem Fall können wir keins der in den vorangegangenen Kapiteln vorgestellten Gradienten-basierten Optimierungsverfahren verwenden, sondern benötigen einen anderen methodischen Ansatz.

Zunächst benötigen wir aber noch einige Definition um konvexe Funktionen analytisch besser beschreiben zu können.

DEFINITION 2.50: Subdifferential und Subgradient.

Sei $H : \mathbb{R}^n \to \mathbb{R}$ eine konvexe Funktion. Dann ist das **Subdifferential** an der Stelle $x \in \mathbb{R}^n$ definiert durch

$$\partial H(x) \;:= \{p \in \mathbb{R}^n \mid H(x) + \langle p, y - x\rangle \le H(y) \quad \forall\, y \in \mathbb{R}^n\}.$$

Ein Element des Subdifferentials nennen wir **Subgradient**.

Man sieht ein, dass ein globales Minimum der konvexen Funktion H durch die Bedingung $\mathbf{0} \in \partial H(x^*)$ charakterisiert wird, denn mit Definition 2.50 erhalten wir in diesem Fall:

$$\mathbf{0} \in \partial H(x^*) \;\Leftrightarrow\; H(x^*) + \underbrace{\langle \mathbf{0}, y - x^*\rangle}_{=0} \le H(y) \quad \forall y \in \mathbb{R}^n. \tag{2.52}$$

Folgendes Beispiel illustriert das Subdifferential einer nicht differenzierbaren, eindimensionalen Funktion.

BEISPIEL 2.51: Subdifferential der Betragsfunktion.

Wir betrachten die eindimensionale Betragsfunktion $H(x) := |x|$. In diesem Fall ist das Subdifferential von H für alle Punkte $x \neq 0$ gegeben als

$$\partial H(x) \;=\; \{\operatorname{sgn}(x)\}.$$

Im Punkt $x = 0$, in dem die Betragsfunktion bekanntlich nicht differenzierbar ist, gilt für das Subdifferential hingegen

$$\partial H(0) \;=\; [-1, 1].$$

Basierend auf dieser Beobachtung im Eindimensionalen können wir auch das Subdifferential der ℓ^1-Norm im $\mathbb{R}^n$ im folgenden Beispiel ausrechnen.

BEISPIEL 2.52: Subdifferential der ℓ^1-Norm.

Wir betrachten die ℓ^1-Norm, die durch die konvexe Funktion $H(x) = \|x\|_{\ell^1}$ gegeben ist. Dann können wir das Subdifferential der ℓ^1-Norm für alle Punkte $x \in \mathbb{R}^n$ angeben als

$$\partial H(x) \;=\; \{p \in [-1, 1]^n \mid p_i = \operatorname{sgn}(x_i) \text{ für } x_i \neq 0\}.$$

Basierend auf der Optimalitätsbedingung für die konvexe Funktion H in (2.52) können wir eine entsprechende notwendige Bedingung für ein lokales Minimum für die Summe einer differenzierbaren Funktion G und einer konvexen Funktion H herleiten, wie das folgende Lemma besagt.

LEMMA 2.53: Optimalitätsbedingung.

Sei $F\colon \mathbb{R}^n \to \mathbb{R}$ eine Zielfunktion, die sich schreiben lässt als $F = G + H$, wobei G stetig differenzierbar und H konvex ist. Sei außerdem $x^* \in \mathbb{R}^n$ ein lokaler Minimierer von F.
Dann gilt $\mathbf{0} \in \nabla G(x^*) + \partial H(x^*)$, d.h., es existiert ein Subgradient $p \in \partial H(x^*)$ mit $\nabla G(x^*) + p = \mathbf{0}$.

Beweis. Sei $x^* \in \mathbb{R}^n$ ein lokaler Minimierer von F. Dann gilt für jeden Punkt $x \in \mathbb{R}^n$ in einer lokalen Umgebung des Minimierers x^* mit einer Taylor-Entwicklung erster Ordnung der Funktion G schon

$$\begin{aligned} G(x^*) + H(x^*) \;&=\; F(x^*) \;\leq\; F(x) \;=\; G(x) + H(x) \\ &=\; G(x^*) + \langle \nabla G(x^*), x - x^* \rangle + r(x)\|x - x^*\| + H(x) \end{aligned}$$

mit einem Fehlerterm $r\colon \mathbb{R}^n \to \mathbb{R}$ für den gilt $r(x) \to 0$ wenn $x \to x^*$. Wählen wir nun den speziellen Vektor $p = -\nabla G(x^*)$ und stellen die Gleichung um, so gilt also

$$\frac{H(x^*) + \langle p, x - x^* \rangle - H(x)}{\|x - x^*\|} \;=\; r(x).$$

Dann können wir unter Ausnutzung der Konvexität von H folgende Abschätzung treffen

$$\begin{aligned} 0 \;=\; \limsup_{x\to x^*} r(x) \;&\geq\; \limsup_{x\to x^*} \frac{H(x^*) + \langle p, x - x^*\rangle - H(x)}{\|x - x^*\|} \\ &\geq\; \lim_{\lambda\to 0} \frac{H(x^*) + \langle p, \lambda(x - x^*)\rangle - H((1-\lambda)x^* + \lambda x)}{\lambda\|x - x^*\|} \\ &\geq\; \lim_{\lambda\to 0} \frac{H(x^*) + \lambda\langle p, x - x^*\rangle - (1-\lambda)H(x^*) - \lambda H(x)}{\lambda\|x - x^*\|} \\ &=\; \lim_{\lambda\to 0} \frac{\langle p, x - x^*\rangle + H(x^*) - H(x)}{\|x - x^*\|} \;=\; \frac{\langle p, x - x^*\rangle + H(x^*) - H(x)}{\|x - x^*\|}. \end{aligned}$$

Durch Umstellen erhalten wir schließlich

$$H(x^*) + \langle p, x - x^*\rangle \;\leq\; H(x).$$

Dies bedeutet aber schon nach Definition 2.50, dass $p \in \partial H(x^*)$ gilt. □

Basierend auf Lemma 2.53 könnten wir also von der notwendigen Optimalitätsbedingung $\nabla G(x^*) + p = 0$ ausgehen und damit ein Analogon zum Gradientenabstiegsverfahren aus Abschnitt 2.2.1 aufstellen mit

$$x_{k+1} = x_k - \alpha_k(\nabla G(x_k) + p_k), \qquad p_k \in \partial H(x_k).$$

Allerdings ist nicht klar welchen Subgradienten $p_k \in \partial H(x_k)$ wir in jeder Iteration auswählen sollen (falls mehr als einer existiert) oder ob überhaupt ein Subgradient existiert. Deshalb werden wir im Folgenden eine Variante betrachten bei der automatisch Iterationsschritte $x_{k+1} \in \mathbb{R}^n$ mit nichtleerem Subdifferential und ebenso ein $p_{k+1} \in \partial H(x_{k+1})$ ausgewählt werden.

2.5.1 Proximales Splitting

Die Idee des sogenannten proximalen Splitting, auch *Forward-Backward Splitting* genannt, ist es den differenzierbaren Teil genauso wie beim Gradientenabstiegsverfahren auszuwerten, d.h., vorwärts ausgehend vom Punkt $x_k \in \mathbb{R}^n$, während der Subgradient hingegen bezüglich der nächsten Iterierten ausgewertet wird, d.h., rückwärts ausgehend vom Punkt $x_{k+1} \in \mathbb{R}^n$.

Somit lässt sich die Iterationsvorschrift für das proximale Splitting schreiben als

$$x_{k+1} \;=\; x_k - \alpha_k(\nabla G(x_k) + p_{k+1}), \qquad p_{k+1} \in \partial H(x_{k+1}). \tag{2.53}$$

Diese Gleichung können wir nun selbstverständlich nicht mehr explizit auswerten, da der Subgradient $p_{k+1} \in \partial H(x_{k+1})$ vom bisher unbestimmten Iterationsschritt $x_{k+1} \in \mathbb{R}^n$ abhängt.

Dennoch können wir die Iterationsvorschrift (2.53) weiter umschreiben zu

$$\frac{1}{\alpha_k}(x_{k+1} - x_k) + \nabla G(x_k) + p_{k+1} \;=\; 0.$$

Die zentrale Idee ist es nun eine Funktion zu finden, deren Ableitung den Ausdruck auf linken Seite der obigen Gleichung liefert. Man sieht ein, dass die obige Gleichung die hinreichende Optimalitätsbedingung für die Minimierung der folgenden strikt konvexen Funktion darstellt

$$F_k(x) \;=\; \frac{1}{2\alpha_k}\|x - x_k + \alpha_k \nabla G(x_k)\|^2 + H(x).$$

Hierbei stellt die Iterierte $x_{k+1} \in \mathbb{R}^n$ also einen stationären Punkt der Zielfunktion $F_k(x)$ dar, während $p_{k+1} \in \mathbb{R}^n$ aus dem Subdifferential $\partial H(x)$ stammt.

Wir können das Iterationsverfahren in (2.53) also durchführen, wenn wir strikt konvexe Funktionen der Form

$$\Phi(x) \;:=\; \frac{1}{2\alpha}\|x - y\|^2 + H(x)$$

effizient minimieren können.

Bei der Optimierung der Funktion Φ versucht man anschaulich einen Kompromiss einzugehen zwischen dem Ziel die Funktion H zu minimieren und gleichzeitig nahe dem Punkt $y \in \mathbb{R}^n$ bezüglich der Euklidischen Norm zu sein. Der Parameter $\alpha > 0$ steuert hierbei die Gewichtung zwischen diesen beiden Kriterien.

Da es sich bei Φ um eine strikt konvexe Funktion handelt existiert ein eindeutiges Minimum. Dieses lässt sich durch den sogenannten Proximaloperator beschreiben, den wir im Folgenden definieren wollen.

DEFINITION 2.54: Proximaloperator.

Sei $H\colon \mathbb{R}^n \to \mathbb{R}$ eine konvexe, unterhalbstetige Funktion und sei $\alpha > 0$ ein positiver Parameter. Dann definieren wir den **Proximaloperator** bezüglich der Funktion H im Punkt $y \in \mathbb{R}^n$ als

$$\operatorname{prox}_{\alpha H}(y) \;:=\; \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \left\{ \frac{1}{2\alpha}||x - y||^2 + H(x) \right\}.$$

Wir wollen zunächst das Konzept des Proximaloperators im folgenden Beispiel für die Betragsfunktion bzw. die ℓ^1-Norm genauer verstehen.

BEISPIEL 2.55: Shrinkage-Operator.

Sei $H(x) = |x|$ in diesem Beispiel zunächst die Betragsfunktion. Dann ist der Proximaloperator $z = \operatorname{prox}_{\alpha H}(y)$ der eindeutige Minimierer der strikt konvexen Funktion

$$\Phi(x) \;=\; \frac{1}{2\alpha}(x - y)^2 + |x|$$

mit der notwendigen Optimalitätsbedingung 1. Ordnung

$$\frac{1}{\alpha}(y - z) \in \partial|z|. \tag{2.54}$$

Aus Beispiel 2.51 wissen wir bereits, dass $\partial|z| = \{\operatorname{sgn}(z)\} = \{-1, 1\}$ für $z \neq 0$ gilt und $\partial|0| = [-1, 1]$ in der Null.

Betrachten wir zunächst den Fall $z > 0$. Dann ist klar, dass $\partial|z| = 1$ gilt und somit wird aus der notwendigen Optimalitätsbedingung (2.54)

$$\frac{1}{\alpha}(y - z) = 1 \quad \Leftrightarrow \quad z = y - \alpha.$$

Da wir $z > 0$ angenommen haben, ist dies nur möglich für $y > \alpha > 0$.
Analog gilt für den Fall $z < 0$, dass $\partial|z| = -1$ ist und somit erhalten wir aus der notwendigen Optimalitätsbedingung (2.54)

$$\frac{1}{\alpha}(y - z) = -1 \quad \Leftrightarrow \quad z = \alpha + y.$$

Dies ist aber für $z < 0$ nur möglich, wenn $y < -\alpha < 0$ gilt.
Für den letzten Fall mit $z = 0$ lauten die notwendigen Optimalitätsbedingung (2.54)

$$\frac{1}{\alpha}(y - z) = \frac{y}{\alpha} \in [-1, 1] = \partial|0|.$$

Dies ist äquivalent zu der Bedingung $-1 \leq \frac{y}{\alpha} \leq 1$, die nur dann erfüllt ist wenn $-\alpha \leq y \leq \alpha$ gilt.
Somit lässt sich der Proximaloperator $z = \text{prox}_{\alpha H}(y)$ für die Betragsfunktion $H(x) := |x|$, der auch **Shrinkage-Operator** genannt wird, wie folgt angeben:

$$\text{prox}_{\alpha|\cdot|}(y) = \text{shrink}_\alpha(y) := \begin{cases} y - \alpha & \text{falls } y > \alpha \\ 0 & \text{falls } y \in [-\alpha, \alpha] \\ y + \alpha & \text{falls } y < -\alpha. \end{cases}$$

Die kontrahierende Wirkung des Shrinkage-Operators wird in **??** illustriert.
Analog können wir auch den Proximaloperator für die ℓ^1-Norm $H(x) = \|x\|_{\ell^1}$ im $\mathbb{R}^n$ angeben als

$$\text{prox}_{\alpha H}(y) = (\text{shrink}_\alpha(y_i))_{i=1,\dots,n}.$$

Mit Hilfe des Proximaloperators können wir das proximale Splitting in (2.53) schreiben als

$$x_{k+1} = \text{prox}_{\alpha_k H}(x_k - \alpha_k \nabla G(x_k)).$$

Wie wir in Beispiel 2.55 gesehen haben ist der Shrinkage-Operator in gewisser Weise kontraktiv. Diese Beobachtung gilt im Allgemeinen für Proximaloperatoren wie folgendes Lemma feststellt.

LEMMA 2.56: Kontraktivität des Proximaloperators.

Sei H eine konvexe, unterhalbstetige Funktion. Dann ist der Proximaloperator prox_H Lipschitz stetig mit Lipschitz-Konstante 1.

Beweis. Wir betrachten zunächst die beiden Punkte, die für $i = 1, 2$ gegeben sind durch $x_i = \text{prox}_H(y_i)$. Dann gilt wegen der Optimalitätsbedingung des Proximaloperators $x_i + p_i = y_i$, für einen Subgradienten $p_i \in \partial H(x_i)$. Subtrahieren wir die beiden Identitäten für $i = 1, 2$, so erhalten wir

$$x_1 - x_2 + p_1 - p_2 \;=\; y_1 - y_2.$$

Multiplizieren wir diese Gleichung von links mit dem Zeilenvektor $(x_1 - x_2)^T \in \mathbb{R}^n$ so gilt mit Hilfe der *Cauchy-Schwarz Ungleichung*:

$$\|x_1 - x_2\|^2 + \langle x_1 - x_2, p_1 - p_2\rangle = \langle x_1 - x_2, y_1 - y_2\rangle \;\leq\; \|y_1 - y_2\| \cdot \|x_1 - x_2\|. \quad (2.55)$$

Wegen Definition 2.50 des Subgradienten können wir festhalten, dass gilt

$$\begin{aligned}\langle p_1, x_1 - x_2\rangle &\;\geq\; H(x_1) - H(x_2)\\ \langle p_2, x_2 - x_1\rangle &\;\geq\; H(x_2) - H(x_1).\end{aligned}$$

Addieren wir diese beiden Ungleichungen, so erhalten wir insgesamt $\langle p_1 - p_2, x_1 - x_2\rangle \geq 0$. Damit können wir die linke Seite der Ungleichung (2.55) weiter abschätzen und erhalten durch Teilen mit dem Wert $||x_1 - x_2||$ auf beiden Seiten schließlich

$$\|x_1 - x_2\| \;\leq\; \|y_1 - y_2\|,$$

was genau die gewünschte Lipschitz-Stetigkeit des Proximaloperators mit Lipschitz-Konstante $L = 1$ bedeutet. □

Analog zum Beweis von Theorem 2.47 können wir auch vorgehen um die Konvergenz des proximalen Splitting Verfahrens im folgenden Theorem zu zeigen.

THEOREM 2.57: Konvergenz des proximalen Splitting-Verfahrens.

Sei $F := \mathbb{R}^n \to \mathbb{R}$ eine nach unten beschränkte Zielfunktion, das heißt, dass die Niveaumenge $K := \{x \in \mathbb{R}^n : F(x) \leq F(x_0)\}$ beschränkt ist. Außerdem lasse sich F als Summe zweier Funktionen schreiben mit $F(x) = G(x) + H(x)$ für eine zweimal stetig differenzierbare Funktion $G\colon \mathbb{R}^n \to \mathbb{R}$ und eine konvexe, unterhalbstetige Funktion $H\colon \mathbb{R}^n \to \mathbb{R}$.
Dann konvergiert das proximale Splitting-Verfahren der Form

$$x_{k+1} \;=\; \text{prox}_{\alpha_k H}(x_k - \alpha_k \nabla G(x_k)).$$

Beweis. Wir nutzen zunächst für einen Subgradienten $p_{k+1} \in \partial H(x_{k+1})$ die explizite Iterationsvorschrift (2.53) für das proximale Splitting

$$\frac{1}{\alpha_k}(x_{k+1} - x_k) + \nabla G(x_k) + p_{k+1} \;=\; 0.$$

Multiplizieren wir diese Gleichung von links mit dem Zeilenvektor $(x_{k+1} - x_k)^T \in \mathbb{R}^n$, dann folgt

$$\frac{1}{\alpha_k}\|x_{k+1} - x_k\|^2 + \langle x_{k+1} - x_k, \nabla G(x_k)\rangle + \langle x_{k+1} - x_k, p_{k+1}\rangle \;=\; 0. \tag{2.56}$$

Da die Funktion G nach Voraussetzung zweimal stetig differenzierbar ist folgt mit einer Taylorapproximation erster Ordnung

$$\langle x_{k+1} - x_k, \nabla G(x_k)\rangle \;=\; G(x_{k+1}) - G(x_k) - r_k.$$

Hierbei lässt sich der Fehlerterm r_k abschätzen durch

$$r_k \;\leq\; \frac{C_k}{2}\|x_{k+1} - x_k\|^2,$$

wobei C_k eine obere Schranke für die Norm der Hesse-Matrix von G in einer Umgebung von x_k mit Radius $\|x_{k+1} - x_k\|$ ist.

Aus Definition 2.50 folgt für den Subgradienten $p_{k+1} \in \partial H(x_{k+1})$ außerdem

$$\langle p_{k+1}, x_{k+1} - x_k\rangle \;\geq\; H(x_{k+1}) - H(x_k).$$

Nutzen wir diese Abschätzungen nun für die Identität (2.56), so erhalten wir insgesamt

$$\begin{aligned}
0 \;&=\; \frac{1}{\alpha_k}\|x_{k+1} - x_k\|^2 + \langle x_{k+1} - x_k, \nabla G(x_k)\rangle + \langle x_{k+1} - x_k, p_{k+1}\rangle \\
&\geq\; \frac{1}{\alpha_k}\|x_{k+1} - x_k\|^2 + G(x_{k+1}) - G(x_k) - \frac{C_k}{2}\|x_{k+1} - x_k\|^2 + H(x_{k+1}) - H(x_k) \\
&=\; \left(\frac{1}{\alpha_k} - \frac{C_k}{2}\right)\|x_{k+1} - x_k\|^2 + F(x_{k+1}) - F(x_k)
\end{aligned}$$

Theoretisch können wir die Schrittweiten $\alpha_k > 0$ so klein wählen, dass $\frac{1}{\alpha_k} - \frac{C_k}{2} > \epsilon$ für ein beliebiges $\epsilon > 0$ gilt. Durch Umstellen sehen wir also, dass gilt

$$F(x_k) \;\geq\; \left(\frac{1}{\alpha_k} - \frac{C_k}{2}\right)\|x_{k+1} - x_k\|^2 + F(x_{k+1}) \;>\; \epsilon \cdot \|x_{k+1} - x_k\|^2 + F(x_{k+1}).$$

Damit ist offensichtlich, dass $F(x_{k+1}) < F(x_k)$ für alle $k \in \mathbb{N}$ gilt. Damit haben wir gezeigt, dass das proximale Splitting-Verfahren die Zielfunktion F in jedem Schritt verkleinert.

Analog zum Beweis von Theorem 2.47 sehen wir ein, dass für die Folge der Iterationsschritte $(x_k)_{k\in\mathbb{N}} \subset K$ gilt und nach dem *Satz von Bolzano-Weierstrass* eine konvergente Teilfolge besitzt, deren Grenzwert in der kompakten Menge K liegen muss.

Es gilt ebenfalls

$$\sum_{k=0}^{\infty}\|x_{k+1} - x_k\|^2 \;<\; \infty.$$

Da die Iterierten auf einer beschränkten Menge bleiben und G zweimal stetig differenzierbar ist, ist die Norm der Hesse-Matrix von G ebenfalls uniform für alle $k \in \mathbb{N}$ beschränkt

für eine Konstante $C > 0$ mit $0 < C_k < C$ und somit kann die Folge der Schrittweiten $(\alpha_k)_{k\in\mathbb{N}}$ ebenfalls uniform nach unten beschränkt werden.

Aus der Konvergenz

$$0 \;=\; \lim_{k\to\infty} \frac{1}{\alpha_k}(x_{k+1} - x_k) \;=\; \lim_{k\to\infty} \nabla G(x_k) + p_{k+1}, \qquad p_{k+1} \in \partial H(x_{k+1})$$

können wir schließlich folgern, dass jeder Häufungspunkt die Optimalitätsbedingung aus Lemma 2.53 erfüllt. □

2.5.2 Primal-Duale Verfahren

In vielen Fällen lässt sich der Proximaloperator einer beliebigen konvexen, unterhalbstetigen Funktion H analytisch nicht gut berechnen und muss gegebenenfalls numerisch approximiert werden. In solchen Fällen ist das proximale Splitting-Verfahren aus Abschnitt 2.5.1 häufig weniger effizient anzuwenden.

In vielen interessanten Fällen hat die Funktion H die Form $H(x) = J(Bx)$, mit einer Matrix $B \in \mathbb{R}^{n\times n}$, die nicht diagonal ist, und einer konvexen, unterhalbstetigen Funktion J, deren Proximaloperator man analytisch gut berechnen kann. Ein typisches Beispiel für solch ein Problem ist eine Variante des Lasso-Problems aus Beispiel 2.49, in dem man folgende Optimierung durchführen will

$$\min_{x\in\mathbb{R}^n} \left\{ F(x) \;=\; \frac{1}{2}\|Ax - b\|^2 + \alpha\|Bx\|_{\ell^1} \right\}.$$

Obwohl für den Spezialfall der Identität $B = I$ bereits in Beispiel 2.55 den Proximaloperator analytisch angeben konnten, lässt sich dieser für $H(x) = \|Bx\|_{\ell^1}$ im Fall beliebiger Matrizen B im Allgemeinen nicht explizit angeben.

Um dieses Problem zu umgehen, ist eine Idee, eine Nebenbedingung mit einer zusätzlichen Variable einzuführen, die wir im Folgenden mit $y \in \mathbb{R}^n$ bezeichnen. Setzen wir nämlich $y = Bx$, so können wir eine modifizierte Zielfunktion

$$\tilde{F}(x, y) \;=\; G(x) + J(y)$$

unter der Nebenbedingung $y = Bx$ minimieren. Um eine Nebenbedingung dieser Form bei der Minimierung von $\tilde{F}$ einfach zu berücksichtigen, können wir die Idee der **Lagrange Multiplikatoren** verwenden (siehe [Data Science 2, Kapitel 7.2]). Hierzu definieren wir zunächst das folgende Lagrange-Funktional

$$L(x, y, z) \;:=\; \tilde{F}(x, y) + \langle z, Bx - y\rangle,$$

wobei der Vektor $z \in \mathbb{R}^n$ die Lagrange Multiplikatoren für die Nebenbedingung $Bx = y$ darstellt.

Man sieht nun ein, dass gilt

$$\inf_{\substack{x,y\in\mathbb{R}^n \\ Bx=y}} \tilde{F}(x, y) \;=\; \inf_{x,y\in\mathbb{R}^n} \sup_{z\in\mathbb{R}^n} L(x, y, z).$$

Dies liegt daran, dass das Supremum über $L(x, y, z)$ unendlich wird, sobald $Bx \neq y$ ist, das heißt, dass das Infimum von $\tilde{F}$ in solchen Fällen sicher nicht angenähert wird. Daher reicht es das Lagrange-Funktional L auf der Menge $\{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : Bx = y\}$ zu betrachten. Auf dieser Menge wird der zweite Term von L jedoch Null und somit stimmt das Lagrange-Funktional schon mit der modifizierten Zielfunktion überein, d.h. es gilt $L(x, y, z) = \tilde{F}(x, y)$.

Wir sehen also, dass wir im Fall der Optimierung mit Nebenbedingungen eigentlich ein **Sattelpunktproblem** der folgenden Form lösen wollen

$$\inf_{x,y\in\mathbb{R}^n} \sup_{z\in\mathbb{R}^n} (G(x) + J(y) + \langle z, Bx - y\rangle).$$

Nehmen wir zunächst an, dass wir das Infimum und das Supremum vertauschen können (was möglich ist, falls ein Sattelpunkt existiert) und nutzen die Relationen $\inf_{x,y} -E(x, y) = -\sup_{x,y} E(x, y)$ und $\sup_z -F(z) = -\inf_z F(z)$ dann lösen wir

$$\begin{aligned}
&\sup_{z\in\mathbb{R}^n} \inf_{x,y\in\mathbb{R}^n} (G(x) + J(y) + \langle z, Bx - y\rangle) \\
=\ &\sup_{z\in\mathbb{R}^n} \inf_{x,y\in\mathbb{R}^n} -(\langle z, y\rangle - J(y) + \langle -B^T z, x\rangle - G(x)) \\
=\ &-\inf_{z\in\mathbb{R}^n} \sup_{x,y\in\mathbb{R}^n} (\langle z, y\rangle - J(y) + \langle -B^T z, x\rangle - G(x)) \\
=\ &-\inf_{z\in\mathbb{R}^n} (\sup_{y\in\mathbb{R}^n} (\langle z, y\rangle - J(y)) + \sup_{x\in\mathbb{R}^n} (\langle -B^T z, x\rangle - G(x))).
\end{aligned}$$

Bevor wir eine sogenannte primal-duale Formulierungen des ursprünglichen Problems herleiten, werden wir zunächst die Konvex-Konjugierte definieren.

DEFINITION 2.58: Legendre-Fenchel-Transformation.

Sei $J\colon \mathbb{R}^n \to \mathbb{R}$ eine strikt konvexe Funktion. Dann ist die **Legendre-Fenchel-Transformation** J^* von J (auch Konvex-Konjugierte genannt) definiert als

$$J^*(z) \ := \sup_{y\in\mathbb{R}^n} (\langle z, y\rangle - J(y)).$$

Mit Definition 2.58 der Konvex-Konjugierten lässt sich das ursprünglichen Problem äquivalent umformulieren zu

$$\inf_{z\in\mathbb{R}^n} \sup_{x\in\mathbb{R}^n} (J^*(z) + \langle -B^T z, x\rangle - G(x)), \tag{2.57}$$

wobei wir bemerken, dass wir in dieser Formulierung die explizite Abhängigkeit von der Hilfsvariable $y = Bx$ vermeiden konnten.

Die neue Formulierung des Sattelproblems in (2.57) dient als Grundlage vieler primal-dualer Verfahren in der Numerik. Beispielsweise können wir zur Approximation eines Sattelpunkts abwechselnd einen Abstiegsschritt mittels Proximaloperator bezüglich der Variable $z \in \mathbb{R}^n$ und einen Gradientenaufstiegsschritt bezüglich der Variable $x \in \mathbb{R}^n$

durchführen. Dies führt zu einer Variante des sogenannten **Uzawa-Verfahrens** mit Schrittweiten $\tau_k, \sigma_k \in \mathbb{R}^+$:

$$\begin{aligned} z_{k+1} &= \operatorname{prox}_{\tau_k J^*}(z_k + \tau_k B x_k), \\ x_{k+1} &= x_k - \sigma_k(\nabla G(x_k) + B^T z_k). \end{aligned} \tag{2.58}$$

In diesem Fall müssen wir die Matrix B und B^T nur einmal mit einem Vektor multiplizieren in jedem Schritt, sowie den Proximaloperator der Konvex-Konjugierten J^* ausrechnen. Letzteres ist genau dann einfach, wenn wir den Proximaloperator von J effizient ausrechnen können, wie folgende Bemerkung festhält.

BEMERKUNG 2.59 (Moreau-Zerlegung für Proximaloperatoren). Sei $J\colon \mathbb{R}^n \to \mathbb{R}$ eine konvexe, unterhalbstetige Funktion und $J^*\colon \mathbb{R}^n \to \mathbb{R}$ die zugehörige Konvex-Konjugierte von J. Sei außerdem $\tau > 0$ ein positiver Parameter.
Dann kann man die folgende Identität, genannt Moreau-Zerlegung, für die Proximaloperatoren von J und J^* zeigen:

$$x = \operatorname{prox}_{\tau J^*}(x) + \tau \operatorname{prox}_{J/\tau}(x/\tau).$$

△

Somit können wir den Proximaloperator $\operatorname{prox}_{\tau_k J^*}$ der Konvex-Konjugierten in (2.58) berechnen als:

$$\operatorname{prox}_{\tau_k J^*}(z_k + \tau_k B x_k) = z_k + \tau_k B x_k - \tau_k \operatorname{prox}_{\frac{1}{\tau_k} J}\left(\frac{z_k + \tau_k B x_k}{\tau_k}\right)$$

Kapitel 3

Numerische Lösungsverfahren für Anfangswertprobleme

Differentialgleichungen spielen eine wichtige Rolle in der Modellierung vieler naturwissenschaftlicher Phänomene. Insbesondere in der Physik und den Ingenieurswissenschaften ist das Verständnis und das Lösen von Differentialgleichungen essentiell, da sich viele physikalische Gesetze und Zusammenhänge durch diese beschreiben lassen. Leider existieren für viele Differentialgleichungen keine geschlossen darstellbaren analytischen Lösungen, so dass man Lösungen mit Hilfe numerischer Verfahren approximieren muss. In diesem und nächstem Kapitel widmen wir uns daher der numerischen Lösung von verschiedenen Problemen, die Differentialgleichungen beinhalten. Dies sind insbesondere Anfangswert- und Randwertprobleme.

Mathematisch gesehen ist eine Differentialgleichung eine Gleichung, in der eine unbekannte Funktion und ihre Ableitungen auftreten. Hierbei werden folgende Kriterien zur Unterscheidung von Differentialgleichungen genutzt.

1. Die größte auftretende Ableitung der unbekannten Funktion bestimmt die **Ordnung der Differentialgleichung**.

2. Je nachdem ob Ableitungen der unbekannten Funktion bezüglich einer einzigen Variablen oder unterschiedlicher Variablen auftreten sprechen wir von einer **gewöhnlichen Differentialgleichung** oder einer **partiellen Differentialgleichung**.

3. Werden gleich mehrere solche Funktionen durch mehrere Gleichungen beschrieben, so spricht man von einem **Differentialgleichungssystem**.

Wir werden in den folgenden Abschnitten zunächst die wichtigsten theoretischen Erkenntnisse für gewöhnliche Differentialgleichungen wiederholen und insbesondere Existenz- und Eindeutigkeitsaussagen diskutieren. Diese werden für die spätere Konstruktion und Analyse von numerischen Lösungsverfahren hilfreich sein.

3.1 Theorie für Anfangswertprobleme gewöhnlicher Differentialgleichungen

Im Folgenden beschäftigen wir uns zunächst mit Anfangswertproblemen für gewöhnliche Differentialgleichungen, im Englischen *„Ordinary differential equations (ODE)“* genannt.

DEFINITION 3.1: Anfangswertproblem.

Ein mathematischen Problem, bei dem wir eine Lösung $u\colon \mathbb{R}^+ \to \mathbb{R}^n$ eines gewöhnlichen Differentialgleichungssystems 1. Ordnung mit gegebener Startwertbedingung der folgenden allgemeinen Form suchen

$$\begin{aligned} \frac{du}{dt} \;=\; u'(t) \;&=\; F(t, u(t)), \\ u(0) \;&=\; u_0 \in \mathbb{R}^n, \end{aligned} \tag{3.1}$$

wobei $F : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}^n$ eine stetige Funktion ist, nennen wir ein **Anfangswertproblem**.

Anfangswertprobleme der Form (3.1) treten häufig in der Modellierung von Zuständen auf, die sich über die Zeit ändern. Nimmt man an, dass der zu modellierende Zustand zum Zeitpunkt $t > 0$ durch die Funktion $u\colon \mathbb{R}^+ \to \mathbb{R}^n$ beschrieben wird, möchte man verstehen wie sich dieser in hinreichend kleiner Zeit durch die Wirkung der Funktion F ändern wird. Das heißt wir schreiben

$$u(t + \tau) \;\approx\; u(t) + \tau \cdot F(t, u(t)).$$

Im Grenzwert für immer kleinere Zeitschrittweiten $\tau \to 0$ erhalten wir dann die Differentialgleichung $u'(t) = F(t, u(t))$.

Solche Differentialgleichungen lassen sich beispielsweise in der Physik herleiten, wie folgenden Beispiel zeigt.

BEISPIEL 3.2: Newtonsche Bewegungsgleichung.

Wir betrachten in diesem Beispiel die Bewegung eines Teilchens mit Hilfe des zweiten Newtonschen Gesetzes. Beschreibt $x(t) \in \mathbb{R}^3$ den Ort des Teilchens zur Zeit $t > 0$ im dreidimensionalen Raum, $v(t) \in \mathbb{R}^3$ seine Geschwindigkeit, $a(t)$ seine Beschleunigung und F ein Kraftfeld in dem sich das Teilchen bewegt, so sind folgende Zusammenhänge aus dem physikalischen Mechanik bekannt:

- Geschwindigkeit ist Änderung des Orts pro Zeit, d.h., $v(t) = \dot{x}(t) = \frac{dx}{dt}(t)$
- Beschleunigung ist Änderung der Geschwindigkeit pro Zeit, d.h. es gilt $a(t) = \dot{v}(t) = \frac{dv}{dt}(t)$
- Kraft ist Masse mal Beschleunigung, d.h., $m \cdot a(t) = F(x(t), v(t), t)$

Diese Gesetze können wir auch als Differentialgleichungen für den Ort x und die Geschwindigkeit v des Teilchens (oder dessen Impuls $p(t) = m \cdot v(t)$) interpretieren, wenn wir die Beschleunigung a eliminieren. Es gilt dann

$$\frac{d}{dt}(x(t), m \cdot v(t)) = (v(t), F(x(t), v(t), t)).$$

Kennen wir den Anfangsort $x(0) \in \mathbb{R}^3$ des Teilchens und dessen Anfangsgeschwindigkeit $v(0) \in \mathbb{R}^3$, so haben wir ein Anfangswertproblem für ein System von Differentialgleichungen mit insgesamt sechs Gleichungen für sechs Unbekannte.
In großen Systemen mit $n \in \mathbb{N}$ Teilchen hat man dann Gleichungen für jede Position $x_1, \ldots, x_n \in \mathbb{R}^3$ und die respektiven Geschwindigkeiten $v_1, \ldots, v_n \in \mathbb{R}^3$ und sucht Lösungen für das folgende System von Differentialgleichungen:

$$\frac{d}{dt}(x_i(t), v_i(t)) = (v_i(t), F(x_1(t), v_1(t), \ldots, x_n(t), v_n(t), t)), \quad i = 1, \ldots, n.$$

Zur Beschreibung realer Vorgänge mit vielen Teilchen (Moleküle, Zellen, Tiere in Herden, Fußgänger, Autos ...) erhält man also schnell beliebig komplexe Systeme von Differentialgleichungen.

3.1.1 Existenz und Eindeutigkeit von Lösungen

Um die in dieser Vorlesung diskutierten Verfahren zur numerischen Lösung von gewöhnlichen Differentialgleichungen besser zu verstehen müssen wir im Folgenden zunächst einige theoretische Grundlagen von gewöhnlichen Differentialgleichungen zusammenfassen. Wir beginnen mit einer allgemeinen Theorie der Existenz und Eindeutigkeit von Lösungen solcher Differentialgleichungen. Die wesentliche Grundlage hierfür wird eine Umformulierung der gewöhnlichen Differentialgleichung in eine Fixpunktform sein, so dass wir dann einfach einen passenden Fixpunktsatz anwenden können.

Nehmen wir zunächst an, dass $u \in C^1([0, T])$ eine Lösung des Anfangswertproblems (3.1) sei. Dann gilt nach dem Hauptsatz der Differential- und Integralrechnung auch folgende Gleichung

$$u(t) \;=\; u_0 + \int_0^t F(s, u(s))\,\mathrm{d}s, \qquad 0 \le t \le T.$$

Diese Darstellung können wir als Fixpunktgleichung $u = \mathcal{F}(u)$ in einem Banachraum interpretieren. Um Existenz und Eindeutigkeit eines Fixpunktes u und somit einer Lösung des Anfangswertproblems zu zeigen, lassen sich nun zwei unterschiedliche Arten von Fixpunktsätzen anwenden.

Die erste Art (*Satz von Schauder* oder andere Varianten) basiert auf Kompaktheit, d.h. wenn man den Operator $\mathcal{F}$ auf Funktionen u anwendet, liegt das Bild in einer topologischen Menge mit einem Fixpunkt des Operators. In unserem Fall erhält man

die Kompaktheit aus dem *Satz von Arzela–Ascoli*, denn man kann zeigen, dass für eine beschränkte Folge von Funktionen $(u_n)_{n\in\mathbb{N}}$ die Folge $(\mathcal{F}(u_n))_{n\in\mathbb{N}}$ immer eine konvergente Teilfolge besitzt. Dies liefert den sogenannten **Satz von Peano**, der die Existenz einer Lösung für eine stetige Funktion F garantiert. Da man hier jedoch recht abstrakt und mittels Teilfolgen argumentiert, lässt sich mit dieser Methode leider nicht die Eindeutigkeit eines Fixpunkts nachweisen.

Die zweite Art um Existenz und Eindeutigkeit eines Fixpunkts zu zeigen basiert eigentlich immer auf dem *Banachschen Fixpunktsatz* (siehe [Numerik 1, Kapitel 4.1]). Diese Herangehensweise wollen wir im Folgenden etwas näher diskutieren. Dazu beachten wir, dass falls die Funktion F im zweiten Argument Lipschitz-stetig ist (mit Lipschitz-Konstante $L > 0$), folgendes gilt

$$\begin{aligned}
\|\mathcal{F}(u_1) - \mathcal{F}(u_2)\|_\infty &= \max_{0\le t\le T} \left| \int_0^t F(s, u_1(s)) - F(s, u_2(s))\,\mathrm{d}s \right| \\
&\le \max_{0\le t\le T} \int_0^t |F(s, u_1(s)) - F(s, u_2(s))|\,\mathrm{d}s \\
&\le L \cdot \max_{0\le t\le T} \int_0^t |u_1(s) - u_2(s)|\,\mathrm{d}s \\
&\le LT \cdot \max_{0\le s\le T} |u_1(s) - u_2(s)| \;=\; LT \cdot \|u_1 - u_2\|_\infty.
\end{aligned}$$

Wir erkennen an dieser Abschätzung, dass die Abbildung $\mathcal{F} : C^1([0,T]) \to C^1([0,T])$ kontraktiv ist, wenn $T > 0$ klein genug gewählt wird, so dass $LT < 1$ gilt. Mit Hilfe der Kontraktivität von $\mathcal{F}$ liefert der Banach'sche Fixpunktsatz die Existenz und Eindeutigkeit eines Fixpunkts u (und damit einer Lösung der Differentialgleichung) in $C^1([0,T])$ für T hinreichend klein. Dies ist die erste Version des **Satzes von Picard-Lindelöf,** die uns die Existenz und Eindeutigkeit für kleine Zeiten $T > 0$ liefert.

In unserem Fall können wir jedoch ein noch besseres Resultat erreichen, in dem wir einfach die Norm in unserer Abschätzung passend wählen. Die Idee dazu liefert zunächst eine einfache Differentialgleichung im folgenden Beispiel.

BEISPIEL 3.3: Einfache Differentialgleichung mit konstantem Faktor.

Wir beschäftigen uns in diesem Beispiel mit einer einfachen gewöhnlichen Differentialgleichung mit einer Konstanten $L > 0$ der Form

$$u'(t) \;=\; L \cdot u(t).$$

Nehmen wir $u\colon \mathbb{R}_+ \to \mathbb{R}$ als positive Lösung an, dann folgt mit der Kettenregel der Differentialgleichung

$$\frac{d}{dt}(\log u(t)) \;=\; \frac{u'(t)}{u(t)} \;=\; L.$$

Dies können wir mit Hilfe des Hauptsatzes der Differential- und Integralrechnung schreiben als

$$\log u(t) - \log u(0) \;=\; \log\left(\frac{u(t)}{u_0}\right) \;=\; L \cdot t.$$

Wenden wir auf beide Seiten die Exponentialfunktion an, erhalten wir die folgende analytische Lösung der Differentialgleichung:

$$u(t) \;=\; u_0 \cdot e^{Lt}.$$

Wählen wir nun den Vorfaktor L in Beispiel 3.3 als die Lipschitzkonstante im zweiten Argument der Funktion F, dann erkennen wir, dass wir ein exponentielles Wachstum mit $e^{L \cdot t}$ erwarten müssen.

Diese Beobachtung ist auch im Allgemeinen gültig, wie das folgende Lemma zeigt.

LEMMA 3.4: Lemma von Gronwall.

Sei $v(t)$ eine nichtnegative, stetige Funktion, die die folgende Ungleichung

$$v(t) \;\leq\; a + \int_0^t b \cdot v(s)\,\mathrm{d}s, \quad \forall\, 0 \leq t \leq T, \tag{3.2}$$

mit $a > 0$ und $b \in \mathbb{R}$ erfüllt.
Dann lässt sich die Funktion v nach oben abschätzen durch

$$v(t) \;\leq\; a \cdot e^{bt}, \quad \forall\, 0 \leq t \leq T.$$

Beweis. Wir definieren zunächst die Hilfsfunktion $w(t) := e^{-bt} \cdot v(t) - a$. Damit können wir v auch schreiben als $v(t) = e^{bt} \cdot (w(t) + a)$. Da die Ungleichung (3.2) nach Voraussetzung gilt können wir folgende Abschätzung machen:

$$\begin{aligned}
w(t) &= e^{-bt} \cdot v(t) - a \\
&\leq e^{-bt} \cdot a + e^{-bt} \cdot \int_0^t b \cdot v(s)\,\mathrm{d}s - a \\
&= a \cdot (e^{-bt} - 1) + \int_0^t e^{-bt} \cdot b \cdot v(s)\,\mathrm{d}s \\
&= a \cdot (e^{-bt} - 1) + \int_0^t e^{b(s-t)} \cdot b \cdot (w(s) + a)\,\mathrm{d}s \\
&= a \cdot (e^{-bt} - 1) + ab \cdot e^{-bt} \cdot \int_0^t e^{bs}\,\mathrm{d}s + b \cdot \int_0^t e^{b(s-t)} \cdot w(s)\,\mathrm{d}s \\
&\leq a \cdot (e^{-bt} - 1) + ab \cdot e^{-bt} \cdot \left[\frac{1}{b} e^{bs}\right]_0^t + b \cdot \int_0^t \underbrace{e^{b(s-s)}}_{=1} \cdot w(s)\,\mathrm{d}s \\
&= \underbrace{a \cdot (e^{-bt} - 1) + a \cdot (1 - e^{-bt})}_{=0} + b \int_0^t w(s)\,\mathrm{d}s \\
&= \int_0^t b \cdot w(s)\,\mathrm{d}s.
\end{aligned}$$

Für die letzte Ungleichung haben wir verwendet, dass $e^{-bs} \geq e^{-bt}$ gilt für alle $0 \leq s \leq t$. Aus dieser Ungleichung können wir für $t = 0$ folgern, dass $w(0) \leq 0$ und somit $v(0) \leq a \cdot e^0$ gilt. Somit gilt die Abschätzung des Lemmas bereits in diesem Fall.

Da v als stetige Funktion vorausgesetzt wurde ist w ebenfalls stetig und somit existiert ein maximaler Zeitpunkt $T \geq T_0 > 0$ für den $w(t) \leq 0$ für alle $t \leq T_0$ gilt. Ist $T_0 = T$, so haben wir die Abschätzung des Lemmas bereits gezeigt, da aus $w(t) \leq 0$ für alle $0 \leq t \leq T$ durch Umstellen schon folgt $v(t) \leq a \cdot e^{bt}$.

Nehmen wir nun an, dass $0 < T_0 < T$ gilt, so gibt es ein hinreichend kleines Zeitintervall $(T_0, T_0 + \delta)$, in dem w positiv ist. Wegen der Positivität von w folgt dann für ein beliebiges $t \in (T_0, T_0 + \delta)$ die Ungleichung

$$w(t) \;\leq\; \int_{T_0}^{t} b \cdot w(s)\,\mathrm{d}s \;\leq\; \delta \cdot \max_{T_0 \leq s \leq T_0+\delta} b \cdot w(s)$$

und somit auch

$$\max_{T_0 \leq t \leq T_0+\delta} w(t) \;\leq\; \delta \cdot \max_{T_0 \leq s \leq T_0+\delta} b \cdot w(s) \;=\; \delta b \cdot \max_{T_0 \leq s \leq T_0+\delta} w(s).$$

Für $\delta b < 1$ ist dies aber ein Widerspruch zur Positivität von w. Also muss $w(t) \leq 0$ und damit $u(t) \leq ae^{bt}$ für alle $0 \leq t \leq T$ gelten. □

Abstrakt gesehen liefert das obige Lemma von Gronwall eine Stabilitätsaussage für die Differentialgleichung, denn es besagt, dass die Norm der Lösung in endlicher Zeit nicht beliebig wachsen kann. Betrachten wir nämlich eine Differentialgleichung mit einer Funktion F, die Lipschitz-stetig im zweiten Argument ist, so folgt mit der Wahl der Funktion $v(t) = |u(t) - u_0|$ nämlich

$$\begin{aligned}
v(t) \;=\; |u(t) - u_0| \;&=\; \left| \int_0^t F(s, u(s)\,\mathrm{d}s \right| \;\leq\; \int_0^t |F(s, u(s)|\,\mathrm{d}s \\
&= \int_0^t |F(s, u(s) - F(s, u_0) + F(s, u_0)|\,\mathrm{d}s \\
&\leq \int_0^t |F(s, u(s)) - F(s, u_0)| + |F(s, u_0)|\,\mathrm{d}s \\
&\leq \int_0^t |F(s, u(s)) - F(s, u_0)|\,\mathrm{d}s + \underbrace{T \cdot \max_{0 \leq t \leq T} |F(t, u_0)|}_{=:\, a} \\
&\leq L \cdot \int_0^t |u(s) - u_0|\,\mathrm{d}s + a \;=\; b \cdot \int_0^t v(s)\,\mathrm{d}s + a.
\end{aligned}$$

Wir sehen also, dass die Voraussetzungen für Lemma 3.4 erfüllt sind für die Funktion $v(t)$ mit $b := L$ und $a := T \cdot \max_{0 \leq t \leq T} |F(t, u_0)|$. Somit gilt also nach dem Lemma von Gronwall die Abschätzung

$$|u(t) - u_0| \;\leq\; a \cdot e^{bt} \;=\; T \cdot \max_{0 \leq t \leq T} |F(t, u_0)| \cdot e^{Lt}.$$

Verwenden wir die umgekehrte Dreiecksungleichung $|u(t)| - |u_0| \leq |u(t) - u_0|$ sehen wir, dass $u(t)$ beschränkt ist und höchstens wie e^{Lt} wächst:

$$|u(t)| \;\leq\; |u(t) - u_0| + |u_0| \;\leq\; T \cdot \max_{0 \leq t \leq T} |F(t, u_0)| \cdot e^{Lt} + |u_0|.$$

Dies führt zu der Idee, die folgende gewichtete Norm zu betrachten:

$$\|u\|_{\infty,L} \;:=\; \max_{0 \leq t \leq T} e^{-Lt} \cdot |u(t)|.$$

Da e^{-Lt} nach oben durch Eins und nach unten durch $e^{-LT} \geq 0$ beschränkt ist, ist dies eine äquivalente Norm im Raum der stetig differenzierbaren Funktionen $C^1([0,T])$.

Wir wiederholen also unsere Abschätzung des Fixpunktoperators in dieser konstruierten Norm und erhalten somit

$$\begin{aligned}
\|\mathcal{F}(u_1) - \mathcal{F}(u_2)\|_{L,\infty} &= \max_{0 \leq t \leq T} e^{-Lt} \cdot \left| \int_0^t F(s, u_1(s)) - F(s, u_2(s))\, \mathrm{d}s \right| \\
&\leq \max_{0 \leq t \leq T} \int_0^t e^{L(s-t)} \cdot e^{-Ls} \cdot |F(s, u_1(s)) - F(s, u_2(s))|\, \mathrm{d}s \\
&\leq \int_0^T e^{-L\tau}\, \mathrm{d}\tau \cdot \max_{0 \leq s \leq T} e^{-Ls} \cdot L \cdot |u_1(s) - u_2(s)| \\
&= (1 - e^{-LT}) \cdot \|u_1 - u_2\|_{L,\infty}.
\end{aligned}$$

Der Operator ist nun also kontraktiv bezüglich der gewählten Norm für beliebiges $T > 0$, da stets $1 - e^{-LT} < 1$ gilt.

Wir haben mit Hilfe des Banachschen Fixpunktsatzes also die folgende Variante des Satzes von Picard-Lindelöf bewiesen.

THEOREM 3.5: Satz von Picard-Lindelöf.

Sei F eine stetige Funktion, die Lipschitz-stetig bezüglich der zweiten Variable ist.
Dann besitzt das Anfangswertproblem (3.1) eine eindeutige Lösung in $C^1([0,T])$.

Wir werden sehen, dass wir auch bei numerischen Verfahren zur Lösung von Anfangswertproblemen ähnliche Aussagen und insbesondere eine Variante des Lemma von Gronwall in Lemma 3.4 benötigen werden, um die Stabilität dieser Verfahren garantieren zu können.

3.1.2 Analytische Lösungsverfahren

Wir betrachten im Folgenden einige spezielle Fälle von gewöhnlichen Differentialgleichungen in denen wir analytisch eine geschlossene Form der Lösung berechnen können.

Andererseits gibt es viele gewöhnliche Differentialgleichungen für die sich analytisch keine Lösung angeben lässt, wie zum Beispiel die folgende simple, nichtlineare Gleichung:

$$u'(t) \;=\; t^2 + u^2(t).$$

Ein weiteres bekanntes Phändomen, dass durch ein Differentialgleichungssystem beschrieben ist, jedoch nicht analytisch lösbar ist, ist das das *N-Körper-Problem* [N-Körper], welches auch schon in Beispiel 3.2 durch die geeignete Wahl des Kraftfeldes beschrieben wird. Aus diesem Grund wollen wir in dieser Vorlesung numerische Lösungsmethoden für gewöhnliche Differentialgleichungen diskutieren und untersuchen.

Hierbei werden wir nur kurz die verschiedenen Herangehensweisen zur Herleitung einer Lösung skizzieren. Für eine formale Beschreibung dieser Ansätze verweisen wir auf beispielsweise auf [Data Science 2, Kapitel 8].

Lineare Differentialgleichungen

Eine Klasse von gewöhnlichen Differentialgleichungen, die in folgender Definition erklärt sind, stellen sich als gut verständlich und analytisch lösbar heraus.

DEFINITION 3.6: Lineare Differentialgleichung.

Wir nennen ein gewöhnliches Differentialgleichungssystem n-ter Ordnung **linear**, wenn es sich in folgender Form schreiben lässt:

$$\sum_{i=0}^{n} a_i(t)u^{(i)}(t) \;=\; a_0(t)\cdot u(t) + a_1(t)\cdot u'(t) + \ldots + a_n(t)\cdot u^{(n)}(t) \;=\; b(t).$$

Hierbei sind die $a_i\colon \mathbb{R}_+ \to \mathbb{R}^{n\times n}, i = 0,\ldots,n$ und $b\colon \mathbb{R}_+ \to \mathbb{R}^n$ stetige Funktionen, die nicht von u abhängen, und $u^{(i)}$ bezeichnet die i-te Ableitung der unbekannten Funktion u.

Ein wichtiger Spezialfall von Definition 3.6, der uns auch kanonische Beispiele für numerische Verfahren liefert, ist ein lineares Differentialgleichungssystem 1. Ordnung mit konstanten Koeffizienten der Form

$$u'(t) \;=\; A\cdot u(t) + b(t), \tag{3.3}$$

mit einer gegebenen Matrix $A \in \mathbb{R}^{n\times n}$.

Matrixexponential

Wir betrachten zunächst den homogenen Fall für $b \equiv \mathbf{0}$ des linearen Differentialgleichungssystem in (3.3). Ist die Matrix A diagonalisierbar durch $A = B^{-1}\cdot D\cdot B$ mit einer Diagonalmatrix $D \in \mathbb{R}^{n\times n}$, so können wir analog eine unbekannte Hilfsvariable $v(t) = B\cdot u(t)$ betrachten. Damit können wir nun das lineare Differentialgleichungssystem (3.3) schreiben als

$$\begin{aligned} u'(t) \;&=\; A\cdot u(t) \;=\; B^{-1}\cdot D\cdot B\cdot u(t) \;=\; B^{-1}\cdot D\cdot v(t) \\ \Rightarrow\; v'(t) \;&=\; (B\cdot u(t))' \;=\; B\cdot u'(t) \;=\; D\cdot v(t). \end{aligned}$$

Somit können wir also äquivalent das folgende simple Differentialgleichungssystem lösen:

$$v'(t) \;=\; D\cdot v(t).$$

Da D eine Diagonalmatrix ist, sind die Gleichungen entkoppelt und das bedeutet, dass wir für jede Zeile eine gewöhnliche Differentialgleichung mit konstantem Vorfaktor der Form $v_i'(t) = D_{ii}v_i(t)$ erhalten. Für diese können wir analog zu Beispiel 3.3 eine explizite Lösung angeben als

$$v_i(t) \;=\; v_i(0) \cdot e^{D_{ii} \cdot t}.$$

Die unbekannte Lösung u der urspünglichen Differentialgleichung (3.3) erhalten wir anschließend durch

$$u(t) \;=\; B^{-1} \cdot v(t).$$

Zur Vereinfachung lässt sich dieser Lösungsansatz mit Hilfe des sogenannten Matrixexponentials schreiben.

DEFINITION 3.7: Matrixexponential.

Sei $A \in \mathbb{R}^{n\times n}$ eine quadratische Matrix. Das **Matrixexponential** $e^A \in \mathbb{R}^{n\times n}$ ist dann definiert durch die folgende Potenzreihe

$$e^A \;:=\; \sum_{k=0}^{\infty} \frac{A^k}{k!} \;=\; I_n + A + \frac{A^2}{2} + \dots$$

Für den Fall einer Diagonalmatrix $D \in \mathbb{R}^{n\times n}$ und einem Skalar $t \in R$ ist das Matrixexponential e^{Dt} in Definition 3.7 eine Diagonalmatrix mit Einträgen $(e^{Dt})_{i,i} = e^{D_{i,i}\cdot t}$ für $i = 1, \dots, n$. Mit dieser Erkenntnis lässt sich die Lösung der Differentialgleichung (3.3) kompakt angeben als:

$$u(t) \;=\; B^{-1} \cdot v(t) \;=\; B^{-1} \cdot e^{Dt} \cdot v(0) \;=\; B^{-1} \cdot e^{Dt} \cdot B \cdot u(0) \;=:\; e^{At} \cdot u(0). \qquad (3.4)$$

Wir bekommen durch Diagonalisieren der Matrix A also eine simple Form des Matrixexponentials, die uns direkt eine Lösung des Anfangswertproblems liefert. Im Fall einer nicht diagonalisierbaren Matrix ist ein ähnliches Vorgehen über die Jordan'sche Normalform möglich, was an dieser Stelle aber über den Rahmen dieser Vorlesung hinaus gehen würde.

Variation der Konstanten

Im inhomogenen Fall $b \neq \mathbf{0}$ eines linearen Differentialgleichungssystems erster Ordnung in (3.3) lässt sich die sogenannte Variation der Konstanten benutzen um eine Lösung zu berechnen. Die Idee hierbei ist es einen Produktansatz der unbekannten Lösung u zu betrachten mit

$$u(t) \;=\; e^{At} \cdot w(t).$$

Anstatt der Konstanten $u(0) = u_0 \in \mathbb{R}^n$ in der homogenen Lösung in (3.4) betrachten wir also jetzt eine zeitabhängige Funktion $w(t)$.

Mit Hilfe der Produktregel für Differentiation und der Eigenschaften des Matrixexponentials lässt sich somit die Ableitung der Funktion u schreiben als

$$u'(t) \;=\; \frac{d}{dt}(e^{At}\cdot w(t)) \;=\; A\cdot \underbrace{e^{At}\cdot w(t)}_{=u(t)} + e^{At}\cdot w'(t) \;=\; A\cdot u(t) + e^{At}\cdot w'(t),$$

Vergleichen wir nun diesen Ausdruck der Ableitung der unbekannten Funktion u mit rechten Seite der Differentialgleichung (3.3), so sehen wir ein, dass $b(t) = e^{At}\cdot w'(t)$ gelten muss. Durch Umstellen sehen wir ein, dass für die Ableitung der unbekannten Hilfsfunktion $w'(t) = e^{-At}\cdot b(t)$ gilt. Integrieren wir diesen Ausdruck, so lässt sich die unbekannte Lösung u durch die folgende Identität berechnen:

$$u(t) \;=\; e^{At}\cdot w(t) \;=\; e^{At}\cdot\left(c + \int_0^t w'(s)\,\mathrm{d}s\right) \;=\; e^{At}\cdot\left(c + \int_0^t e^{-As}\cdot b(s)\,\mathrm{d}s\right).$$

Hierbei lässt sich das Matrixexponential $e^{\pm At} \in \mathbb{R}^{n\times n}$ wie im homogenen Fall beschrieben bestimmen und die Integrationskonstante $c \in \mathbb{R}$ erhält man durch Anwendung der Anfangswertbedingung $u(0) = u_0 \in \mathbb{R}^n$.

Trennung der Variablen

Ein weiterer Spezialfall für die analytische Bestimmung von Lösungen gewöhnlicher Differentialgleichungen in Dimension $n = 1$ sind sogenannte separable Gleichungen der folgenden Form:

$$u'(t) \;=\; g(u(t))\cdot h(t),$$

wobei $g, h\colon \mathbb{R}_+ \to \mathbb{R}$ zwei stetige Funktionen sind. Nehmen wir an, dass $g(u(t))$ nicht Null wird, so können wir durch diesen Term teilen und erhalten

$$\frac{u'(t)}{g(u(t))} \;=\; h(t).$$

Ist G eine Stammfunktion von $\frac{1}{g}$ und H eine Stammfunktion von h, so schreiben wir diese Gleichung mit Hilfe der Kettenregel der Differentiation als

$$\frac{d}{dt}G(u(t)) \;=\; G'(u(t))\cdot u'(t) \;=\; H'(t).$$

Diese Gleichung können wir nun aufintegrieren zu $G(u(t)) = H(t) + c$. Die Integrationskonstante $c \in \mathbb{R}$ können wir anschließend aus dem Anfangswert mit $c = G(u_0) - H(0)$ berechnen.

Setzen wir weiter voraus, dass die Stammfunktion G von g^{-1} invertierbar ist, lässt sich die unbekannte Lösung u nun wie folgt berechnen:

$$u(t) \;=\; G^{-1}(H(t) - H(0) + G(u_0)).$$

Gradientenfluss

Zum Abschluss betrachten wir noch eine Klasse von gewöhnlichen Differentialgleichungen, die wir aus dem Gradientenabstiegsverfahren der numerischen Optimierung in Abschnitt 2.2.1 erhalten. Hierbei interpretieren wir die Iterationsschritte $x_k \in \mathbb{R}^n$ als Wert einer Funktion $u\colon \mathbb{R}_+ \to \mathbb{R}^n$ zur Zeit $t > 0$ und nehmen an, dass die Zielfunktion F stetig differenzierbar ist. Damit können wir das Gradientenabstiegsverfahren schreiben als

$$u(t + \alpha_k) \;=\; u(t) - \alpha_k \nabla F(u(t)), \qquad \alpha_k > 0.$$

Lassen wir nun die Schrittweiten $\alpha_k > 0$ gegen Null gehen, so erhalten wir im Grenzwert eine als Gradientenfluss bekannte Differentialgleichung der Form

$$u'(t) \;=\; -\nabla F(u(t)).$$

Hierbei bleibt die Abstiegseigenschaft erhalten, denn es gilt

$$\frac{d}{dt} F(u(t)) \;=\; \langle \nabla F(u(t)), u'(t) \rangle \;=\; -\|\nabla F(u(t))\|^2 \;=\; -\|u'(t)\|^2 \;\leq\; 0.$$

Ist die Zielfunktion F zusätzlich konvex, d.h. wir wissen

$$\langle v - u, \nabla F(u) \rangle \;\leq\; F(v) - F(u), \qquad \forall\, v, u \in \mathbb{R}^n,$$

dann gilt für einen Minimierer $u^*(t) \equiv u^* \in \mathbb{R}^n$ von F (der gleichzeitig eine stationäre Lösung des Gradientenflusses darstellt) sogar

$$\begin{aligned}
\frac{d}{dt}\left(\frac{1}{2}\|u(t) - u^*(t)\|^2\right) \;&=\; \langle u(t) - u^*(t), u'(t) - u^{*\prime}(t) \rangle \\
&=\; -\langle u(t) - u^*(t), \nabla F(u(t)) - \underbrace{\nabla F(u^*(t))}_{=\mathbf{0}} \rangle \\
&=\; \langle u^*(t) - u(t), \nabla F(u(t)) \rangle \\
&\leq\; F(u^*(t)) - F(u(t)) \;\leq\; 0.
\end{aligned}$$

Hieran erkennen wir, dass der Abstand der Lösung u zum Minimierer u^* monoton fällt in der Zeit und wir die Norm der unbekannten Lösung sogar gleichmäßig durch die Anfangswertbedingung $u(0) = u_0 \in \mathbb{R}^n$ beschränken können.

3.2 Einschrittverfahren für Anfangswertprobleme

In diesem Abschnitt wollen wir zunächst eine einfache Möglichkeit zur numerischen Berechnung von Lösungen gewöhnlicher Differentialgleichungen der folgenden Form eines Anfangswertproblems behandeln:

$$\begin{aligned}
u'(t) \;&=\; F(t, u(t)), \qquad \forall\, 0 \leq t \leq T, \\
u(0) \;&=\; u_0 \in \mathbb{R}^n,
\end{aligned} \tag{3.5}$$

Die grundlegende Idee ist es sukzessive (d.h., aufeinander aufbauend) die Lösung zu verschiedenen Zeitschritten zu approximieren. Hierzu führen wir eine sogenannte **Zeitdiskretisierung der Differentialgleichung** durch indem wir das kontinuierliche Intervall $[0,T] \subset \mathbb{R}^+$ in $N \in \mathbb{N}$ Teilintervalle mit $N+1$ Zeitpunkten aufteilen. Der Einfachheit halber wählen wir im Folgenden uniforme Zeitschritte mit einer festgewählten Schrittweite, d.h., wir diskretisieren das Intervall $\Omega := [0,T]$ durch

$$\Omega_\tau \;:=\; \{t_k \in \Omega \,|\, t_k := k \cdot \tau,\; k = 0, \ldots, N\} \quad \text{mit} \quad \tau \;:=\; \frac{T}{N}.$$

Es sei bemerkt, dass ein analoges Vorgehen auch für nicht äquidistante Schrittweiten möglich ist.

Basierend auf der Diskretisierung der Differentialgleichung können wir dann sukzessive numerische Approximationen $u_\tau(t_k) \in \mathbb{R}^n$ für die unbekannte Lösung u des Anfangswertproblems zu diesen diskreten Zeitschritten berechnen. Hierbei haben wir die Möglichkeit zeitliche Ableitung $\frac{d}{dt}$ durch geeignete Differenzenquotienten zu den diskreten Zeitpunkten zu approximieren. Alternativ lässt sich auch eine numerische Quadraturformel (siehe [Numerik 1, Kapitel 6]) für die zugehörige Integralgleichung in diesen Zeitschritten anwenden. Hierbei macht man sich die folgende auf dem Hauptsatz der Differential- und Integralrechnung basierende Beobachtung zu Nutze:

$$u(t_{k+1}) \;=\; u(t_k) + \int_{t_k}^{t_{k+1}} u'(t)\,\mathrm{d}t \;=\; u(t_k) + \int_{t_k}^{t_{k+1}} F(t,u(t))\,\mathrm{d}t, \qquad t_k \in \Omega_\tau. \tag{3.6}$$

Im Falle von Einschrittverfahren, die wir in diesem Abschnitt zunächst behandeln wollen, verwenden wir dabei eine Differenzenformel, die nur einen einzigen Zeitschritt berücksichtigt, d.h. zur Berechnung von $u_\tau(t_{k+1})$ wird lediglich der Wert $u_\tau(t_k)$ des vorherigen Zeitschritts verwendet.

Im Folgenden wollen wir zunächst eine allgemeine Form für Einschrittverfahren zur numerischen Lösung von Anfangswertproblemen angeben.

DEFINITION 3.8: Einschrittverfahren und Verfahrensfunktion.

Sei eine Zeitdiskretisierung Ω_τ des Anfangswertproblems (3.5) für das Zeitinterval $[0,T] \subset \mathbb{R}^+$ mit $n+1$ äquidistanten Zeitschritten $t_k = k \cdot \tau$ und $\tau = T/n$ gegeben. Dann lassen sich **Einschrittverfahren** zur numerischen Lösung des Anfangswertproblems in folgender allgemeiner Form angeben:

$$u_\tau(t_{k+1}) \;=\; u_\tau(t_k) + \tau \cdot f_\tau(t_k, u_\tau(t_k)), \tag{3.7}$$

wobei $f_\tau \colon \Omega_\tau \times \mathbb{R}^n \to \mathbb{R}^n$ eine numerische Approximation der Funktion F auf der rechten Seite der gewöhnlichen Differentialgleichung in (3.5) darstellt und allgemein **Verfahrensfunktion** genannt wird.
Hängt die Verfahrensfunktion f_τ nur vom bekannten Wert $u_\tau(t_k)$ ab, so nennen wir das Einschrittverfahren **explizit**, da es eine explizite Vorschrift zur Berechnung des nächsten Zeitschritts basierend auf den Werten des aktuellen Zeitschritts liefert. Hängt andererseits die Verfahrensfunktion f_τ auch vom unbekannten Wert

$u_\tau(t_{k+1})$ ab, so heißt das Verfahren **implizit**, da es nur eine implizite Bedingung (im Allgemeinen eine nichtlineare Gleichung) für den neuen Zeitschritt liefert.

Ob ein Einschrittverfahren explizit oder implizit ist, hängt häufig von der Wahl der Approximation der Zeitableitung mittels finiter Differenzen ab, die in folgender Definition eingeführt werden.

DEFINITION 3.9: Finite Differenzen.

Die Grundidee der sogenannten **finiten Differenzen** ist die Approximation der Ableitung $u'(t)$ durch Differenzenquotienten mit Werten auf einem Diskretisierungsgitter Ω_τ für eine feste Schrittweite $\tau := \frac{T}{N}, N \in \mathbb{N}$. Differenzenbildung auf einem Gitter. Hierbei haben wir verschiedene Möglichkeiten die Zeitableitung zu approximieren und wir unterscheiden folgende Fälle:

$$
\begin{aligned}
u'(t_k) &\approx \frac{u(t_{k+1}) - u(t_k)}{\tau} = \frac{u(t_k + \tau) - u(t_k)}{\tau} &&=: D^+u(t_k),\\
u'(t_k) &\approx \frac{u(t_k) - u(t_{k-1})}{\tau} = \frac{u(t_k) - u(t_k - \tau)}{\tau} &&=: D^-u(t_k),\\
u'(t_k) &\approx \frac{u(t_{k+1}) - u(t_{k-1})}{2\tau} = \frac{u(t_k + \tau) - u(t_k - \tau)}{2\tau} &&=: D^cu(t_k).
\end{aligned}
$$

Hierbei werden die obigen Differenzenquotienten D^+, D^-, D^c **Vorwärts-, Rückwärts-** und **zentrale Differenz** genannt.

Mittels Taylorformel sieht man ein, dass alle drei finiten Differenzen in Definition 3.9 eine gute numerische Approximation für die erste Ableitung der Funktion u liefern, wenn die Schrittweite $\tau > 0$ der Diskretisierung hinreichend klein gewählt wird. Da der Fehler dieser Approximation für $\tau \to 0$ bei allen drei Varianten gegen Null konvergiert spricht man auch von *konsistenten Verfahren* (mehr dazu später). Bei einer quantitativen Analyse des Fehlerglieds in der Taylorentwicklung lässt sich zeigen, dass die Vorwärts- und Rückwärtsdifferenz D^+ und D^- jeweils die Konsistenzordnung 1 besitzen, d.h., der Fehler verschwindet linear bzgl. der Schrittweite τ, während die zentrale Differenz D^c sogar Konsistenzordnung 2 besitzt, d.h., der Fehler verschwindet quadratisch bzgl. der Schrittweite τ.

Wir betrachten nun einige bekannte Einschrittverfahren, die häufig zur numerischen Lösung von Anfangswertproblemen genutzt werden.

BEISPIEL 3.10: Einschrittverfahren.

Im Folgenden wollen wir drei unterschiedliche Einschrittverfahren diskutieren, die häufig zur numerischen Lösung von Anfangswertproblemen genutzt werden.

1. Das einfachste Beispiel eines Einschrittverfahrens ist das **Vorwärts-Euler-Verfahren**

$$u_\tau(t_{k+1}) \;=\; u_\tau(t_k) + \tau \cdot F(t_k, u_\tau(t_k)), \tag{3.8}$$

welches auf einer Approximation der Ableitung $u'(t_k)$ mittels Vorwärtsdifferenz aus Definition 3.9 basiert. Das Vorwärts-Euler Verfahren hat die Verfahrensfunktion

$$f_\tau(t_k, u_\tau(t_k)) \;:=\; F(t_k, u_\tau(t_k)),$$

und ist entsprechend nach Definition 3.8 ein explizites Verfahren. Daher wird es häufig auch explizites Euler-Verfahren genannt.

In der Integralform der Differentialgleichung bedeutet das, dass wir die folgende Approximation benutzen:

$$u(t_{k+1}) - u(t_k) \;=\; \int_{t_k}^{t_{k+1}} F(t, u(t))\,\mathrm{d}t \;\approx\; \tau \cdot F(t_k, u_\tau(t_k)),$$

d.h., dass wir das Integral durch die Intervallänge mal dem Wert am linken Intervallrand annähern. Dies entspricht einer numerischen Quadraturformel mit nur einer Stützstelle (siehe [Numerik 1, Kapitel 6]).

2. Ein wenig komplizierter ist das **Rückwärts-Euler-Verfahren**

$$u_\tau(t_{k+1}) \;=\; u_\tau(t_k) + \tau \cdot F(t_{k+1}, u_\tau(t_{k+1})), \tag{3.9}$$

welches auf einer Approximation der Ableitung $u'(t_k)$ mittels Rückwärtsdifferenz aus Definition 3.9 basiert. In diesem Fall müssen wir zunächst eine (möglicherweise nichtlineare) Gleichung für den neuen Zeitschritt lösen.

Das Rückwärts-Euler Verfahren hat die Verfahrensfunktion

$$f_\tau(t_k, u_\tau(t_k)) \;:=\; F(t_{k+1}, u_\tau(t_{k+1})),$$

und ist entsprechend nach Definition 3.8 ein implizites Verfahren. Daher wird es häufig auch implizites Euler-Verfahren genannt.

In der Integralform der Differentialgleichung bedeutet das, dass wir die folgende Approximation benutzen:

$$u(t_{k+1}) - u(t_k) \;=\; \int_{t_k}^{t_{k+1}} F(t, u(t))\,\mathrm{d}t \;\approx\; \tau \cdot F(t_{k+1}, u_\tau(t_{k+1})),$$

d.h., dass wir das Integral durch die Intervallänge mal dem Wert am rechten Intervallrand annähern. Auch dies entspricht einer numerischen Quadraturformel mit nur einer Stützstelle wie beim expliziten Euler-Verfahren.

3. Ein Kompromiss aus den beiden obigen Verfahren ist das sogenannte **Crank-Nicolson Verfahren** der Form

$$u_\tau(t_{k+1}) \;=\; u_\tau(t_k) + \frac{\tau}{2} \cdot \left(F(t_k, u_\tau(t_k)) + F(t_{k+1}, u_\tau(t_{k+1}))\right). \tag{3.10}$$

Hierbei ist die Verfahrensfunktion der *Mittelwert* der Verfahrensfunktionen des Vorwärts- und Rückwärts-Euler-Verfahrens mit

$$f_\tau(t_k, u_\tau(t_k)) \;:=\; \frac{1}{2}\left(F(t_k, u_\tau(t_k)) + F(t_{k+1}, u_\tau(t_{k+1}))\right)$$

und ist entsprechend nach Definition 3.8 ebenfalls ein implizites Verfahren.

In der Integralform der Differentialgleichung bedeutet das, dass wir die folgende Approximation benutzen:

$$\begin{aligned} u(t_{k+1}) - u(t_k) \;&=\; \int_{t_k}^{t_{k+1}} F(t, u(t))\,\mathrm{d}t \\ &\approx\; \frac{\tau}{2}\left(F(t_k, u_\tau(t_k)) + F(t_{k+1}, u_\tau(t_{k+1}))\right), \end{aligned}$$

Dies entspricht einer Approximation des Integrals mit Hilfe der Trapezregel aus [Numerik 1, Kapitel 6.1].

Wir sehen, dass ein explizites Verfahren sofort wohldefiniert ist, falls die Verfahrensfunktion f_τ eine stetige Funktion auf $\mathbb{R}_+ \times \mathbb{R}^n$ ist, während bei impliziten Verfahren noch eine Fixpunktgleichung gelöst werden muss. Die Existenz und Eindeutigkeit dieser Gleichung können wir mit dem **Banachschen Fixpunktsatz** garantieren, wenn wiederum τ klein genug ist.

THEOREM 3.11: Existenz und Eindeutigkeit von Lösungen.

Sei $f_\tau : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ stetig und Lipschitz-stetig bezüglich dem letzten Argument mit Lipschitzkonstante $L > 0$. Dann existiert für $\tau < \frac{1}{L}$ genau eine Lösung $u_\tau(t_{k+1})$ der Fixpunktgleichung

$$u = u_\tau(t_k) + \tau\; f_\tau(t_k, u_\tau(t_k), u).$$

Numerisch müssen wir zur Durchführung eines impliziten Verfahrens immer noch ein System in $\mathbb{R}^n$ lösen. Ist dieses linear, so können wir die üblichen Verfahren für lineare Gleichungssysteme anwenden. Andernfalls bietet sich die Verwendung eines iterativen Verfahrens wie einer Fixpunktiteration oder des Newton-Verfahrens an (beachte, dass unter der obigen Bedingung $\mathbb{1} - \tau f'_\tau$ invertierbar ist für $f_\tau \in C^1$). Mit dem Wert $u_\tau(t_k)$ oder einer einfachen Vorhersage in der Zeit (etwa mit dem expliziten Euler-Verfahren) haben wir dafür auch einen sehr guten Startwert.

Nachdem wir die Wohldefiniertheit und numerische Umsetzung von Einschrittverfahren geklärt haben, widmen wir uns nun der Analyse der Verfahren. Wir wollen dabei den Fehler

$$E_\tau = \max_{k\in\mathbb{N}} \|u_\tau(t_k) - u(t_k)\| \tag{3.11}$$

abschätzen, wobei u die exakte Lösung des Anfangswertproblems ist. Wollen wir den Fehler an anderen Stellen t abschätzen, so können wir ein Interpolationsverfahren und die entsprechenden Abschätzungen anwenden.

Unsere Strategie dabei ist die Folgende: Zunächst schreiben wir eine Gleichung für den Fehler $e_\tau = u_\tau - u$. Es gilt

$$\begin{aligned} e_\tau(t_{k+1}) =& e_\tau(t_k) + \\ & \tau(f_\tau(t_k, u_\tau(t_k), u_\tau(t_{k+1})) - f_\tau(t_k, u(t_k), u(t_{k+1}))) + \\ & \tau\left[f_\tau(t_k, u(t_k), u(t_{k+1})) - \frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t, u(t))\ dt\right]. \end{aligned}$$

Nun benötigen wir zwei zentrale Eigenschaften von Diskretisierungsmethoden:

- **Konsistenz:** Der Fehler

  $$f_\tau(t_k, u(t_k), u(t_{k+1})) - \frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t, u(t))\ dt,$$

  d.h. das Residuum der Lösung des Anfangswertproblems eingesetzt in das numerische Verfahren konvergiert gegen Null für $\tau \to 0$.

- **Stabilität:** Bei der Umsetzung des numerischen Verfahrens mit gegebener rechter Seite wird diese nicht beliebig verstärkt, insbesondere existiert eine Abschätzung unabhängig von τ.

Zusammen ergeben Konsistenz und Stabilität Konvergenz des Verfahrens, d.h. $E_\tau \to 0$. Dies halten wir in einer Definition fest:

DEFINITION 3.12: Konvergenz von Einschrittverfahren.

Sei E_τ definiert durch (3.11), dann heißt das Verfahren

(i) konvergent, wenn $E_\tau \to 0$ für $\tau \to 0$,

(ii) konvergent von der Ordnung p, wenn $E_\tau = \mathcal{O}(\tau^p)$ für $\tau \to 0$, d.h. es gibt eine Konstante C_p, sodass $E_\tau \leq C_p\tau^p$ für τ hinreichend klein.

3.2.1 Konsistenz von Einschrittverfahren

Gemäß der obigen Motivation definieren wir den Konsistenzfehler als

$$K_\tau = \max_{k\in\mathbb{N}} \|g_\tau(t_k)\| \tag{3.12}$$

mit

$$g_\tau(t_k) = f_\tau(t_k, u(t_k), u(t_{k+1})) - \frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t, u(t))\ dt, \tag{3.13}$$

wobei u eine Lösung des Anfangswertproblems (3.1) ist.

DEFINITION 3.13: Konsistenzfehler.

Sei K_τ definiert durch (3.12), dann heißt das Verfahren

(i) konsistent, wenn $K_\tau \to 0$ für $\tau \to 0$,

(ii) konsistent von der Ordnung p, wenn $K_\tau = \mathcal{O}(\tau^p)$ für $\tau \to 0$.

Die Abschätzung des Konsistenzfehlers erfolgt meist durch Taylorentwicklung, wir führen dies an zwei Beispielen durch:

BEISPIEL 3.14.

Wir betrachten das Vorwärts-Euler Verfahren unter der Annahme, dass F bezüglich beider Variablen Lipschitz-stetig ist. Definieren wir $\varphi(t) = F(t, u(t))$, dann ist φ wegen $u \in C^1$ eine Lipschitz-stetige Funktion und es gilt

$$\begin{aligned}
\|g_\tau(t_k)\| &= \|\varphi(t_k) - \frac{1}{\tau}\int_{t_k}^{t_{k+1}} \varphi(t)\ dt\| \\
&= \|\frac{1}{\tau}\int_{t_k}^{t_{k+1}} (\varphi(t_k) - \varphi(t))\ dt\| \\
&\le \frac{1}{\tau}\int_{t_k}^{t_{k+1}} \|\varphi(t_k) - \varphi(t)\|\ dt \\
&\le \frac{1}{\tau}\int_{t_k}^{t_{k+1}} L_\varphi(t - t_k)\ dt = \frac{L_\varphi}{2}\tau.
\end{aligned}$$

Damit das Verfahren die Konsistenzordnung $p = 1$, wir sehen im Beispiel $F(t, u) = t$ auch sofort, dass man im allgemeinen nicht Ordnung zwei erreichen kann.

BEISPIEL 3.15.

Wir betrachten das Crank–Nicholson Verfahren unter der Annahme, dass F bezüglich beider Variablen zweimal stetig differenzierbar ist. Definieren wir $\varphi(t) = F(t, u(t))$, dann ist φ ebenfalls zweimal stetig differenzierbar, da

$$u''(t) = (F(t, u(t)))' = \partial_t F(t, u(t)) + \partial_u F(t, u(t))u'(t)$$

stetig ist. Damit gilt

$$\begin{aligned}\|g_\tau(t_k)\| &= \|\frac{1}{2}(\varphi(t_k)+\varphi(t_{k+1})) - \frac{1}{\tau}\int_{t_k}^{t_{k+1}} \varphi(t)\ dt\| \\ &= \frac{1}{2\tau}\|\int_{t_k}^{t_{k+1}} (\varphi(t_k)+\varphi(t_{k+1})-2\varphi(t))\ dt\| \\ &\leq \frac{1}{2\tau}\|\int_{t_k}^{t_{k+1}} \varphi'(t_k)(t_k+t_{k+1}-2t)+r_k\ dt\|,\end{aligned}$$

mit dem Restglied $r_k = \mathcal{O}(\tau^2)$. Da $\int_{t_k}^{t_{k+1}} \varphi'(t_k)(t_k+t_{k+1}-2t)\ dt = 0$, folgt $\|g_\tau(t_k)\| = \mathcal{O}(\tau^p)$. Damit das Verfahren die Konsistenzordnung $p=2$, wir sehen im Beispiel $F(t,u)=t^2$ auch sofort, dass man im allgemeinen nicht Ordnung zwei erreichen kann.

Um eine Konvergenzordnung p zu erhalten, benötigen wir, dass F, aber auch u p-mal stetig differenzierbar ist. Aus den Eigenschaften von F folgt Letzteres aber sofort: wir haben gesehen, dass für F stetig auch u stetig differenzierbar folgt. Mit dem Argument aus dem letzten Beispiel sehen wir, dass für F stetig differenzierbar auch u zweimal stetig differenzierbar ist. Induktiv können wir durch weiteres differenzieren zeigen, dass u p-mal stetig differenzierbar ist, wenn F $p-1$-mal stetig differenzierbar ist.

3.2.2 Stabilität und Konvergenz

Wir widmen uns nun der Frage der Stabilität von Einschrittverfahren. Hierbei verwenden wir eine diskrete Version des Lemmas von Gronwall.

LEMMA 3.16: Diskretes Lemma von Gronwall.

Es sei $\beta_j \geq 0, j \in \mathbb{N}_0$ eine Folge nicht-negativer Zahlen und für die Folge $u_j \in \mathbb{R}, j \in \mathbb{N}_0$ gelte

$$\begin{aligned}u_0 &\leq \alpha \in \mathbb{R}_0^+ \\ u_k &\leq \alpha + \sum_{j=0}^{k-1} \beta_j u_j\end{aligned}$$

für $k \in \mathbb{N}$, dann gilt die Abschätzung

$$u_k \leq \alpha \exp\left(\sum_{j=0}^{k-1} \beta_j\right).$$

Beweis. Übung. □

Wir zeigen zunächst uniforme Schranken an u_τ.

LEMMA 3.17.

Sei F_τ stetig und Lipschitz-stetig bezüglich des dritten Arguments (d.h. $u_\tau(t_{k+1})$ mit Modul unabhängig von τ. Dann existiert eine Konstante $M(T)$ unabhängig von τ, sodass

$$\max_{t_k \le T} \|u_\tau(t_k)\| \le M$$

gilt für alle τ hinreichend klein.

Beweis. Aus der Definition des Verfahrens folgt

$$u_\tau(t_{k+1}) - u_0 = u_\tau(t_k) - u_0 + \tau(f_\tau(t_k, u_\tau(t_k), u_\tau(t_{k+1})) - f_\tau(t_k, u_0, u_0)) + \tau f_\tau(t_k, u_0, u_0)$$

und mit der Dreiecksungleichung folgt für $v_k = \|u_\tau(t_k) - u_0\|$

$$\begin{aligned} v_{k+1} &\le v_k + \tau\|f_\tau(t_k, u_\tau(t_k), u_\tau(t_{k+1})) - f_\tau(t_k, u_0, u_0)\| + \tau\|f_\tau(t_k, u_0, u_0)\| \\ &\le v_k + \tau L(v_k + v_{k+1}) + \tau C. \end{aligned}$$

Hier haben wir benutzt, dass f_τ stetig ist, damit folgt $f_\tau(t, u_0, u_0)$ ist auf dem kompakten Intervall $[0, T]$ durch eine Konstante C beschränkt. Dazu bezeichnet L den Lipschitz-Modul von f_τ bezüglich zweitem und drittem Argument. Sei nun $\tau \le \frac{1}{2L}$, d.h. $1 - \tau L \ge \frac{1}{2}$, dann folgt

$$v_{k+1} \le 2(1 + \tau L)v_k + 2\tau C.$$

Das diskrete Lemma von Gronwall impliziert dann die Beschränktheit von v_k. □

Mit einem ähnlichen Beweis können wir auch die Stabilität zeigen.

THEOREM 3.18: Stabilität von Einschrittverfahren.

Sei $u_\tau \colon \Omega_\tau \to \mathbb{R}^n$ die Lösung eines Einschrittverfahrens mit Lipschitz-stetiger Verfahrensfunktion f_τ für eine Zeitdiskretisierung $\Omega_\tau := \{t_k \in [0, T] \colon t_k := \frac{k \cdot T}{N}, k = 0, \dots, N\}$ mit Schrittweite $\tau := \frac{T}{N} > 0$. Sei außerdem $u \colon [0, T] \to \mathbb{R}$ die Lösung des Anfangswertproblems (3.1) mit gleichem Anfangswert $u_0 \in \mathbb{R}^n$. Der lokale Konsistenzfehler $g_\tau(t_k)$ und der globale Konsistenzfehler K_τ seien gegeben wie in Definition 3.13.
Dann existiert eine Konstante $C > 0$, so dass für $\tau > 0$ hinreichend klein die folgende Abschätzung für den Fehler des Einschrittverfahrens gilt:

$$E_\tau \;=\; \max_{t_k} \|u_\tau(t_k) - u(t_k)\| \;\le\; C \cdot \max_{t_k} \|g_\tau(t_k)\| \;=\; C \cdot K_\tau.$$

Beweis. Wir definieren uns die Hilfsfunktion $v_k := \|u_\tau(t_k) - u(t_k))\|$. Durch Anwendung der Dreiecksungleichung und dem Ausnutzen der Lipschitz-Stetigkeit von f_τ erhalten

wir dann

$$\begin{aligned} v_{k+1} &= \|u_\tau(t_{k+1}) - u(t_{k+1})\| \\ &= \left\| u_\tau(t_k) + \tau \cdot f_\tau(t_k, t_{k+1}, u_\tau(t_k), u_\tau(t_{k+1})) - u(t_k) - \int_{t_k}^{t_{k+1}} F(t, u(t))\,\mathrm{d}t \right\| \\ &\leq v_k + \tau \cdot \|f_\tau(t_k, t_{k+1}, u_\tau(t_k), u_\tau(t_{k+1})) - f_\tau(t_k, t_{k+1}, u(t_k), u(t_{k+1}))\| \\ &\quad + \tau \cdot \left\| f_\tau(t_k, t_{k+1}, u(t_k), u(t_{k+1})) - \frac{1}{\tau} \int_{t_k}^{t_{k+1}} F(t, u(t))\,\mathrm{d}t \right\| \\ &\leq v_k + \tau \cdot L \cdot (v_k + v_{k+1}) + \|g_\tau(t_k)\|. \end{aligned}$$

Somit erhalten wir durch Umstellen also insgesamt

$$\begin{aligned} & v_{k+1} \leq v_k + L \cdot \tau \cdot (v_k + v_{k+1}) + \|g_\tau(t_k)\| \\ \Rightarrow \quad & v_{k+1} \leq \frac{1 + \tau L}{1 - \tau L} \cdot v_k + \max_{t_k} \|g_\tau(t_k)\|. \end{aligned}$$

Für hinreichend kleine Schrittweiten $\tau < \frac{1}{2L}$ erhalten wir die gewünschte Schranke wieder direkt aus dem diskreten Lemma 3.16 von Gronwall. □

Eine direkte Folgerung von Theorem 3.18 ist die Äquivalenz von Konsistenz und Konvergenz für Einschrittverfahren, wie wir im folgenden Korollar feststellen.

KOROLLAR 3.19: Äquivalenz von Konsistenz und Konvergenz.

Für ein Einschrittverfahren mit Lipschitz-stetiger Verfahrensfunktion gilt: ist das Verfahren konsistent von der Ordnung p, so ist es auch konvergent von der Ordnung p.

Mit Hilfe dieses Korollars und der Abschätzung des Konsistenzfehlers stellen wir fest, dass das Vorwärts- und Rückwärts-Euler Verfahren konvergent von der Ordnung eins sind. Das Crank–Nicholson Verfahren hingegen ist sogar konvergent von der Ordnung zwei.

3.2.3 Runge–Kutta Verfahren

Bisher haben wir Verfahren kennengelernt, die auf einzelne Funktionsauswertungen von F an den Zeitschritten t_k und t_{k+1} zurückgreifen. Damit haben wir bereits die Konsistenzordnung Eins erreicht für die beiden Euler-Verfahren und als maximale Konsistenzordnung Zwei im Falle des Crank–Nicholson Verfahrens. Eine höhere Konsistenzordnung ist mit so einem Ansatz nicht möglich. Wir widmen uns im Folgenden also der Frage wie wir Einschrittverfahren höherer Ordnung konstruieren können. Aus der Analyse der Konsistenzordnung lässt sich ableiten, dass wir Verfahrensfunktionen der Art konstruieren müssen, so dass die Taylorentwicklung ein Restglied höherer Ordnung liefert.

Eine erste Idee zur Steigerung der Konsistenzordnung ist es Ableitungen von F in den Zeitschritten t_k und t_{k+1} zu berücksichtigen, womit man offensichtlich die Taylor-Entwicklung besser approximieren und somit eine höhere Ordnung erreichen kann. Die

Berechnung von Ableitungen von F ist jedoch potentiell numerisch aufwändig und instabil, deswegen geht alternativ einen anderen Weg und verwendet geschachtelte Funktionsauswertungen.

Die grundlegende Idee ist es das Integral der Anfangswertaufgabe in (3.6) durch eine geeignete Quadraturformel (siehe [Numerik 1, Kapitel 6]) der folgenden Form zu approximieren:

$$\frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t,u(t))\,\mathrm{d}t \;\approx\; f_\tau(t_k,u_\tau(t_k)) \;:=\; \sum_{i=1}^{s} b_i f_i^k.$$

Hierbei stellen die $b_i \in \mathbb{R}$ Gewichte der Quadraturformel dar und die Zwischenwerte $f_i^k \in \mathbb{R}^n$ sollen eine Approximation der Funktionswerte von F liefern für $i = 1,\dots,s$ mit

$$f_i^k \;\approx\; F(t_k + c_i\tau, u(t_k + c_i\tau)).$$

Hierbei sind die $0 \le c_i \le 1, i = 1,\dots,s$ aufsteigende Parameter, die Stützstellen für die jeweiligen Funktionsauswertung definieren. Da wir die Werte der unbekannten Funktion $u_\tau(t_k + c_i\tau)$ an den Stützstellen nicht kennen, benötigen wir ebenfalls Approximationen mit Hilfe von Quadraturformeln der folgenden Form:

$$u(t_k + c_i\tau) \;=\; u(t_k) + \int_{t_k}^{t_k+c_i\tau} F(t,u(t))\,\mathrm{d}t \;\approx\; u_\tau(t_k) + \tau\sum_{j=1}^{s} a_{ij} f_j^k.$$

Diese Idee führt zur Definition von sogenannten Runge-Kutta Verfahren.

DEFINITION 3.20: Runge–Kutta Verfahren der Stufe s.

Bei einem **Runge–Kutta Verfahren der Stufe s**, $s \in \mathbb{N}^+$ berechnet man

$$u_\tau(t_{k+1}) \;=\; u_\tau(t_k) + \tau f_\tau(t_k,u_\tau(t_k)) \;=\; u_\tau(t_k) + \tau\sum_{i=1}^{s} b_i f_i^k.$$

Die Zwischenwerte f_i^k lassen sich als Funktionsauswertungen von F wie folgt berechnen

$$f_i^k \;:=\; F\Big(t_k + c_i\tau, u_\tau(t_k) + \tau\sum_{j=1}^{s} a_{ij} f_j^k\Big), \qquad i = 1,\dots,s. \tag{3.14}$$

Um in der Zeit vorwärts zu gehen wählt man die Koeffizienten $0 \le c_i \le 1, i = 1,\dots,s$ als aufsteigende Folge und die Matrix $(a_{ij})_{i,j=1,\dots,s}$ als linke, untere Dreiecksmatrix.
Man nennt ein Runge–Kutta Verfahren **explizit** falls für alle Einträge der Matrix $A \in \mathbb{R}^{s\times s}$ gilt $a_{ij} = 0, i = 1,\dots,s$ wenn $j \ge i$.

Die zu bestimmenden Koeffizienten $b_i, c_i \in \mathbb{R}$ und $a_{i,j} \in \mathbb{R}$ für $i,j = 1,\dots,s$ in Definition 3.20 erhält man durch einen Vergleich mit der Taylorentwicklung des zu approximierenden Integrals $\frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t,u(t))\,\mathrm{d}t$. Das folgende Beispiel soll die Idee zur

Herleitung eines Runge–Kutta Verfahren für den einfachen Fall der Stufe $s = 1$ verdeutlichen.

BEISPIEL 3.21: Runge–Kutta Verfahren der Stufe 1.

Für ein Runge–Kutta Verfahren der Stufe $s = 1$ ist die Verfahrensfunktion gegeben durch

$$f_\tau(t_k, u_\tau(t_k)) \;:=\; b_1 f_1^k,$$

und für den Zwischenwert $f_1^k \in \mathbb{R}^n$ gilt entsprechend:

$$f_1^k \;=\; F(t_k + c_1\tau, u_\tau(t_k) + \tau a_{11} f_1^k).$$

Wollen wir nun ein **explizites Verfahren** herleiten, so muss nach Definition 3.20 schon $a_{11} = 0$ gelten. Die Verfahrensfunktion ist also von der Form

$$f_\tau(t_k, u_\tau(t_k)) \;=\; b_1 f_1^k \;=\; b_1 F(t_k + c_1\tau, u_\tau(t_k)).$$

Die einzige sinnvolle Wahl für den Koeffizienten c_1 ist Null, da wir sonst die Funktion F zu einer anderen Zeit auswerten als die unbekannte Funktion u. Um ein konsistentes Verfahren zu erhalten sieht man außerdem ein, dass $b_1 = 1$ gelten muss. Also erhalten wir schlussendlich das Vorwärts-Euler Verfahren.

Im **impliziten Fall** können wir interessanterweise eine höhere Ordnung erreichen. Wir berechnen hierzu die Taylor-Entwicklung des Zwischenwerts f_1^k wie folgt

$$\begin{aligned} f_1^k &= F(t_k + c_1\tau, u_\tau(t_k) + \tau a_{11} f_1^k) \\ &= F(t_k, u(t_k)) + c_1\tau \partial_t F(t_k, u(t_k)) + \tau a_{11} D_u F(t_k, u(t_k)) f_1^k + \mathcal{O}(\tau^2). \end{aligned} \tag{3.15}$$

Hierbei bezeichnet D_u die Jacobimatrix bezüglich des zweiten Arguments u von F. Setzen wir auf der rechten Seite nochmal den ersten Approximationsterm für f_1^k ein, so folgt

$$f_1^k \;=\; F(t_k, u(t_k)) + c_1\tau \partial_t F(t_k, u(t_k)) + \tau a_{11} D_u F(t_k, u(t_k)) F(t_k, u(t_k)) + \mathcal{O}(\tau^2).$$

Andererseits gilt mit einer Taylorentwicklung des unbekannten Funktionswerts $u(t_{k+1})$ im Punkt t_k:

$$\begin{aligned} \frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t, u(t))\,\mathrm{d}t &= \frac{1}{\tau}\left(u(t_{k+1}) - u(t_k)\right) \\ &= \frac{1}{\tau}\left(u(t_k) + \tau u'(t_k) + \frac{\tau^2}{2}u''(t_k) + \mathcal{O}(\tau^3) - u(t_k)\right) \\ &= u'(t_k) + \frac{\tau}{2}u''(t_k) + \mathcal{O}(\tau^2) \\ &= F(t_k, u(t_k)) + \frac{\tau}{2}\partial_t F(t_k, u(t_k)) \\ &\qquad + \frac{\tau}{2} D_u F(t_k, u(t_k)) F(t_k, u(t_k)) + \mathcal{O}(\tau^2). \end{aligned} \tag{3.16}$$

Damit gilt

$$f_\tau(t_k, u_\tau(t_k)) \;=\; b_1 f_1^k \;\approx\; \frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t, u(t))\,\mathrm{d}t,$$

vergleichen wir die beiden Formeln (3.15) und (3.16) und sehen ein, dass wir eine Konsistenzordnung von Zwei erreichen können, wenn die Koeffizienten als $b_1 = 1$, $c_1 = \frac{1}{2}$ und $a_{11} = \frac{1}{2}$ gewählt werden.
Die Verfahrensfunktion ist also gegeben durch die Lösung von

$$f_\tau(t_k, u_\tau(t_k)) \;=\; f_1^k \;=\; F(t_k + \frac{\tau}{2}, u_\tau(t_k) + \frac{\tau}{2} f_1^k).$$

Wir können dieses Runge-Kutta Verfahren als *Mittelpunktsregel* im Intervall $[t_k, t_{k+1}]$ interpretieren, wobei der unbekannte Wert von u_τ am Mittelpunkt $t_k + \frac{\tau}{2}$ des Intervalls durch das *Rückwärts-Euler-Verfahren* bestimmt wird.

Im Folgenden betrachten wir noch 2-stufige Runge–Kutta Verfahren.

BEISPIEL 3.22: Runge–Kutta Verfahren der Stufe 2.

Für ein Runge–Kutta Verfahren der Stufe $s = 2$ ist die Verfahrensfunktion gegeben durch

$$f_\tau(t_k, u_\tau(t_k)) \;:=\; b_1 f_1^k + b_2 f_2^k,$$

und für die Zwischenwerte $f_1^k, f_2^k \in \mathbb{R}^n$ gilt im **expliziten Fall** mit $a_{11} = a_{22} = 0$:

$$f_1^k \;=\; F(t_k + c_1\tau, u_\tau(t_k)), \quad f_2^k \;=\; F(t_k + c_2\tau, u_\tau(t_k) + \tau a_{21} f_1^k).$$

Analog zur Argumentation im expliziten Fall von Beispiel 3.21 sehen wir wieder, dass nur $c_1 = 0$ eine sinnvolle Wahl ist. Eine Taylor-Entwicklung des Zwischenwertes f_2^k liefert dann

$$\begin{aligned} b_1 f_1^k + b_2 f_2^k \;=\; & (b_1 + b_2) F(t_k, u(t_k)) + b_2 c_2 \tau \partial_t F(t_k, u(t_k)) \\ & + b_2 a_{21} \tau D_u F(t_k, u(t_k)) F(t_k, u(t_k)) + \mathcal{O}(\tau^2). \end{aligned}$$

Vergleichen wir dies wieder mit der Taylor-Entwicklung des Integrals in (3.16), so erhalten wir folgendes Gleichungssystem

$$b_1 + b_2 \;=\; 1, \qquad b_2 c_2 \;=\; \frac{1}{2}, \qquad b_2 a_{21} \;=\; \frac{1}{2}.$$

Eine einfache Lösung ist beispielsweise $b_1 = 0$, $b_2 = 1$, $c_2 = a_{21} = \frac{1}{2}$. Diese Wahl der Koeffizienten liefert uns also ein Verfahren der Konsistenzordnung zwei mit der Verfahrensfunktion

$$f_\tau(t_k, u_\tau(t_k)) \;=\; F(t_k + \frac{\tau}{2}, u_\tau(t_k) + \frac{\tau}{2} F(t_k, u(t_k))).$$

Wir können dieses zweistufige Runge-Kutta Verfahren als eine *Mittelpunktsregel* im Intervall $[t_k, t_{k+1}]$ interpretieren, wobei der unbekannte Wert von u_τ am Mittelpunkt $t_k + \frac{\tau}{2}$ durch das *Vorwärts-Euler-Verfahren* bestimmt wird.

Bei der Konstruktion eines Runge-Kutta Verfahrens in Beispiel 3.21 und Beispiel 3.22 haben wir die Koeffizienten $b_i, c_i \in \mathbb{R}$ und $a_{i,j} \in \mathbb{R}$ für $i, j = 1, \dots, s$ basierend auf vernünftigen Überlegungen zu wählen. Wie wir speziell im Fall des zweistufigen Runge-Kutta Verfahrens gesehen haben führt dies im Allgemeinen jedoch nicht zu eindeutigen Lösungen. Daher ist es sinnvoll Kriterien an die implizierten Runge-Kutta Verfahren zu stellen, so dass wir eindeutige Koeffizienten ableiten können.

Am einfachsten ist ein Kriterium für die Konsistenz des Runge-Kutta Verfahrens für die unbekannten Koeffizienten abzuleiten, wie das folgende Lemma festhält.

LEMMA 3.23: Konsistenzbedingung von Runge-Kutta-Verfahren.

Ein Runge-Kutta Verfahren der Stufe $s \in \mathbb{N}^+$ ist genau dann konsistent, wenn die folgende Bedingung erfüllt ist:

$$\sum_{i=1}^{s} b_i \;=\; 1.$$

Beweis. Um zu zeigen, dass ein Runge-Kutta Verfahren konsistent ist müssen wir zeigen, dass der Konsistenzfehler

$$K_\tau \;=\; \max_{k=0,\dots,N} \left\| f_\tau(t_k, u(t_k)) - \frac{1}{\tau} \int_{t_k}^{t_{k+1}} F(t, u(t))\,\mathrm{d}t \right\|$$

mindestens von der Fehlerordnung $\mathcal{O}(\tau)$ ist und für $\tau \to 0$ gegen Null geht.

Die Verfahrensfunktion des Runge-Kutta Verfahrens der Stufe s gegeben ist als

$$f_\tau(t_k, u_\tau(t_k)) \;=\; \sum_{i=1}^{s} b_i f_i^k$$

und daher können wir die Taylorapproximation erster Ordnung jedes Zwischenwerts $f_i^k \in \mathbb{R}^n$ für $i = 1, \dots, s$ analog zu (3.15) betrachten und erhalten somit

$$\sum_{i=1}^{s} b_i f_i^k \;=\; \sum_{i=1}^{s} (b_i F(t_k, u(t_k)) + \mathcal{O}(\tau)) \;=\; \sum_{i=1}^{s} b_i F(t_k, u(t_k)) + \mathcal{O}(\tau).$$

Die Taylorapproximation erster Ordnung des Integrals liefert nach (3.16) außerdem

$$\frac{1}{\tau} \int_{t_k}^{t_{k+1}} F(t, u(t))\,\mathrm{d}t = F(t_k, u(t_k)) + \mathcal{O}(\tau).$$

Nun gilt also für den Konsistenzfehler

$$\begin{aligned} K_\tau \;&=\; \max_{k=0,\dots,N} \left\| f_\tau(t_k, u(t_k)) - \frac{1}{\tau} \int_{t_k}^{t_{k+1}} F(t, u(t))\,\mathrm{d}t \right\| \\ &=\; \max_{k=0,\dots,N} \left\| \sum_{i=1}^{s} b_i F(t_k, u(t_k)) - F(t_k, u(t_k)) + \mathcal{O}(\tau) \right\|. \end{aligned}$$

Wir sehen, dass für den Konsistenzfehler $K_\tau \in \mathcal{O}(\tau)$ genau dann gilt, wenn $\sum_{i=1}^{s} b_i = 1$ ist. □

Umgekehrt ist die Konsistenzordnung eines Runge-Kutta Verfahrens nach oben durch dessen Stufe beschränkt, wie folgendes Lemma zeigt.

LEMMA 3.24: Maximale Konsistenzordnung von Runge-Kutta-Verfahren.

Für die Konsistenzordnung $p \in \mathbb{N}^+$ eines expliziten Runge-Kutta Verfahren der Stufe $s \in \mathbb{N}^+$ für ein Anfangswertproblem mit einer rechten Seite der Differentialgleichung $F \in C^\infty([0,T] \times \mathbb{R}^n; \mathbb{R}^n)$ gilt, dass die Konsistenzordnung durch die Stufe des Verfahrens nach oben beschränkt ist, d.h., es gilt $0 < p \leq s$.

Beweis. Wir zeigen die Behauptung für den reellen Fall $n = 1$ und wählen die besonders einfache rechte Seite $F(t,u) = u$ mit $u_0 = 1 \in \mathbb{R}$. Dann ist die Lösung des Anfangswertproblems

$$u'(t) \;=\; F(t,u(t)) \;=\; u(t), \qquad u(0) \;=\; 1,$$

gegeben durch $u(t) = u_0 \cdot e^t = e^t$.

Mit der Taylorentwicklung der Exponentialfunktion $u(t) = e^t = \sum_{j=0}^{\infty} \frac{t^j}{j!}$ können wir das Integral als Polynom in τ schreiben mit:

$$\begin{aligned}
\frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t,u(t))\,\mathrm{d}t \;&=\; \frac{1}{\tau}\int_{t_k}^{t_{k+1}} u(t)\,\mathrm{d}t \\
&= \frac{1}{\tau}\int_{t_k}^{t_{k+1}} \underbrace{u(t_k - t_k)}_{=1} \cdot u(t)\,\mathrm{d}t \\
&= \frac{1}{\tau}\int_{t_k}^{t_{k+1}} u(t_k)\cdot u(t-t_k)\,\mathrm{d}t \\
&= \frac{u(t_k)}{\tau}\int_{t_k}^{t_{k+1}} \sum_{j=0}^{p} \frac{(t-t_k)^j}{j!} + \mathcal{O}(\tau^{p+1})\,\mathrm{d}t \\
&= \frac{u(t_k)}{\tau}\sum_{j=0}^{p} \frac{(t_{k+1}-t_k)^{j+1}}{(j+1)!} + \mathcal{O}(\tau^p) \\
&= \sum_{j=0}^{p} \tau^j \frac{u(t_k)}{(j+1)!} + \mathcal{O}(\tau^p).
\end{aligned}$$

Andererseits sehen wir für ein beliebiges explizites Runge-Kutta Verfahren der Stufe $s \in \mathbb{N}^+$ mit Konsistenzordnung $p \leq s$, dass die Zwischenwerte gegeben sind durch:

$$\begin{aligned}
f_1^k \;&=\; u_\tau(t_k), \\
f_2^k \;&=\; u_\tau(t_k) + \tau a_{21} u_\tau(t_k), \\
f_3^k \;&=\; u_\tau(t_k) + \tau a_{31} u_\tau(t_k) + \tau a_{32} u_\tau(t_k) + \tau^2 a_{32} a_{21} u_\tau(t_k), \\
&\;\;\vdots
\end{aligned}$$

Wir sehen also, dass jeder Zwischenwert $f_i^k \in \mathbb{R}^n$ für $i = 1,\ldots,s$ ein Polynom in τ vom Grad kleiner gleich $i-1$ ist und somit ist die Verfahrensfunktion $f_\tau(t_k, u_\tau(t_k)) = \sum_{i=1}^{s} b_i f_i^k$ ein Polynom in τ vom Grad kleiner gleich $s-1$.

Damit können wir keine höhere Ordnung erreichen, da der Fehler ab der Ordnung $\mathcal{O}(\tau^p)$ im Allgemeinen nicht verschwindet. □

Man sieht ein, dass sich mittels der Forderung von Konsistenz in Lemma 3.23 und dem Bestreben nach maximaler Konsistenzordnung $p = s$ eines s-stufigen Runge-Kutta Verfahrens, das enstehende Gleichungssystem lösen lässt. Allerdings bleibt hierbei immer noch eine Uneindeutigkeit der Koeffizienten $c_i \in \mathbb{R}, i = 1, \dots, s$ übrig. Um hier eine systematische Wahl zu treffen, fordert man im Allgemeinheit die Invarianz des Verfahrens gegenüber sogenannter *Autonomisierung*. Hierzu führen wir zunächst den Begriff einer autonomen Differentialgleichung ein.

DEFINITION 3.25: Autonome gew. Differentialgleichung.

Sei $u'(t) = F(t, u(t)),\ u(0) = u_0 \in \mathbb{R}^n$ ein Anfangswertproblem einer gewöhnlichen Differentialgleichung erster Ordnung.
Wir nennen die Differentialgleichung **autonom**, wenn die Funktion F auf der rechten Seite nicht explizit von t abhängt, d.h., es gilt $F(t, u(t)) = F(u(t))$.

Das allgemeine Anfangswertproblem (3.1) lässt sich autonomisieren, d.h. in ein äquivalentes System autonomer Differentialgleichungen für $\tilde{u} = v, u)$ mit einer Hilfsfunktion v umschreiben. Hierzu betrachten wir das Differentialgleichungssystem

$$\begin{aligned} u'(t) &= F(v(t), u(t)) = F(\tilde{u}(t)), & u(0) &= u_0 \in \mathbb{R}^n, \\ v'(t) &= 1, & v(0) &= 0. \end{aligned}$$

Man erkennt sofort, dass $v(t) = t$ gelten muss, was die Äquivalenz zu (3.1) impliziert.

Wir nennen die obige Transformation des Anfangswertproblems Autonomisierung und fordern, dass das Runge-Kutta Verfahren unter der Autonomisierung invariant sein soll. Dies bedeutet, dass wir fordern, dass die Anwendung des numerischen Verfahrens auf das neue System die selbe Lösung u_τ liefern soll, wie die Anwendung auf das ursprüngliche Anfangswertproblem (3.1). Darüber hinaus fordern wir, dass die Invarianz unter Autonomisierung für jede geeignete Funktion F auf der rechten Seite gelten soll.

Schreiben wir die Zwischenwerte für beide Formulierungen, so gilt für das *ursprüngliche Anfangswertproblem*

$$F(t_k + c_i\tau, u(t_k + c_i\tau)) \approx F(t_k + c_i\tau, u(t_k) + \tau \sum_{j=1}^{s} a_{ij} f_j^k) =: f_i^k,$$

und für die *autonome Variante* andererseits

$$F(v(t_k + c_i\tau), u(t_k + c_i\tau)) \approx F(v(t_k) + \tau \sum_{j=1}^{s} a_{ij}, u(t_k) + \tau \sum_{j=1}^{s} a_{ij} f_j^k) =: f_i^k.$$

Da wegen der exakten numerischen Integration der linearen Funktion $t \mapsto t$ immer $v(t_k) = t_k$ gilt, sehen wir durch Koeffizientenvergleich, dass die Invarianz gegenüber

Autonomisierung äquivalent ist zu der Bedingung

$$c_i \;=\; \sum_{j=1}^{s} a_{ij}, \qquad i = 1, \ldots, s. \tag{3.17}$$

Aus diesem Grund bestimmt man die Koeffizienten $c_i \in \mathbb{R}$ immer aus Gleichung (3.17) und lediglich die Einträge a_{ij} der Matrix A werden aus der Konsistenzbedingung mittels Taylorentwicklung bestimmt. Damit ist es möglich ein explizites s-stufiges Runge-Kutta Verfahren der Konsistenzordnung $s \in N^+$ zu konstruieren

Butcher-Schema

Allgemein lässt sich ein s-stufiges Runge-Kutta Verfahren durch die Matrix A und die Vektoren b, c eindeutig repräsentieren. Diese lassen sich kompakt in Form eines Schemas angeben.

DEFINITION 3.26: Butcher-Schema.

Sei $s \in \mathbb{N}^+$ die Stufe eines Runge-Kutta Verfahrens mit Koeffizienten, die gegeben sind durch eine Matrix $A \in \mathbb{R}^{s\times s}$ und den beiden Vektoren $\mathbf{c}, \mathbf{b} \in \mathbb{R}^s$. Dann lässt sich das Runge-Kutta Verfahren kompakt durch das folgende **Butcher-Schema** repräsentieren:

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b}^T \end{array} \;=\; \begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} .$$

Im folgenden Beispiel geben wir die Butcher-Schemata für vier explizite Runge-Kutta Verfahren im Vergleich an.

BEISPIEL 3.27: Explizite Runge–Kutta Verfahren.

$s = 1$

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

Vorwärts-Euler Verfahren mit:
$u_\tau(t_{k+1}) = u_\tau(t_k) + \tau F(t_k, u_\tau(t_k))$.
Konsistenzordnung $p = 1$.

$s = 2$

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Verbessertes Euler Verfahren mit:
$u_\tau(t_{k+1}) = u_\tau(t_k) + \frac{\tau}{2} F(t_k, u_\tau(t_k)) + \frac{\tau}{2} F(t_k + \tau, u_\tau(t_k) + \tau F(t_k, u_\tau(t_k)))$.
Konsistenzordnung $p = 2$.

$s = 3$

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & -1 & 2 & 0 \\ \hline & \frac{1}{6} & \frac{4}{6} & \frac{1}{6} \end{array}$$

$$\begin{aligned} u_\tau(t_{k+1}) &= u_\tau(t_k) + \tfrac{\tau}{6}(f_1^k + 4f_2^k + f_3^k), \\ f_1^k &= F(t_k, u_\tau(t_k)), \\ f_2^k &= F(t_k + \tfrac{\tau}{2}, u_\tau(t_k)) + \tfrac{\tau}{2} f_1^k), \\ f_3^k &= F(t_k + \tau, u_\tau(t_k)) - \tau f_1^k + 2\tau f_2^k). \end{aligned}$$

Konsistenzordnung $p = 3$.

$s = 4$

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array}$$

Standard Runge-Kutta Verfahren mit:

$$\begin{aligned} u_\tau(t_{k+1}) &= u_\tau(t_k) + \tfrac{\tau}{6}(f_1^k + 2f_2^k + 2f_3^k + f_4^k), \\ f_1^k &= F(t_k, u_\tau(t_k)), \\ f_2^k &= F(t_k + \tfrac{\tau}{2}, u_\tau(t_k)) + \tfrac{\tau}{2} f_1^k), \\ f_3^k &= F(t_k + \tfrac{\tau}{2}, u_\tau(t_k)) + \tfrac{\tau}{2} f_2^k), \\ f_4^k &= F(t_k + \tau, u_\tau(t_k)) - \tau f_3^k). \end{aligned}$$

Konsistenzordnung $p = 4$.

3.3 Mehrschrittverfahren für Anfangswertprobleme

Im vorangegangenen Abschnitt haben wir Einschrittverfahren betrachtet, die zur Berechnung einer numerischen Approximation der unbekannten Lösung $u_\tau(t_{k+1}) \in \mathbb{R}^n$ lediglich Informationen aus dem letzten Zeitschritt $t_k \in \Omega_\tau \subset [0, T]$ verwendet haben. Dies erinnert ein wenig an Markow-Prozesse aus der Stochastik, die nur auf Informationen des aktuellen Zustands zurückgreifen und ansonsten gedächtnislos sind. Da sich das Verhalten einer Lösungsfunktion u_τ nur schwer mit Hilfe eines einzigen Datenpunkts vorhersagen lässt, ist es sinnvoll auch die Werte der vorangehenden Zeitschritte zu berücksichtigen. Dieser Ansatz führt zu den sogenannten Mehrschrittverfahren für Anfangswertprobleme.

Konkret verwenden wir nun zur Berechnung der numerischen Approximation $u_\tau(t_{k+s})$ die bereits berechneten Werte der Zeitschritte $t_k, \ldots, t_{k+s-1} \in \Omega_\tau$ für eine Zeitdiskretisierung Ω_τ des Intervalls $[0, T]$. Hierbei bezeichnen wir mit $s \in \mathbb{N}^+$ die Stufe des Mehrschrittverfahrens. Die bereits in Abschnitt 3.2 diskutierten Einschrittverfahren lassen sich dementsprechend als Mehrschrittverfahren der Stufe $s = 1$ interpretieren. Wir werden in diesem Abschnitt die **abkürzende Notation** $F_k := F(t_k, u_\tau(t_k))$ verwenden.

Wir betrachten zunächst einige einfache Beispiele für Mehrschrittverfahren der Stufe $s = 2$.

BEISPIEL 3.28: Mehrschrittverfahren für Stufe $s = 2$.

Die beiden folgenden Beispiele erklären Mehrschrittverfahren der Stufe $s = 2$, d.h., wir verwenden zur Berechnung der numerischen Approximation $u_\tau(t_{k+2})$ die beiden vorigen Zeitschritte $t_k, t_{k+1} \in \Omega_\tau$. Die Grundidee ist es passende Quadraturformeln (siehe [Numerik 1, Kapitel 6]) für die Approximation des Integrals auf der rechten Seite folgender Gleichung zu benutzen:

$$u_\tau(t_{k+2}) - u_\tau(t_k) \;=\; \int_{t_k}^{t_{k+2}} u'_\tau(t)\,\mathrm{d}t \;=\; \int_{t_k}^{t_{k+2}} F(t, u_\tau(t))\,\mathrm{d}t.$$

1. Verwenden wir zunächst die simple **Mittelpunktsregel** zur Approximation des Integrals im Intervall $[t_k, t_{t+2}]$, so erhalten wir

$$\begin{aligned} u_\tau(t_{k+2}) - u_\tau(t_k) &= \int_{t_k}^{t_{k+2}} F(t, u_\tau(t))\,\mathrm{d}t \\ &\approx (t_{k+2} - t_k) \cdot F\left(\frac{t_{k+2} + t_k}{2}, u_\tau\left(\frac{t_{k+2} + t_k}{2}\right)\right) \\ &= 2\tau \cdot F(t_{k+1}, u_\tau(t_{k+1})). \end{aligned}$$

   Durch Umstellen erhalten wir so ein *explizites Verfahren* zur Berechnung des nächsten Werts $u_\tau(t_{k+2})$ der numerischen Approximation aus den letzten beiden Werten, nämlich

$$u_\tau(t_{k+2}) \;=\; u_\tau(t_k) + 2\tau \cdot F_{k+1}.$$

2. Verwenden wir stattdessen die **Simpsonregel** als interpolatorische Quadraturformel für die Approximation des Integrals im Intervall $[t_k, t_{t+2}]$, so erhalten wir

$$\begin{aligned} u_\tau(t_{k+2}) - u_\tau(t_k) &= \int_{t_k}^{t_{k+2}} F(t, u_\tau(t))\,\mathrm{d}t \\ &\approx \frac{t_{k+2} - t_k}{6} \cdot \Big(F(t_k, u_\tau(t_k)) \\ &\qquad + 4 \cdot F\left(\frac{t_{k+2} + t_k}{2}, u_\tau\left(\frac{t_{k+2} + t_k}{2}\right)\right) \\ &\qquad + F(t_{k+2}, u_\tau(t_{k+2}))\Big) \\ &= \frac{\tau}{3} \cdot (F(t_k, u_\tau(t_k)) + 4F(t_{k+1}, u_\tau(t_{k+1})) + F(t_{k+2}, u_\tau(t_{k+2}))). \end{aligned}$$

   Durch Umstellen erhalten wir so ein *implizites Verfahren* zur Berechnung des nächsten Werts $u_\tau(t_{k+2})$ der numerischen Approximation aus den letzten beiden Werten, nämlich

$$u_\tau(t_{k+2}) \;=\; u_\tau(t_k) + \frac{\tau}{3} \cdot (F_k + 4 \cdot F_{k+1} + F_{k+2}).$$

Die in Beispiel 3.28 illustrierten Mehrschrittverfahren sind beide von einer linearen Gestalt, die in folgender Definition verallgemeinert wird.

DEFINITION 3.29: Lineare Mehrschrittverfahren.

Sei eine Zeitdiskretisierung $\Omega_\tau := \{t_k \in [0,T] \colon t_k := k \cdot \tau, k = 0, \ldots, N\}$ für ein gegebenes Intervall $[0,T] \subset \mathbb{R}^+$ mit Zeitschrittweite $\tau := \frac{T}{N}$ für $N \in \mathbb{N}^+$ eines Anfangswertproblems der Form $u'(t) = F(t, u(t))$ für $t \in [0,T]$ mit $u(0) := u_0 \in \mathbb{R}^n$ gegeben.
Wir nennen ein Mehrschrittverfahren der Stufe $s \in \mathbb{N}^+$ **linear**, wenn es sich in folgender Form schreiben lässt:

$$\begin{aligned} \sum_{i=0}^{s} \alpha_i u_\tau(t_{k+i}) &= \alpha_0 u_\tau(t_k) + \alpha_1 u_\tau(t_{k+1}) + \ldots + \alpha_s u_\tau(t_{k+s}) \\ &= \tau \cdot (\beta_0 F_k + \beta_1 F_{k+1} + \ldots + \beta_s F_{k+s}) = \tau \cdot \sum_{i=0}^{s} \beta_i F_{k+i}, \end{aligned} \tag{3.18}$$

mit Koeffizienten $\alpha_i, \beta_i \in \mathbb{R}$ für $i = 0, \ldots, s$.
Wir nennen das lineare Mehrschrittverfahren **explizit** falls $\beta_s = 0$ gilt und ansonsten **implizit**.

Lineare Mehrschrittverfahren sind mit Abstand die gebräuchlichsten Methoden in der Numerik und daher werden wir uns im Folgenden auf diese Klasse von Verfahren einschränken. Damit wir wirklich ein s-stufiges lineares Mehrschrittverfahren vorliegen haben, werden wir immer annehmen, dass $\alpha_s \neq 0$ und $|\alpha_0| + |\beta_0| > 0$ in Definition 3.29 gilt. An der Form eines linearen Mehrschrittverfahrens in (3.18) sehen wir ebenfalls einen allgemeinen Vorteil gegenüber den Einschrittverfahren in Abschnitt 3.2: Wir können in jeder Iteration $k = s, \ldots, N - s$ zur Bestimmung des Werts $u_\tau(t_{k+s}) \in \mathbb{R}^n$ auf bereits berechnete Funktionsauswertungen $F_k, \ldots, F_{k+s-1} \in \mathbb{R}^n$ zurückgreifen und müssen nur eine einzige neue Funktionsauswertung $F_{k+s} \in \mathbb{R}^n$ durchführen.

Bei der numerischen Berechnung gehen wir analog wie bei Einschrittverfahren vor. Wir müssen nur zusätzlich zu $u_\tau(t_{k+s-1}) \in \mathbb{R}^n$ auch die Werte $u_\tau(t_k), \ldots, u_\tau(t_{k+s-2}) \in \mathbb{R}^n$ verwenden. Hierdurch ensteht ein effektiver Unterschied zu Einschrittverfahren mit Bezug auf die benötigten Anfangswerte. Um ein Mehrschrittverfahren der Stufe $s > 1$ durchzuführen benötigen wir nicht nur den gegebenen Anfangswert $u_0 \in \mathbb{R}^n$, sondern auch numerische Approximationen der folgenden $s - 1$ Werte $u_\tau(t_1), \ldots, u_\tau(t_{s-1})$ zu Berechnung des Wertes $u_\tau(t_s)$. Da diese numerischen Approximationen zunächst unbekannt sind müssen wir sie erst durch ein anderes Verfahren, etwa ein Einschrittverfahren, berechnen. Dabei müssen wir insbesondere darauf achten, dass das gewählte Verfahren von der selben Konvergenzordnung wie das Mehrschrittverfahren gewählt wird, um diese insgesamt nicht zu verkleinern. Eine Diskussion der Konsistenz- und Konvergenzordnung von Mehrschrittverfahren folgt in Abschnitt 3.3.1.

BEMERKUNG 3.30 (Wohldefiniertheit von Mehrschrittverfahren). Bei expliziten Verfahren, d.h. $\beta_s = 0$, ist die Wohldefiniertheit des Mehrschrittverfahrens natürlich gegeben, da eine numerische Approximation $u_\tau(t_{k+1}) \in \mathbb{R}^n$ explizit bestimmt werden kann. Im impliziten Fall, d.h. für $\beta_s \neq 0$, ist die Existenz einer Lösung des auftretenden nichtlinearen Gleichungssystems jedoch im Allgemeinen nicht gesichert. Hier benötigen wir wieder das in Abschnitt 3.1.1 diskutierte Fixpunktargument um die Wohldefiniertheit der beschriebenen Verfahren zu gewährleisten.
Es lässt sich zeigen, dass folgende Aussage gilt: Für eine stetige Funktion F des Anfangswertproblems $u'(t) = F(t, u(t)), u(0) = u_0 \in \mathbb{R}^n$, die bezüglich des zweiten Arguments Lipschitz-stetig mit Lipschitz-Konstante $L > 0$ ist, existiert eine eindeutige Lösung $u_\tau(t_{k+s}) \in \mathbb{R}^n$ des Mehrschrittverfahrens (3.18) falls für die Schrittweite $\tau < \frac{|\beta_s|}{|\alpha_s|} \cdot L$ gilt. △

3.3.1 Konsistenz von Mehrschrittverfahren

Während wir die Konvergenzordnung analog zu Einschrittverfahren definieren können, benötigen wir noch eine passende Definition von Konsistenz für Mehrschrittverfahren. Wir führen dazu zunächst den lokalen Konsistenzfehler eines Mehrschrittverfahrens ein

$$\begin{aligned} g_\tau(t_k) &:= \frac{1}{\tau}\sum_{i=0}^{s} \alpha_i u(t_{k+i}) - \sum_{i=0}^{s} \beta_i F(t_{k+i}, u(t_{k+i})) \\ &= \frac{1}{\tau}\sum_{i=0}^{s} \alpha_i u(t_{k+i}) - \sum_{i=0}^{s} \beta_i u'(t_{k+i}) \qquad k = 0, \dots, N-s. \end{aligned} \tag{3.19}$$

Hierbei haben wir analog zum Einschrittverfahren wieder die echte Lösung $u\colon [0,T] \to \mathbb{R}^n$ in das Mehrschrittverfahren eingesetzt. Basierend auf dem lokalen Konsistenzfehler (3.19) können wir nun definieren wann ein Mehrschrittverfahren konsistent ist.

DEFINITION 3.31: Konsistenz von linearen Mehrschrittverfahren.

Ein lineares Mehrschrittverfahren heißt **konsistent**, falls der globale Konsistenzfehler

$$K_\tau := \max_{k=0,\dots,N-s} ||g_\tau(t_k)||$$

gegen Null konvergiert für $\tau \to 0$.
Das Verfahren heißt konsistent von der Ordnung p, falls $K_\tau = \mathcal{O}(\tau^p)$ für $\tau \to 0$.

Wir können zunächst ein schönes Resultat zur Charakterisierung der Konsistenz eines linearen Mehrschrittverfahren herleiten.

LEMMA 3.32: Konsistenzbedingung für Mehrschrittverfahren.

Ein lineares Mehrschrittverfahren der Stufe $s \in \mathbb{N}^+$ ist konsistent von der Ordnung $p \in \mathbb{N}^+$, wenn die folgenden Gleichungen gelten

$$\sum_{i=0}^{s} \alpha_i i^m \;=\; m \cdot \sum_{i=0}^{s} \beta_i i^{m-1}, \qquad m = 0, 1, \ldots, p.$$

Hierbei verwenden wir die Konvention $0^0 := 1$.

Beweis. Es sei $s \in \mathbb{N}^+$ die Stufe des Mehrschrittverfahrens und $u\colon [0,T] \to \mathbb{R}^n$ die echte Lösung des Anfangswertproblems (3.1). Um zu zeigen, dass das Verfahren konsistent von der Ordnung $p \in \mathbb{N}^+$ ist, müssen wir gemäß Definition 3.31 zeigen, dass für den global Konsistenzfehler gilt:

$$K_\tau \;:=\; \max_{k=0,\ldots,N-s} \left\| \frac{1}{\tau} \sum_{i=0}^{s} \alpha_i u(t_{k+i}) - \sum_{i=0}^{s} \beta_i u'(t_{k+i}) \right\| \in \mathcal{O}(\tau^p).$$

Sei im weiteren $k \in \{0, \ldots, N-s\}$ der Index des Zeitschritts für den die Norm des lokalen Konsistenzfehlers $g_\tau(t_k)$ maximal wird. Wir betrachten zunächst die Taylorentwicklung des linken Terms im Zeitschritt $t_k \in \Omega_\tau$ wie folgt:

$$\begin{aligned}
\frac{1}{\tau} \sum_{i=0}^{s} \alpha_i u(t_{k+i}) \;=\; \frac{1}{\tau}\Big(& \alpha_0 u(t_k) + \alpha_1 (u(t_k) + \tau u'(t_k) + \frac{\tau^2}{2} u''(t_k) + \ldots) \\
& + \alpha_2 (u(t_k) + 2\tau u'(t_k) + \frac{(2\tau)^2}{2} u''(t_k) + \ldots) \\
& \vdots \\
& + \alpha_s (u(t_k) + s\tau u'(t_k) + \frac{(s\tau)^2}{2} u''(t_k) + \ldots) \Big)
\end{aligned}$$

Dies können wir zusammenfassen zu

$$\begin{aligned}
\frac{1}{\tau} \sum_{i=0}^{s} \alpha_i u(t_{k+i}) \;&=\; \frac{1}{\tau} \sum_{m=0}^{p} \sum_{i=0}^{s} \alpha_i \frac{i^m \tau^m}{m!} u^{(m)}(t_k) + \mathcal{O}(\tau^p) \\
&=\; \sum_{m=0}^{p} \sum_{i=0}^{s} \alpha_i i^m \frac{\tau^{m-1}}{m!} u^{(m)}(t_k) + \mathcal{O}(\tau^p)
\end{aligned}$$

Betrachten wir nun die Taylorentwicklung des rechten Terms im Zeitschritt $t_k \in \Omega_\tau$:

$$\begin{aligned}
\sum_{i=0}^{s} \beta_i u'(t_{k+i}) \;=\; & \beta_0 u'(t_k) + \beta_1 (u'(t_k) + \tau u''(t_k) + \frac{\tau^2}{2} u'''(t_k) + \ldots) \\
& + \beta_2 (u'(t_k) + 2\tau u''(t_k) + \frac{(2\tau)^2}{2} u'''(t_k) + \ldots) \\
& \vdots \\
& + \beta_s (u'(t_k) + s\tau u''(t_k) + \frac{(s\tau)^2}{2} u'''(t_k) + \ldots)
\end{aligned}$$

Dies können wir wiederum zusammenfassen zu

$$\begin{aligned}
\sum_{i=0}^{s} \beta_i u'(t_{k+i}) &= \sum_{m=0}^{p} \sum_{i=0}^{s} \beta_i \frac{i^m \tau^m}{m!} u^{(m+1)}(t_k) + \mathcal{O}(\tau^p) \\
&= \underbrace{0 \cdot \sum_{i=0}^{s} \beta_i i^{-1} \frac{\tau^0}{(0+1)!} u^0(t_k)}_{=0} \\
&\qquad + \sum_{m=0}^{p} (m+1) \cdot \sum_{i=0}^{s} \beta_i i^m \frac{\tau^m}{(m+1)!} u^{(m+1)}(t_k) + \mathcal{O}(\tau^p) \\
&= 0 \cdot \sum_{i=0}^{s} \beta_i i^{-1} \frac{\tau^0}{(0+1)!} u^0(t_k) \\
&\qquad + \sum_{m=1}^{p} m \cdot \sum_{i=0}^{s} \beta_i i^{m-1} \frac{\tau^{m-1}}{m!} u^{(m)}(t_k) + \mathcal{O}(\tau^p) \\
&= \sum_{m=0}^{p} m \cdot \sum_{i=0}^{s} \beta_i i^{m-1} \frac{\tau^{m-1}}{m!} u^{(m)}(t_k) + \mathcal{O}(\tau^p)
\end{aligned}$$

Insgesamt gilt also für den globalen Konsistenzfehler

$$K_\tau = \left\| \sum_{m=0}^{p} \sum_{i=0}^{s} \alpha_i i^m \frac{\tau^{m-1}}{m!} u^{(m)}(t_k) - \sum_{m=0}^{p} m \cdot \sum_{i=0}^{s} \beta_i i^{m-1} \frac{\tau^{m-1}}{m!} u^{(m)}(t_k) + \mathcal{O}(\tau^p) \right\|$$

Ein Koeffizientenvergleich der beiden Terme zeigt uns, dass $K_\tau \in \mathcal{O}(\tau^p)$ gilt wenn folgende Bedingungen erfüllt sind:

$$\sum_{i=0}^{s} \alpha_i i^m = m \cdot \sum_{i=0}^{s} \beta_i i^{m-1}, \qquad m = 0, \ldots, p.$$

□

Um eine möglichst hohe Konsistenzordnung für ein s-stufiges Mehrschrittverfahren zu erhalten müssen wir wir analog zu den Einschrittverfahren in Abschnitt 3.2 Gleichungssysteme für die Koeffizienten lösen, wie wir im folgenden Beispiel sehen werden.

BEISPIEL 3.33: Konsistenzordnung von Mehrschrittverfahren.

In diesem Beispiel wollen wir Mehrschrittverfahren der Stufen $s = 1$ und $s = 2$ bezüglich ihrer Konsistenzordnung $p \in \mathbb{N}^+$ untersuchen. Hierzu verwenden wir die Konsistenzbedingung aus Lemma 3.32 mit

$$\sum_{i=0}^{s} \alpha_i i^m = m \cdot \sum_{i=0}^{s} \beta_i i^{m-1}, \qquad m = 0, \ldots, p.$$

1. Im Fall eines Mehrschrittverfahrens der Stufe $\mathbf{s = 1}$ gilt für die erste Bedingung mit $m = 0$ und der Konvention $0^0 := 1$:

$$\alpha_0 \cdot 0^0 + \alpha_1 \cdot 1^0 \;=\; \alpha_0 + \alpha_1 \;\overset{!}{=}\; 0.$$

   Hieraus lässt sich also ableiten, dass $\alpha_0 = -\alpha_1$ gelten muss. Ohne Beschränkung der Allgemeinheit können wir $\alpha_1 = 1$ und $\alpha_0 = -1$ normieren. Betrachten wir nun die zweite Bedingung für $m = 1$, so erhalten wir:

$$\alpha_0 \cdot 0^1 + \alpha_1 \cdot 1^1 \;=\; \alpha_1 \;\overset{!}{=}\; 1 \cdot (\beta_0 \cdot 0^0 + \beta_1 \cdot 1^0) \;=\; \beta_0 + \beta_1.$$

   Durch die Normierung von $\alpha_1 = 1$ sehen wir also, dass $\beta_0 + \beta_1 = 1$ gelten muss, um Konsistenzordnung $p = s = 1$ zu erreichen. Diese Bedingungen werden unter anderem durch das *Vorwärts-Euler Verfahren* ($\beta_0 = 1, \beta_1 = 0$), das *Rückwärts-Euler Verfahren* ($\beta_0 = 0, \beta_1 = 1$) und das *Crank-Nicholson Verfahren* ($\beta_0 = \beta_1 = \frac{1}{2}$) aus Abschnitt 3.2 erfüllt.

2. Im Fall eines Mehrschrittverfahrens der Stufe $\mathbf{s = 2}$ gilt für die erste Bedingung mit $m = 0$ und der Konvention $0^0 := 1$:

$$\alpha_0 \cdot 0^0 + \alpha_1 \cdot 1^0 + \alpha \cdot 2^0 \;=\; \alpha_0 + \alpha_1 + \alpha_2 \;\overset{!}{=}\; 0.$$

   Für die zweite Bedingung mit $m = 1$ erhalten wir

$$\alpha_0 \cdot 0^1 + \alpha_1 \cdot 1^1 + \alpha \cdot 2^1 \;=\; \alpha_1 + 2\alpha_2 \;\overset{!}{=}\; 1 \cdot (\beta_0 \cdot 0^0 + \beta_1 \cdot 1^0 + \beta_2 \cdot 2^0) \;=\; \beta_0 + \beta_1 + \beta_2.$$

   Und für die dritte Bedingung mit $m = 2$ erhalten wir

$$\alpha_0 \cdot 0^2 + \alpha_1 \cdot 1^2 + \alpha \cdot 2^2 \;=\; \alpha_1 + 4\alpha_2 \;\overset{!}{=}\; 2 \cdot (\beta_0 \cdot 0^1 + \beta_1 \cdot 1^1 + \beta_2 \cdot 2^1) \;=\; 2\beta_1 + 4\beta_2.$$

   Insgesamt müssen wir also für ein Mehrschrittverfahren der Stufe $s = 2$ und Konsistenzordnung $p = 2$ folgendes unterbestimmtes Gleichungssystem lösen:

$$\alpha_0 + \alpha_1 + \alpha_2 \;=\; 0, \qquad \alpha_1 + 2\alpha_2 \;=\; \beta_0 + \beta_1 + \beta_2, \qquad \alpha_1 + 4\alpha_2 \;=\; 2\beta_1 + 4\beta_2.$$

   Wir sehen, dass die *Mittelpunktsregel* mit $\alpha_0 = -1, \alpha_1 = 0, \alpha_2 = 1$ sowie $\beta_0 = \beta_2 = 0, \beta_1 = 2$ eine mögliche Lösung dieses Systems liefert.

3.3.2 Stabilität von Mehrschrittverfahren

Nachdem wir die Konsistenzordnung von Mehrschrittverfahren untersucht haben stellt sich die Frage, ob wir die gleiche Konvergenzordnung erreichen können. Im Fall von Einschrittverfahren hatten wir festgestellt, dass für Lipschitz-stetige Funktionen F auf der rechten Seite des Anfangswertproblems (3.1) aus Konsistenz bereits Konvergenz folgt.

Leider ist die Situation für Mehrschrittverfahren deutlich komplizierter, wie das folgende Beispiel zeigt.

BEISPIEL 3.34: Instabilität eines Mehrschrittverfahrens.

Es werde ein Verfahren möglichst hoher Ordnung der Form

$$u_\tau(t_{k+2}) - (1+\alpha)u_\tau(t_{k+1}) + \alpha u_\tau(t_k) \;=\; \tau \cdot \left(\frac{3-\alpha}{2}F_{k+1} - \frac{1+\alpha}{2}F_k\right)$$

gesucht für eine Funktion $F \in C^3([0,T]\times\mathbb{R}^n;\mathbb{R}^n)$.
Wir bestimmen den Koeffizienten $\alpha \in \mathbb{R}$ mittels Taylorentwicklung so, dass eine möglichst hohe Konsistenzordnung erreicht wird indem wir wieder den lokalen Konsistenzfehler in $t_k \in \Omega_\tau$ betrachten:

$$\begin{aligned} g_\tau(t_k) &= \frac{1}{\tau}\cdot(u(t_{k+2}) - (1+\alpha)u(t_{k+1}) + \alpha u(t_k)) - \frac{3-a}{2}u'(t_{k+1}) - \frac{1+\alpha}{2}u'(t_k) \\ &= u'(t_k)\cdot\underbrace{\left(2-(1+\alpha)+\alpha-\frac{3-\alpha}{2}+\frac{1+\alpha}{2}\right)}_{=0} \\ &+ u''(t_k)\cdot\underbrace{\left(\frac{4\tau}{2}-(1+\alpha)\frac{\tau}{2}-\frac{3-\alpha}{2}\tau\right)}_{=0} \\ &+ u'''(t_k)\cdot\underbrace{\left(\frac{8}{6}\tau^2-(1+\alpha)\frac{\tau^2}{6}-\frac{3-\alpha}{2}\cdot\frac{\tau^2}{2}\right)}_{=\frac{\tau^2}{12}(5+\alpha)} + \mathcal{O}(\tau^3). \end{aligned}$$

Also erhalten wir eine Konsistenzordnung von $p=3$ für $\alpha=-5$ und $p=2$ in allen anderen Fällen. Wir erwarten eine entsprechende Konvergenzordnung für das Verfahren. Ein numerisches Experiment zeigt uns jedoch ein völlig anderes Ergebnis.
Wie man an den beiden Illustrationen in Abb. 3.1 erkennen kann ist das Verhalten des numerischen Algorithmus für unterschiedliche Wahlen von $\alpha \in \mathbb{R}$ vollkommen unterschiedlich. Bei näherer Betrachtung stellt sich heraus, dass das Mehrschrittverfahren für $\alpha > 1$ und $\alpha < -1.5$ divergiert trotz einer Konsistenzordnung von mindestens Zwei. Dies liegt daran, dass das lineare Mehrschrittverfahren in diesen Fällen nicht stabil ist.

Motiviert durch das obige Beispiel wollen wir uns also im Folgenden mit der Stabilitätsanalyse von Mehrschrittverfahren beschäftigten. Wir folgen dabei grob der Struktur in [NumDGL].

Die grundlegende Idee ist es nun die Stabilität eines linearen Mehrschrittverfahrens für eine möglichst einfache Differentialgleichung zu untersuchen. Wenn ein konsistentes Verfahren konvergent ist, dann muss es insbesondere konvergent sein für die einfachste

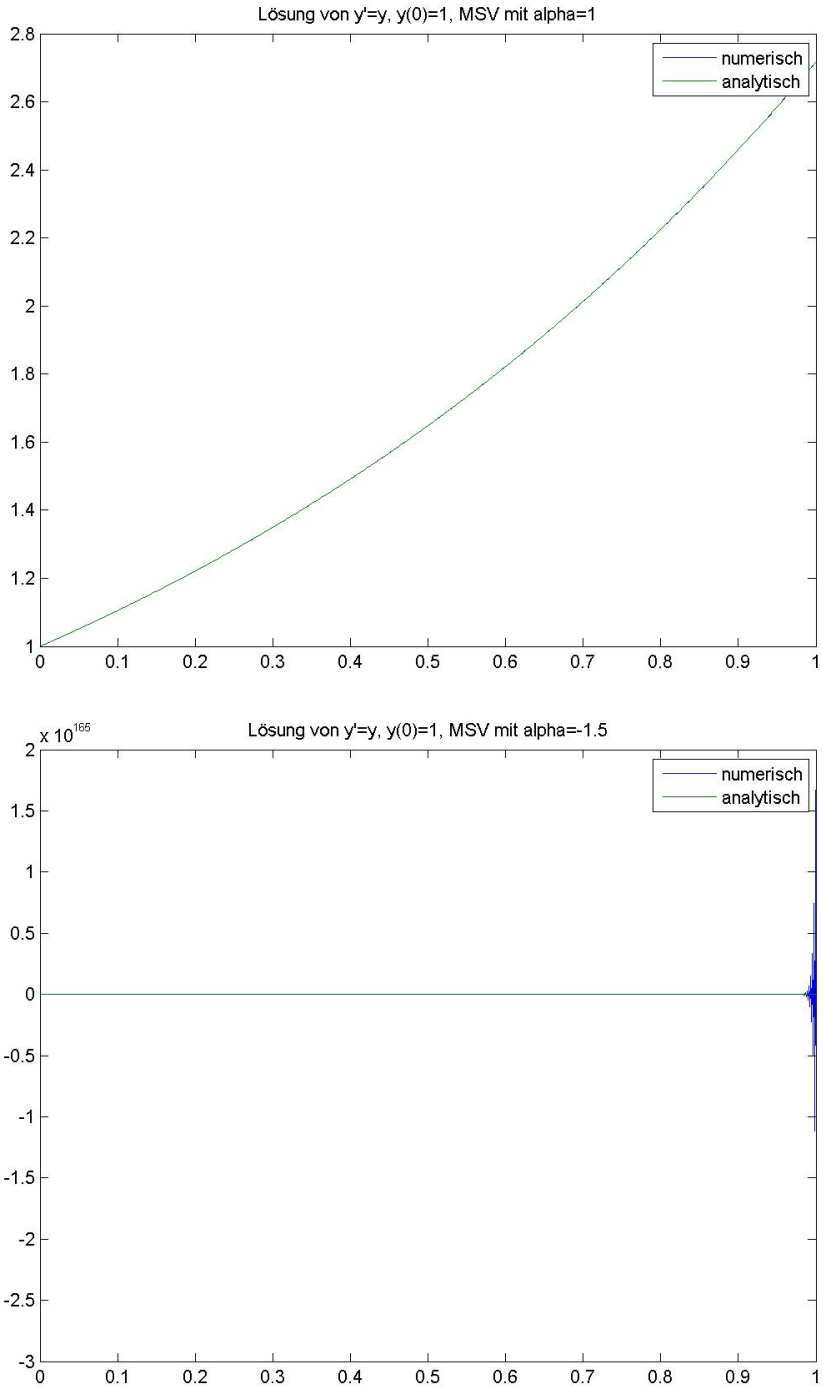

Abbildung 3.1: Visualisierung der numerischen Approximation einer Lösung eines Anfangswertproblems mit dem linearen Mehrschrittverfahren aus Beispiel 3.34 für die Wahl des Parameters $\alpha = 1$ (oben) und $\alpha = -1.5$ (unten). Entnommen aus [NumAna]

.

aller Anfangswertaufgaben. Daher betrachten wir das sogenannte **Modellproblem**

$$u'(t) \;=\; 0, \qquad u(0) \;=\; 0, \tag{3.20}$$

mit der Lösung $u(t) \equiv 0$. Die durch das numerische Verfahren gelieferte Lösung muss also für $\tau \to 0$ gleichmäßig gegen die Nullfunktion konvergieren. Das folgende Lemma besagt, dass diese Bedingung schon ausreicht, damit das numerische Verfahren für alle Anfangswertaufgaben konvergent ist.

LEMMA 3.35: Reduktion auf das Modellproblem.

Gegeben sei ein lineares Mehrschrittverfahren der Stufe $s \in \mathbb{N}^+$. Sei $(\Omega_{\tau_k})_{k\in\mathbb{N}}$ eine Folge von äquidistanten Zeitdiskretisierungen mit Zeitschrittweiten $\tau_k > 0, k \in \mathbb{N}$. Falls für jede Wahl der Startwerte $u_\tau(t_i) \in \mathbb{R}^n$ mit

$$u_{\tau_k}(t_i) \overset{\tau_k\to 0}{\longrightarrow} 0, \qquad i = 0, \dots, s-1$$

die Folge $(u_{\tau_k})_{k\in\mathbb{N}}$ der numerischen Approximationen für das Modellproblem gegen die Nullfunktion konvergiert, so ist das Mehrschrittverfahren stabil für alle Anfangswertaufgaben.

Beweis. Siehe [NumAna, Satz 5.20]. □

Wenn wir also nun ein lineares Mehrschrittverfahren für das Modellproblem (3.20) betrachten, so stellen wir fest, dass für die rechte Seite der gewöhnlichen Differentialgleichung $F(t, u(t)) \equiv 0$ gilt und somit für das Mehrschrittverfahren $\sum_{i=0}^{s} \beta_i F_{k+i} = 0$ ist für alle $k = 0, \dots, N-s$. Somit reduziert sich das lineare Mehrschrittverfahren mit gegebenen Koeffizienten $\alpha_i \in \mathbb{R}, i = 0, \dots, s$ auf die Lösung des folgenden Gleichungssystems:

$$\sum_{i=0}^{s} \alpha_i u_\tau(t_{k+i}) \;=\; 0, \qquad k = 0, \dots, N-s. \tag{3.21}$$

Ein Problem dieser Form ist bekannt als **lineare, homogene Differenzengleichung** mit konstanten Koeffizienten $\alpha_i \in \mathbb{R}, i = 0, \dots, s$.

Aus diesem Grund werden wir uns im Folgenden auf einige mathematische Resultate aus der *Theorie der linearen Differenzengleichungen* stützen, die wir in dieser Vorlesung jedoch nur im Ansatz diskutieren können. Wir wollen deswegen mit einer grundlegenden Definition beginnen.

DEFINITION 3.36: Lineare Differenzengleichung.

Eine **lineare Differenzengleichung s-ter Ordnung** über einem Körper $\mathbb{K}$ ist von der Form

$$\sum_{i=0}^{s} \alpha_i(k) f_{k+i} \;=\; \beta_k, \qquad k \in \mathbb{N}, \tag{3.22}$$

wobei die Koeffizienten $\alpha_i \colon \mathbb{N} \to \mathbb{K}, i = 0, \ldots, s$ Folgen sind, für die $\alpha_s(k) \neq 0$ gilt für alle $k \in \mathbb{N}$. Ist $(\beta_k)_{k\in\mathbb{N}} \subset \mathbb{K}$ die konstante Nullfolge, so nennen wir die Differenzengleichung **homogen** und ansonsten **inhomogen**.
Eine unendliche Zahlenfolge $F = f_0, f_1, f_2, \ldots$, die für alle $k \in \mathbb{N}$ die lineare Differenzengleichung (3.22) erfüllt, heißt **Lösung der Differenzengleichung**.

Die folgende Bemerkung erlaubt es uns eine lineare Differenzenfolge in eine explizite Form zu überführen, so dass Folgeglieder der Lösung aus gegebenen Anfangswert berechnet werden können.

BEMERKUNG 3.37 (Explizite Form von Differenzengleichungen). Es ist klar, dass man eine lineare Differenzengleichung mit einem beliebigen Faktor aus $\mathbb{K}$ multiplizieren kann ohne dessen Lösung zu verändern. Wenn man also die Koeffizientenfolgen $(\alpha_i(k))_{k\in\mathbb{N}}$ normiert, so dass $\alpha_s(k) = -1$ für alle $k \in \mathbb{N}$ gilt, so lässt sich eine Rechenvorschrift für das Folgenglied f_{k+s} explizit angeben als

$$f_{k+s} \;=\; \sum_{i=0}^{s-1} \alpha_i(k) f_{k+i} - \beta(k).$$

Für gegebene Startwerte $f_0, \ldots, f_{s-1} \in \mathbb{K}^n$ und Koeffizientenfolgen $(\alpha_i)_{k\in\mathbb{N}}, (\beta_k)_{k\in\mathbb{N}}$ lässt sich so eine Lösung F der Differenzengleichung bestimmen.. △

Die explizite Form einer linearen Differenzengleichung wird zur Konstruktion von bestimmten Folgen in der Mathematik genutzt, wie unter anderem im folgenden Beispiel erklärt ist.

BEISPIEL 3.38: Fibonnaci Zahlen.

Ein berühmtes Beispiel einer linearen Differenzengleichung 2. Ordnung wird durch die *Fibonacci-Folge* gelöst, deren Folgenglieder sich berechnen lassen durch die Differenzengleichung

$$f_{k+2} \;=\; f_k + f_{k+1} \qquad \Leftrightarrow \qquad f_{k+2} - f_{k+1} - f_k \;=\; 0. \tag{3.23}$$

Für die typischerweise gewählten Startwerte $f_0 = f_1 := 1$ ergibt sich hieraus die bekannte Fibonacci-Folge als Lösung der linearen Differenzengleichung mit:

$$F \;=\; 1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots$$

Wir erkennen, dass die Differenzengleichung (3.23) homogen ist und konstante Koeffizientenfolgen $\alpha_0 = \alpha_1 \equiv -1$ und $\alpha_2 \equiv 1$ besitzt.

Wenn wir zeigen können, dass Lösungen der linearen Differenzengleichung (3.21) für alle Startwerte $f_0, \ldots, f_{s-1} \in \mathbb{K}^n$ beschränkt sind, so lässt sich zeigen, dass der globale Konsistenzfehler K_τ für das Modellproblem (3.20) gegen Null konvergiert und das lineare

Mehrschrittverfahren nach Lemma 3.35 stabil ist für alle Anfangswertaufgaben. Um die Beschränktheit von Lösungen der linearen Differenzengleichung (3.21) zu untersuchen führen wir zunächst ein weiteres Hilfsmittel ein.

DEFINITION 3.39: Charakteristisches Polynom.

Für eine homogene, lineare Differenzengleichung der Ordnung $s \in \mathbb{N}^+$ mit konstanten Koeffizienten $\alpha_i \in \mathbb{K}, i = 0, \dots, s$ der Form (3.21) definieren wir das zugehörige **charakteristische Polynom der Differenzengleichung** $\rho\colon \mathbb{K} \to \mathbb{K}$ als:

$$\rho(x) \ := \ \sum_{i=0}^{s} \alpha_i x^i.$$

Das charakteristische Polynom einer Differenzengleichung sagt viel über das Verhalten von zugehörigen Lösungen aus und dient unter Anderem dazu ihre Beschränktheit zu zeigen, wie folgendes bekanntes Theorem aussagt.

THEOREM 3.40: Wurzelbedingung von Dahlquist.

Alle Lösungen einer homogenen linearen Differenzengleichung mit charakteristischem Polynom ρ sind beschränkt genau dann, wenn für die Nullstellen $\lambda_i \in \mathbb{K}, i = 0, \dots, s-1$ von ρ folgende zwei Bedingungen erfüllt sind:

1. $|\lambda_i| \ \leq \ 1,$
2. $|\lambda_i| \ = \ 1 \qquad \Rightarrow \qquad \sigma_i \ = \ 1,$

wobei $\sigma_i \in \mathbb{N}^+$ die Vielfachheit der Nullstelle $\lambda_i \in \mathbb{K}$ ist.

Beweis. Siehe [NumAna, Satz 5.22]. □

Mit Hilfe der Wurzelbedingung von Dahlquist haben wir nun ein überprüfbares Kriterium um die Stabilität von linearen Mehrschrittverfahren zu überprüfen.

KOROLLAR 3.41: Konvergenz von linearen Mehrschrittverfahren.

Falls ein lineares Mehrschrittverfahren konsistent ist von der Ordnung $p \in \mathbb{N}^+$ und die zugehörige lineare Differenzengleichung die Wurzelbedingung von Dahlquist in Theorem 3.40 erfüllt, so ist das Mehrschrittverfahren konvergent von der Ordnung $p \in \mathbb{N}^+$.

Es bleibt immer noch die Frage offen welche Konvergenzordnung wir wirklich erreichen können, d.h. welche Konsistenzordnung ein stabiles Verfahren maximal erreichen kann. Es gibt hierbei tatsächlich eine Einschränkung, wie folgende Bemerkung festhält.

BEMERKUNG 3.42. Ein lineares Mehrschrittverfahren der Stufe $s \in \mathbb{N}^+$ sei konsistent von der Ordnung $p \in \mathbb{N}^+$ und stabil. Dann gelten die folgenden Beschränkungen an die Konsistenzordnung des Verfahrens:

1. $p \leq s + 2$, wenn s gerade ist,
2. $p \leq s + 1$, wenn s ungerade.

Ist $\frac{\beta_s}{\alpha_s} \leq 0$ (also insbesondere bei impliziten Verfahren) dann gilt sogar nur $p \leq s$. △

Wir wollen abschließend das Wurzelkriterium von Dahlquist anwenden, um die Stabiliät einiger Mehrschrittverfahren im Folgenden zu untersuchen.

BEISPIEL 3.43: Stabilität von linearen Mehrschrittverfahren.

Wir werden im Folgenden die Stabilität von linearen Mehrschrittverfahren durch Betrachtung der Nullstellen der assoziierten charakteristischen Polynome analysieren.

1. Wir können Einschrittverfahren als Mehrschrittverfahren mit $s = 1$ interpretieren. Sie sind von der Form

$$u_\tau(t_{k+1}) - u_\tau(t_k) \;=\; f_\tau(t_k, u_\tau(t_k)),$$

und damit ist das charakteristische Polynom der zugehörigen homogenen, linearen Differenzengleichung gegeben durch

$$\rho(x) \;=\; x - 1.$$

Die einzige (einfache) Nullstelle von ρ ist offensichtlich $\lambda_0 = 1$. Also sind Einschrittverfahren stabil, was zu unseren Beobachtungen aus Abschnitt 3.2 passt.

2. Wir betrachten als Nächstes ein lineares Mehrschrittverfahren das gegeben ist durch:

$$u_\tau(t_{k+2}) - 2u_\tau(t_{k+1}) + u_\tau(t_k) \;=\; 0.$$

Unabhängig von der rechten Seite $F(t, u(t))$ der Anfangswertaufgabe, die es zu lösen gilt, sehen wir ein, dass das Mehrschrittverfahren konsistent ist nach Beispiel 3.33 oben. Betrachten wir also das charakteristische Polynom der zugehörigen Differenzengleichung, das gegeben ist durch

$$\rho(x) \;=\; x^2 - 2x + 1.$$

Dieses Polynom besitzt eine doppelte Nullstelle $\lambda_0 = \lambda_1 = 1$ und somit ist die Wurzelbedingung von Dahlquist in Theorem 3.40 verletzt. Dies bedeutet, dass wir unabhängig von der Wahl der $\beta_i \in \mathbb{R}, i = 0, 1, 2$ des linearen Mehrschrittverfahrens keine Stabilität erwarten können.

3. Alle aus der numerischen Integration gewonnenen linearen Mehrschrittverfahren haben für $s \in \mathbb{N}^+$ und $k, r \in \mathbb{N}$ mit $0 \leq r \leq s$ die Form

$$u_\tau(t_{k+s}) - u_\tau(t_{k+s-r}) \;=\; \int_{k+s-r}^{k+s} F(t, u(t))\,\mathrm{d}t \;\approx\; \tau \sum_{i=0}^{r} \beta_i F_{k+s-r+i}.$$

Daher ist das charakteristische Polynom der zugeordneten linearen Differenzengleichung gegeben durch

$$\rho(x) \;=\; x^s - x^{s-r} \;=\; x^{s-r} \cdot (x^r - 1).$$

Dieses Polynom hat $(s-r)$-fache Nullstelle $\lambda = 0$ und zusätzlich die r-ten Einheitswurzeln in $\mathbb{C}$, die alle die Vielfachheit 1 haben. Darum sind alle linearen Mehrschrittverfahren die durch numerische Integration gewonnen werden stabil.

4. Für das lineare Mehrschrittverfahren in Beispiel 3.34 lässt sich das charakteristische Polynom der zugeordneten linearen Differenzengleichung angegeben als

$$\rho(x) \;=\; x^2 - (1+\alpha)x + \alpha.$$

Dieses Polynom besitzt die beiden Nullstellen $\lambda_0 = 1$ und $\lambda_1 = \alpha$. Nach der Wurzelbedingung von Dahlquist müssen wir also für die Stabilität des linearen Mehrschrittverfahrens fordern, dass gilt

$$-1 \;\leq\; \alpha \;<\; 1.$$

Wir haben hierbei den Fall $\alpha = 1$ herausgenommen, da wir sonst eine doppelte Nullstelle erhalten würden, was zu einer Verletzung der Stabilitätsbedingungen führt.

3.4 Weiterführende Themen

Im Folgenden diskutieren wir noch einige weiterführende Themen und Anwendungen zur Numerik von Einschrittverfahren. Dabei beginnen wir zunächst mit einfachen partiellen Differentialgleichungen und gehen dann zur Verbindung zwischen Optimierung und Differentialgleichungen über.

3.4.1 Lineare Transportgleichung

Wir betrachten im Folgenden eine *lineare Transportgleichung*, die unter Anderem verwendet wird um die Ausbreitung eines Stoffes oder eines Zustands zu modellieren. Wir betrachten dieses Modell der Einfachheit halber nur in einer räumlichen Dimension mit $n = 1$ auf dem Diskretisierungsgitter $\Omega_h = h \cdot \mathbb{Z}$ und einer örtlichen Schrittweite $h > 0$. Die Zustandsvariable $u_k(t) \in \mathbb{R}$ beschreibt einen Zustand im Punkt $k \cdot h \in \Omega_h$ zum

Zeitpunkt $t \in [0,T]$. Wir wählen nun ebenfalls eine äquidistante Diskretisierung des Zeitintervalls $[0,T]$ mit fester Zeitschrittweite $\tau > 0$.

Nehmen wir nun an, dass der Zustand mit einer konstanten, positiven Geschwindigkeit $v := \frac{h}{\tau} > 0$, d.h., genau die Breite einer örtliche Gitterzelle h pro Zeitschritt τ transportiert wird, so gilt offensichtlich

$$u_k(t+\tau) \;=\; u_{k-1}(t).$$

Dies können wir auch durch Erweiterung schreiben als

$$u_k(t+\tau) \;=\; \underbrace{u_k(t) - u_k(t)}_{=0} + \underbrace{\frac{\tau}{h} v}_{=1} \cdot u_{k-1}(t) \;=\; u_k(t) - \tau \frac{v}{h} \cdot (u_k(t) - u_{k-1}(t)). \tag{3.24}$$

Diese Darstellung können wir als *Vorwärts-Euler Verfahren* der folgenden gewöhnlichen Differentialgleichung mit einer konstanten, positiven Geschwindigkeit $v \in \mathbb{R}^+$ interpretieren:

$$u_k'(t) \;=\; -\frac{v}{h} \cdot (u_k(t) - u_{k-1}(t)). \tag{3.25}$$

Die rechte Seite hat eine Lipschitz-Konstante der Ordnung $\mathcal{O}(\frac{1}{h})$, welche beliebig groß werden kann für eine immer feinere örtliche Auflösung, d.h. für $h \to 0$.

Nach (3.24) gilt

$$u_k(t+\tau) \;=\; (1 - \tau\frac{v}{h}) \cdot u_k(t) + \tau \frac{v}{h} \cdot u_{k-1}(t),$$

und wir sehen, dass wir die Stabilität des Einschrittverfahrens zeigen können solange $\tau \cdot \frac{v}{h} \leq 1$ gilt. Dann gilt nämlich per Dreiecksungleichung

$$|u_k(t+\tau)| \;\leq\; (1 - \tau\frac{v}{h}) \cdot |u_k(t)| + \tau \frac{v}{h} \cdot |u_{k-1}(t)| \;\leq\; \max\{|u_k(t)|, |u_{k-1}(t)|\}.$$

Daraus folgt sofort eine Abschätzung für alle örtlichen Gitterpunkte mit

$$\|u(t+\tau)\|_\infty \;\leq\; \|u(t)\|_\infty.$$

Alternativ können wir ebenfalls das *Rückwärts-Euler Verfahren* für die gewöhnliche Differentialgleichung (3.25) betrachten

$$u_k(t+\tau) \;=\; u_k(t) - \tau \frac{v}{h} \cdot (u_k(t+\tau) - u_{k-1}(t+\tau)).$$

Wir wollen nun ebenfalls die Stabilität dieses Einschrittverfahrens näher untersuchen. Der Einfachheit halber betrachten wir nur Lösungen mit $u_k(0) = 0$ für $k \leq 0$. Es ist klar, dass diese Bedingung dann auch für alle $t > 0$ gilt. Dies ermöglicht es uns ein gestaffeltes Gleichungssystem für das implizite Einschrittverfahren herzuleiten und zu lösen. Es gilt im ersten Ortspunkt zu beliebiger Zeit $t \in [0,T]$

$$\begin{aligned}
& u_1(t+\tau) \;=\; u_1(t) - \frac{\tau v}{h} \cdot u_1(t+\tau) - \frac{\tau v}{h} \cdot \underbrace{u_0(t+\tau)}_{=0} \\
\Leftrightarrow \quad & \left(1 + \frac{\tau v}{h}\right) \cdot u_1(t+\tau) \;=\; u_1(t) \\
\Leftrightarrow \quad & u_1(t+\tau) \;=\; \frac{h}{h + v\tau} \cdot u_1(t).
\end{aligned}$$

Für beliebige Punkte $k \cdot h \in \Omega_h$ mit $k > 1$ gilt dann

$$\begin{aligned} & u_k(t+\tau) \;=\; u_k(t) - \frac{\tau v}{h} \cdot u_k(t+\tau) + \frac{\tau v}{h} \cdot u_{k-1}(t+\tau) \\ \Leftrightarrow \quad & \left(1 + \frac{\tau v}{h}\right) \cdot u_k(t+\tau) \;=\; u_k(t) + \frac{\tau v}{h} \cdot u_{k-1}(t+\tau) \\ \Leftrightarrow \quad & u_k(t+\tau) \;=\; \frac{h}{h+v\tau} u_k(t) + \frac{v\tau}{h+v\tau} u_{k-1}(t+\tau). \end{aligned}$$

Wir sehen also direkt, dass wir mittels Dreiecksgleichung folgende Abschätzung treffen können

$$|u_k(t+\tau)| \;\leq\; \max\{|u_k(t)|, |u_{k-1}(t+\tau)|\}.$$

Somit folgt induktiv

$$|u_k(t+\tau)| \;\leq\; \max_{i \leq k} |u_i(t)|.$$

Die impliziert wieder insbesondere die Stabilitätsabschätzung für alle örtlichen Gitterpunkte mit

$$\|u(t+\tau)\|_\infty \;\leq\; \|u(t)\|_\infty.$$

Man beachte, dass in diesem Fall das Einschrittverfahren stabil ist ohne jegliche Beschränkung an die Zeitschrittweite $\tau > 0$.

Für $h \to 0$ sehen wir, dass

$$v \cdot \frac{u(k \cdot h, t) - u((k-1) \cdot h, t)}{h} \;\longrightarrow\; v \cdot \partial_x u(x,t)$$

gilt und wir eigentlich eine *partielle Differentialgleichung* approximiert haben, nämlich die sogenannte **lineare Transportgleichung** der Form

$$\partial_t u(x,t) \;=\; -v \cdot \partial_x u(x,t),$$

mit konstanter, positiver Geschwindigkeit $v \in \mathbb{R}^+$.

Im obigen Fall haben wir also die partielle Ableitung in die Ortskoordinate x durch einen Rückwärtsdifferenzenquotienten approximiert. Analog könnten wir auch ein Verfahren mit Vorwärtsdifferenzenquotienten bezüglich der Ortskoordinate x aufschreiben, was zu folgender gewöhnlichen Differentialgleichungen führt:

$$u_j'(t) \;=\; \frac{v}{h} \cdot (u_j(t) - u_{j+1}(t)).$$

Unabhängig von der Zeitdiskretisierung ist hier allerdings das Differentialgleichungssystem schon instabil, wie man formal zeigen kann. Der Grund hierfür liegt anschaulich betrachtet in der ursprünglichen Motivation der Transportgleichung. Mit positiver Geschwindigkeit $v \in \mathbb{R}^+$ beschreibt die Transportgleichung die Ausbreitung eines Zustands in Richtung steigender Werte der Ortskoordinate x. Hierzu benötigt man ausschließlich Informationen von der linken Seite eines örtlichen Punkts der Diskretisierung Ω_h. Dies wird durch den Rückwärtsdifferenzenquotienten korrekt abgebildet, während der Vorwärtsdifferenzenquotient genau die entgegengesetzte Richtung verwendet.

3.4.2 Diffusionsgleichung

Wir betrachten im Folgenden einen einfachen Diffusionsmodell im eindimensionalen Raum für $n = 1$. Solche Modelle beschreiben beispielsweise in der Physik die Verteilung eines Stoffes in einem Medium oder die Ausbreitung von Wärme in einem Material. Wir modellieren ein Teilchen, dass einen Sprungprozess auf dem Gitter $\Omega_h := h \cdot \mathbb{Z} \cap [0,1]$ mit Schrittweite $h := \frac{1}{N} > 0, N \in \mathbb{N}$ und periodischen Randbedingungen durchführt, wobei es zu jedem Zeitpunkt $t \in [0,T]$ mit gleicher Wahrscheinlichkeit $\alpha \in (0, \frac{1}{2})$ nach links oder rechts springen kann. In diesem Zusammenhang nennt man α einen Diffusionskoeffizienten für den Sprungprozess. Die Wahrscheinlichkeit für einen Sprung in einem kleinen Zeitintervall $[t, t+\tau] \subset [0,T]$ mit Zeitschrittweite $0 < \tau < 1$ ist dementsprechend $2\alpha\tau \in (0,1)$. Dann gilt für die Wahrscheinlichkeit $p_k(t) \in [0,1]$, dass das Teilchen zur Zeit t im Gitterpunkt $k \cdot h \in \Omega_h$ ist:

$$p_k(t+\tau) \;=\; \alpha\tau \cdot p_{k-1}(t) + \alpha\tau \cdot p_{k+1}(t) + (1-2\alpha\tau) \cdot p_k(t), \qquad k = 0, \dots, N, \quad (3.26)$$

wobei wegen der periodischen Randbedingungen $p_{-1}(t) = p_N(t)$ und $p_{N+1}(t) = p_0(t)$ gilt.

Für immer kleiner werdende Zeitschrittweiten $\tau \to 0$ erhalten wir entsprechend im Grenzwert das Differentialgleichungssystem

$$p_k'(t) \;=\; \alpha \cdot (p_{k-1}(t) + p_{k+1}(t) - 2p_k(t)), \qquad k = 0, \dots, N. \quad (3.27)$$

In diesem Zusammenhang sehen wir ein, dass (3.26) als Vorwärts-Euler Verfahren zur numerischen Approximation der gewöhnlichen Differentialgleichung (3.27) interpretiert werden kann. Wir erkennen sofort, dass dieses Einschrittverfahren stabil ist für $2\alpha\tau < 1$. Dies ist potentiell eine starke Einschränkung an die Zeitschrittweite τ in Fällen in denen der Diffusionskoeffizient α groß ist.

Insbesondere können wir mit der Hilfsvariable $D := \alpha \cdot h^2$ das Einschrittverfahren wieder als Ortsdiskretisierung einer partiellen Differentialgleichung mittels eins Differenzenquotient zweiter Ordnung der folgenden Form interpretieren

$$\partial_t p(x,t) \;=\; D \cdot \partial_{xx} p(x,t).$$

Hierbei approximiert man die zweite Ableitung nach der Ortskoordinate durch

$$\partial_{xx} p(k \cdot h, t) \;\approx\; \frac{p_{k-1}(t) - 2p_k(t) + p_{k+1}(t)}{h^2}.$$

Somit erhalten wir Stabilität des Einschrittverfahrens für $\tau \sim h^2$, was für sehr kleines $h > 0$ (also für eine sehr feine Ortsauflösung) einen hohen numerischen Aufwand fordert.

Verwenden wir hingegen das *Rückwarts-Euler Verfahren* zur Zeitdiskretisierung von (3.27), so erhalten wir für $k = 0, \dots, N$

$$p_k(t+\tau) \;=\; p_k(t) + \alpha\tau \cdot p_{k-1}(t+\tau) + \alpha\tau \cdot p_{k+1}(t+\tau) - 2\alpha\tau \cdot p_k(t+\tau). \quad (3.28)$$

Es stellt sich heraus, dass dieses implizite Einschrittverfahren wiederum stabil ist ohne Schranke an die Zeitschrittweite $\tau > 0$. Dies sehen wir folgendermaßen: Sei im Folgenden

$P(t) := (p_0(t), \ldots, p_N(t))^T \in \mathbb{R}^{N+1}$, dann können wir das Einschrittverfahren (3.28) kompakt schreiben als

$$(I + \alpha\tau \cdot B) \cdot P(t+\tau) \;=\; P(t), \tag{3.29}$$

mit

$$B = \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & -1 \\ -1 & 2 & -1 & \ldots & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 \\ -1 & 0 & 0 & \ldots & -1 & 2 \end{pmatrix}.$$

Wie man leicht zeigt, gilt für jeden Vektor $x \in \mathbb{R}^{N+1}$

$$\begin{aligned} x^T Bx \;&=\; x^T \cdot \begin{pmatrix} 2x_0 - x_1 - x_N \\ -x_0 + 2x_1 - x_2 \\ \vdots \\ -x_{k-1} + 2x_k - x_{k+1} \\ \vdots \\ -x_0 - x_{N-1} + 2x_N \end{pmatrix} \\ &=\; x_0 \cdot (2x_0 - x_1 - x_N) \\ &\quad + x_1 \cdot (-x_0 + 2x_1 - x_2) \\ &\quad + \ldots \\ &\quad + x_N \cdot (-x_0 - x_{N-1} + 2x_N) \\ &=\; \sum_{i=0}^{N} (x_{i+1} - x_i)^2 \;\geq\; 0. \end{aligned}$$

Damit haben wir gezeigt, dass die Matrix $B \in \mathbb{R}^{(N+1)\times(N+1)}$ positiv semidefinit ist. Dies hätten wir auch mit Hilfe der Gerschgorin-Kreise (vgl. [Numerik 1, Kapitel 7.2]) sehen können, da wir an der Hauptdiagonalen von B und der Summe der Nebendiagonalen ablesen können, dass alle Eigenwerte der Matrix B in der Menge $B_2(2) := \{x \in \mathbb{R} : |x-2| \leq 2\}$ liegen müssen. Da B also positiv semidefinit ist, ist $(I+\tau B) \in \mathbb{R}^{(N+1)\times(N+1)}$ für jedes $\tau > 0$ positiv definit und invertierbar. Insbesondere erhalten wir folgende Stabilitätsabschätzung indem wir (3.29) von links mit dem Vektor $P(t+\tau)^T$ multiplizieren:

$$\begin{aligned} \|P(t+\tau)\|_2^2 + \alpha\tau \cdot P(t+\tau)^T BP(t+\tau) \;&=\; P(t+\tau)^T P(t) \\ &\leq\; \frac{1}{2}\|P(t+\tau)\|_2^2 + \frac{1}{2}\|P(t)\|_2^2. \end{aligned}$$

Die Ungleichung folgt aus der Einsicht, dass gilt

$$0 \;\leq\; \frac{1}{2}(P(t+\tau) - P(t))^2 \;=\; \frac{1}{2}||P(t+\tau)||^2 - P(t+\tau)^T P(t) + \frac{1}{2}||P(t)||^2.$$

Wegen der positiven Semidefinitheit von B folgt dann

$$\begin{aligned} \|P(t+\tau)\|_2^2 &\leq \frac{1}{2}\|P(t+\tau)\|_2^2 + \frac{1}{2}\|P(t)\|_2^2 \\ \Leftrightarrow \quad \frac{1}{2}\|P(t+\tau)\|_2^2 &\leq \frac{1}{2}\|P(t)\|_2^2. \end{aligned}$$

Somit können wir nun induktiv folgende Ungleichungskette folgern:

$$\|P(t+\tau)\|_2 \leq \|P(t)\|_2 \leq \ldots \leq \|P(0)\|_2.$$

Tatsächlich gilt in diesem Fall sogar folgende Abschätzung

$$\min_{k=0,\ldots,N} p_k(0) \leq \min_{k=0,\ldots,N} p_k(t) \leq \max_{k=0,\ldots,N} p_k(t) \leq \max_{k=0,\ldots,N} p_k(0).$$

Abschätzungen dieser Form nennt man im Kontext von partiellen Differentialgleichungen **Minimums-** und **Maximumsprinzip** und wir werden ähnliche Argumente im nächsten Kapitel zu Randwertaufgaben näher diskutieren.

Adjungierte Methode

Nehmen wir nun an, dass wir uns einen Startpunkt $x_{k_0} = k_0 \cdot h \in \Omega_h$ des modellierten Teilchens vorgeben, d.h., die Wahrscheinlichkeit für ein Auftreten des Teilchens ist zu Anfang $p_{k_0}(0) = 1$ und $p_i(t) = 0$ für alle $i \neq k$. Dann lässt sich mit Hilfe einer numerischen Lösung der Differentialgleichung (3.27) im Intervall $t \in (0, T)$ effizient berechnen was die Auftrittswahrscheinlichkeit $p_k(T)$ des Teilches in allen Punkten $k = 0, \ldots, N$ für den Zeitpunkt T ist.

Schwieriger ist jedoch die Beantwortung der umgekehrten Frage. Man könnte sich die Frage stellen, ob man für einen spezifischen Punkt $x_{k_0} = k_0 \cdot h \in \Omega_h$ die Wahrscheinlichkeit berechnen kann, dass das Teilchen zum Zeitpunkt T dort auftritt, d.h., wir interessieren uns für $p_{k_0}(T)$ für alle möglichen Startpunkte $k = 0, \ldots, N$ des Teilchens. Hierzu müssten wir eigentlich das Differentialgleichungssystem mit allen $N + 1$ Startpunkten des Teilchens lösen und die Lösung dann in dem spezifischen Punkt $p_{k_0}(T)$ auswerten.

Eine alternative Berechnungsmöglichkeit für die letzte Fragestellung ist die sogenannte **adjungierte Methode**, welche aus der Theorie der optimalen Steuerung stammt. Dazu berechnen wir die Lösung des adjungierten Problems, welches gegeben ist durch

$$q_k'(t) = \alpha \cdot (2q_k(t) - q_{k+1}(t) - q_{k-1}(t)), \qquad k = 0, \ldots, N,$$

mit vorgegebenem Endwert $q_{k_0}(T) = 1$ und $q_i(T) = 0$ für $i \neq k_0$.

Die Lösung dieses Problems ist genauso zu berechnen wie für das ursprüngliche System (3.27). Dies sehen wir mit der einfachen Variablentransformation $s := T - t$ ein, denn dann haben wir ein Anfangswertproblem mit den gleichen Vorzeichen wie das System für die Funktionen $p_k(t)$ zu lösen. Hat man die eine numerische Lösung berechnet, so

gilt

$$\begin{aligned} p_{k_0}(T) &= \sum_{i=0}^{N} p_i(T) \cdot q_i(T) \\ &= \sum_{i=0}^{N} p_i(0) \cdot q_i(0) + \sum_{i=0}^{N} \int_0^T (p_i(t) \cdot q_i(t))' \, \mathrm{d}t \\ &= \sum_{i=0}^{N} p_i(0) \cdot q_i(0) + \int_0^T \underbrace{\sum_{i=0}^{N} (p_i'(t) \cdot q_i(t) + p_i(t) \cdot q_i'(t))}_{=0} \, \mathrm{d}t \\ &= \sum_{i=0}^{N} p_i(0) \cdot q_i(0). \end{aligned} \tag{3.30}$$

Bei den obigen Umformungen haben wir ausgenutzt, dass die auftretenden Terme sich in der Summe wegheben, d.h. es gilt:

$$\begin{aligned} &\sum_{i=0}^{N} (p_i'(t) \cdot q_i(t) + p_i(t) \cdot q_i'(t)) \\ = &\sum_{i=0}^{N} \alpha \cdot (p_{i-1}(t) + p_{i+1}(t) - 2p_i(t)) \cdot q_i(t) + \alpha \cdot p_i(t) \cdot (2q_i(t) - q_{i-1}(t) - q_{i+1}(t)) \\ = &\, 0. \end{aligned}$$

Nun sind wir in der Lage die obige Frage effizient zu beantworten, denn wenn wir nun eine Lösung mit Anfangswert $p_\ell(0) = 1$ und $p_i(0) = 0$ für $i \neq \ell$ einsetzen, so liefert uns (3.30) die Identität $p_{k_0}(T) = q_\ell(0)$. Im Gegensatz zur direkten Berechnung müssen wir nun wieder nur ein Differentialgleichungssystem für die unbekannten Lösungen $q_i(t), i = 0, \ldots, N$ lösen.

Das zu Grunde liegende allgemeine Prinzip der adjungierten Methode ist das Folgende. Wir nehmen an, wir haben eine gewöhnliche lineare Differentialgleichung der Form

$$u'(t) = A(t) \cdot u(t)$$

gegeben und wir interessieren uns nicht direkt für Werte der Lösung zum Endzeitpunkt $u(T) \in \mathbb{R}^n$, sondern nur für eine lineare Funktion $L^T \cdot u(T) \in \mathbb{R}$. Dann lösen wir stattdessen das adjungierte Problem

$$v'(t) = -A(t) \cdot v(t) \tag{3.31}$$

mit dem Endwert $v(T) = L \in \mathbb{R}^n$. Denn dann gilt

$$\begin{aligned} L^T \cdot u(T) &= v(T)^T \cdot u(T) = v(0)^T \cdot u(0) + \int_0^T (v(t)^T \cdot u(t))' \, \mathrm{d}t \\ &= v(0)^T \cdot u(0) + \int_0^T \underbrace{(v'(t)^T \cdot u(t) + v(t)^T \cdot u'(t))}_{=0} \, \mathrm{d}t \\ &= v(0)^T \cdot u(0). \end{aligned}$$

Hierbei nutzt man aus, dass gilt

$$v'(t)^T \cdot u(t) + v(t)^T \cdot u'(t) = -(A(t)^T v(t))^T \cdot u(t) + v(t)^T \cdot A(t)u(t) \;=\; 0.$$

Haben wir die adjungierte Gleichung für v berechnet, können wir die uns interessierende Größe $L^T \cdot u(T)$ für jeden Anfangswert sofort durch ein Skalarprodukt $v(0)^T \cdot u(0)$ berechnen. Den Wert $v(0) \in \mathbb{R}^n$ erhalten wir in dem wir eine Variablentransformation $s := T - t$ in (3.31) durchführen und das entstehende Differentialgleichungssystem numerisch lösen.

Bei einer numerischen Lösung müssen wir natürlich die gewöhnliche Differentialgleichung mit einer geeigneten Methode (Ein- oder Mehrschrittverfahren) diskretisieren. Es empfiehlt sich in diesem Kontext die adjungierte Gleichung mit einem passenden Verfahren zu lösen um die Eigenschaft der Adjungierten auch im Diskreten zu erhalten. Haben wir für die numerische Approximation u z.B. ein Vorwärts-Euler Verfahren der Form

$$u(t_{k+1}) \;=\; u(t_k) + \tau \cdot Au(t_k)$$

verwendet, so liefert das Vorwärts-Euler Verfahren in umgekehrter Zeit

$$v(t_k) \;=\; v(t_{k+1}) + \tau A^T v(t_k)$$

genau die richtige Diskretisierung. Es gilt dann nämlich

$$\begin{aligned} u(t_N) \cdot v(t_N) \;&=\; u(0) \cdot v(0) + \sum_{k=0}^{N-1} (u(t_{k+1}) \cdot v(t_{k+1}) - u(t_k) \cdot v(t_k)) \\ &=\; u(0) \cdot v(0) + \sum_{k=0}^{N-1} ((u(t_{k+1}) - u(t_k) \cdot v(t_{k+1}) + (v(t_{k+1}) - v(t_k)) \cdot u(t_k)) \\ &=\; u(0) \cdot v(0) + \sum_{k=0}^{N-1} (Au(t_k) \cdot v(t_{k+1}) - (A^T v(t_k)) \cdot u(t_k)) \;=\; u(0) \cdot v(0). \end{aligned}$$

Man sieht ein, dass bei anderen Verfahren für die adjungierte Gleichung, z.B. einem impliziten Euler-Verfahren, eine solche Identität nicht gegeben ist. Die Diskretisierung und Adjungierung kommutieren in diesem Fall nicht. Beim Vorwärts-Euler Verfahren hingegen erhalten wir genau die Adjungierte der Diskretisierung der Differentialgleichung.

3.4.3 Zusammenhang zwischen Optimierung und Differentialgleichungen

Ein häufiges Problem in praktischen Anwendungen ist die Bestimmung von Parametern in gewöhnlichen Differentialgleichungen. Wir betrachten also ein Anfangswertproblem

$$u'(t) \;=\; F(t, u(t), w), \qquad u(0) \;=\; u_0(w),$$

bei dem die rechte Seite der gewöhnlichen Differentialgleichung und der Anfangswert von einem Parametervektor $w \in \mathbb{R}^m$ abhängen. Wir nehmen im Folgenden an, dass F Lipschitz-stetig ist. Dann existiert für gegebene Parameter $w \in \mathbb{R}^m$ eine eindeutige

Lösung $u_w \in C^1([0,T])$. Um die Parameter zu einer bestimmten Lösung zu bestimmen, misst man die Werte einer Funktion $G(u_w) \in R^k$, wobei typischerweise $k > m$ gilt. Alternativ versucht man die Parameter $w \in \mathbb{R}^m$ derart zu optimieren, damit ein gewünschter Zustand $G(u_w) \in \mathbb{R}^k$ erreicht wird. Häufig sind dies Werte der Lösung zu verschiedenen Zeitpunkten, also beispielweise $G(u) := (u(s_1), \dots, u(s_k))$.

Für gegebenen Daten $g \in \mathbb{R}^k$ lässt sich dann ein Optimierungsproblem, etwa das **Kleinste-Quadrate-Problem**

$$\min_{w \in \mathbb{R}^m} \left\{ E(w) := \frac{1}{2}\|H(w) - g\|^2 = \frac{1}{2}\|G(u_w) - g\|^2 \right\}$$

lösen um die Parameter zu bestimmen. Die Frage, die wir uns nun stellen ist, wie wir in diesem Fall die Lösung des Optimierungsproblems durch eines der Optimierungsverfahren dieser Vorlesung, wie z.B. das Gradientenabstiegsverfahrens, bestimmen können. Es ist klar, dass hierbei die effiziente Berechnung von Gradienten ∇_w essentiell ist.

BEISPIEL 3.44: Parameterbestimmung für lineare DGL.

Als einfaches Beispiel betrachten wir einen Fall, bei der wir zwar wissen, dass die rechte Seite zu einer linearen gewöhnlichen Differentialgleichung erster Ordnung gehört, aber nicht den Koeffizienten des linearen Terms und auch den Anfangswert nicht kennen. Somit können wir zwei unbekannte Parameter $w = (w_1, w_2) \in \mathbb{R}^2$ beschreiben durch

$$u_0(w) = w_1, \qquad F(t,u,w) = w_2 \cdot u(t).$$

Für Anfangswertprobleme dieser Art können wir eine explizite Lösung $u_w \in C^1([0,T])$ angeben mit

$$u_w(t) = w_1 \cdot e^{w_2 t}.$$

Geben wir uns nun Werte der Lösung an verschiedenen Zeitpunkten $G(u) = (u(s_1), \dots, u(s_k))$ vor, so erhalten wir dann

$$\begin{aligned} \partial_{w_1} G(u_w) &= (e^{w_2 s_1}, \dots, e^{w_2 s_k}), \\ \partial_{w_2} G(u_w) &= (w_1 s_1 \cdot e^{w_2 s_1}, \dots, w_1 s_k \cdot e^{w_2 s_K}). \end{aligned}$$

In Beispiel 3.44 konnten wir die Ableitung der Funktion G nach den unbekannten Parametern w angeben, da wir eine explizite Lösung der gewöhnlichen Differentialgleichung kannten. Wie gehen wir aber vor, wenn wir die Differentialgleichung nicht explizit lösen können? Dazu betrachten wir zunächst die partielle Ableitung von u_w nach w_i, welche wir definieren als

$$u_w^i := \lim_{\delta \to 0} \frac{u_{w+\delta e_i}(t) - u_w(t)}{\delta},$$

wobei e_i der i-te Einheitsvektor ist. Diese Funktion können wir zwar nicht berechnen, aber wir können ein Anfangswertproblem herleiten, das von ihr gelöst wird. Unter der Annahme, dass die Abbildung $w \mapsto u_0(w)$ differenzierbar ist, sehen wir dass gilt

$$u_w^i(0) = \partial_{w_i} u_0(w).$$

Falls außerdem die rechte Seite F der Differentialgleichung differenzierbar bezüglich u und w ist, so folgt mit der Kettenregel

$$(u_w^i)'(t) \;=\; \partial_u F(t, u_w(t), w) \cdot u_w^i(t) + \partial_w F(t, u_w(t), w).$$

Wir beachten, dass wir diese lineare Differentialgleichungen für jedes u_w^i sind, da wir u_w ja schon vorher durch Lösen der ursprünglichen Anfangswertproblems berechnen können. Ist G ebenfalls differenzierbar, dann folgt

$$\partial_{w_i} H(w) \;=\; G'(u_w) \cdot u_w^i,$$

daraus bekommen wir also die Jacobi Matrix von H bzw. dann den Gradienten von f per Kettenregel.

Wir rechnen diesen Zusammenhang nun für das Problem in Beispiel 3.44 nach. Hier gilt

$$u_w^1(0) \;=\; 1, \qquad u_w^2(0) \;=\; 0,$$

und

$$(u_w^1)'(t) \;=\; w_2 u_w^1(t), \qquad (u_w^2)'(t) \;=\; w_2 u_w^2(t) + u_w,$$

und $\partial_{w_i} G(u) \;=\; (u_w^i(s_1), \ldots, u_w^i(s_K))$. Diese können wir explizit lösen und erhalten $u_w^1(t) \;=\; e^{w_2 t}$ und $u_w^2(t) \;=\; w_1 t e^{w_2 t}$. Natürlich stimmen die Ableitungen dann wieder mit der direkten Differentiation der expliziten Lösung u_w wie in Beispiel 3.44 berechnet überein.

Ist die Anzahl M der Parameter groß, so ist die Berechnung der Ableitungen in dieser Form sehr aufwändig, da wir M lineare Differentialgleichungen lösen müssen. Dies kann aber vermieden werden, wenn wir uns daran erinnern, dass wir eigentlich

$$\partial_{w_i} f(w) \;=\; (G(u_w) - g) \cdot G'(u_w) u_w^i$$

berechnen wollen, also eine lineare Funktion von u_w^i. Es ist dementsprechend naheliegend wieder eine adjungierte Methode zu verwenden. Wir betrachten dies wieder näher für $G(u) = (u_w(s_1), \ldots, u_w(s_K))$. Wir definieren v als die Lösung von

$$v'(t) \;=\; -\partial_u F(t, u_w(t), w) \cdot v(t), \qquad t \in (0, T) \setminus \{s_1, \ldots, s_K\},$$

mit $v(T) = 0$. An den Messstellen setzen wir

$$v(s_k) \;=\; u_w(s_k) - g_k + \lim_{t \downarrow s_k} v(t).$$

Dann gilt mit $s_0 = 0$, $s_{K+1} = T$,

$$\begin{aligned}
\partial_{w_i} f(w) &= \sum_k (u_w(s_k) - g_k) \cdot u_w^i(s_k) \\
&= \sum_k (\lim_{t \uparrow s_k} v(t) - \lim_{t \downarrow s_k} v(t)) \cdot u_w^i(s_k) \\
&= v(0) \cdot u_w^i(0) + \sum_{k=0}^{K} \int_{s_k}^{s_{k+1}} (v(t) \cdot u_w^i(t))' \, \mathrm{d}t \\
&= v(0) \cdot u_w^i(0) + \int_0^T (v(t) \cdot u_w^i(t))' \, \mathrm{d}t \\
&= v(0) \cdot \partial_{w_i} u_0(w) + \int_0^T v(t) \cdot \partial_{w_i} F(t, u_w(t), w) \, \mathrm{d}t.
\end{aligned}$$

Damit genügt zur Berechnung des Gradienten die Lösung einer adjungierten Differentialgleichung, sowie von M Skalarprodukten mit Anfangswerten und M Integralen mit der Lösung v.

3.4.4 Deep Learning

In modernen Anwendungen des Maschinellen Lernens kommen ähnliche Techniken wie bei Differentialgleichungen und deren Optimierung zum Einsatz. Die Idee dabei ist einen parametrisierten Zusammenhang zwischen Eingangsdaten $x \in \mathbb{R}^n$ und Ausgangsdaten $y \in \mathbb{R}^m$ zu konstruieren. Dieser Zusammenhang wird beim sogenannten *Deep Learning* durch ein (künstliches) neuronales Netz mit vielen Schichten (im Englischen: *Layer*) modelliert, das mathematisch als Hintereinanderausführung von affinen Abbildungen (Austausch von Impulsen zwischen Neuronen) und punktweisen Nichtlinearitäten (Aktivierung eines Neurons durch die eingegangenen Impulse) modelliert.

Ein **neuronales Netzwerk** $f_\Theta \colon \mathbb{R}^n \to \mathbb{R}^m$ mit $N \in \mathbb{N}^+$ Schichten $f^k_{\Theta_k}, k = 1, \ldots, N$ modelliert die Relation $x \mapsto y$ dann durch

$$f_\Theta \;:=\; f^N_{\Theta_N} \circ \ldots \circ f^1_{\Theta_1},$$

wobei die berechneten Werte der k-ten Schicht des neuronalen Netzes gegeben sind durch

$$u_{k+1} \;=\; f^k_{\Theta_k}(u_k) \;:=\; \Psi(W_k u_k + b_k), \qquad k = 0, \ldots, L-1,$$

mit frei wählbaren Parametern $\Theta_k := \{W_k \in \mathbb{R}^{n_k \times n_k}, b_k \in \mathbb{R}^{n_k}\}$. Die **Aktivierungsfunktion** eines Neurons bezeichnen wir mit $\Psi : \mathbb{R} \to \mathbb{R}$. Typische Beispiele sind die *Sigmoid-Funktion*

$$\Psi(x) \;:=\; \frac{1}{1 + e^{-x}}$$

oder die *Rectified Linear Unit (ReLU)*

$$\Psi(x) \;:=\; \max\{x, 0\}.$$

Dazu verwenden wir für Vektoren die Notation $\Psi(x) = (\Psi(x_i))_{i=1,\dots,n}$ als punktweise Auswertung. Für die erste Schicht des neuronalen Netzes setzen wir $u_0 = x$ auf die Eingangsdaten und in der letzten Schicht haben wir für die Antwort des neuronalen Netzes lediglich eine affine Transformation der Form

$$y \;=\; C \cdot u_N + d,$$

mit $C \in \mathbb{R}^{m \times n_N}$, $d \in \mathbb{R}^m$.

Wir sehen also eine gewisse Analogie zu expliziten Euler-Verfahren für gewöhnliche Differentialgleichungen. Dies wird noch deutlicher bei sogenannten residualen Netzwerken von der Form

$$u_{k+1} \;=\; u_k + \tau \cdot \Psi(W_k \cdot u_k + b_k), \qquad k = 0, \dots, L-1,$$

die wir direkt als Diskretisierung von

$$u'(t) \;=\; \Psi(W(t) \cdot u(t) + b(t))$$

interpretieren können.

Das Training einens neuronalen Netzwerks ist nun die optimale Bestimmung der freien Parameter $\Theta = ((W_k, b_k)_{k=1,\dots,N}, C, d)$ aus einer großen Menge an Trainingsdaten $(x_i, y_i)_{i=1,\dots,M}$. Dazu wird ein Minimierungsproblem der Form

$$\min_{\Theta} \left\{ E(\Theta) \;:=\; \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}(f_\Theta(x_i), y_i) \right\},$$

wobei $\mathcal{L}$ eine Metrik ist, die den Abstand zwischen der Antwort des neuronalen Netzes und den Trainingsdaten misst (im Englischen: *Loss function*), z.B. einfach

$$\mathcal{L}(f_\Theta(x_i), y_i) \;:=\; \frac{1}{2} \| f_\Theta(x_i) - y_i \|^2.$$

Dies ist für große $M/N, m/n$ ein riesiges Optimierungsproblem, dessen approximative Lösung lange Zeit ein großes Hindernis bei der Umsetzung solcher Lernansätze war. Der heute gängige Ansatz ist die Berechnung mit einem stochastischen Gradientenverfahren, d.h. durch eine Iteration

$$w_{j+1} \;=\; w_j - \alpha_j \nabla_w \mathcal{L}(C \cdot u_N(x_{\pi(j)}; (W_k, b_k)) + d, y_{\pi(j)}),$$

wobei $\pi(j) \in \{1, \dots, M\}$ zufällig gewählt wird (meist gleichverteilt). Durch die Auswahl eines einzelnen Datenpaars in jeder Iteration erspart man sich die M-fache Berechnung von $u_N(x_i; (A_k, b_k))$, hier muss in jedem Schritt die Vorwärts-Schleife nur für einen Anfangswert x_i berechnet werden. Analog zur adjungierten Methode kann man den Gradienten durch Lösung von

$$v_{k-1} \;=\; -(W_k \cdot \Psi'(W_k u_k + b_k))^T \cdot v_k$$

mit $v_N = \nabla_u \mathcal{L}(C \cdot u_N(x_{i(j)}; (W_k, b_k)) + d, y_{i(j)})$ berechnen. Dies ist in diesem Zusammenhang als **Backpropagation** bekannt.

Kapitel 4

Numerische Lösung von Randwertproblemen

In diesem Kapitel der Vorlesung wollen wir uns mit der Lösung von Randwertproblemen für **lineare gewöhnliche Differentialgleichungen zweiter Ordnung** beschäftigen. Insbesondere beschäftigen wir uns mit Lösungen $u \in C^2([0,1])$ für Differentialgleichungen der Gestalt

$$-u''(x) - p(x) \cdot u'(x) + q(x) \cdot u(x) \;=\; g(x), \tag{4.1}$$

für $x \in (0,1)$ mit den zusätzlichen Randbedingungen

$$\alpha_0 \cdot u'(0) + \beta_0 \cdot u(0) \;=\; g_0, \tag{4.2}$$
$$\alpha_1 \cdot u'(1) + \beta_1 \cdot u(1) \;=\; g_1. \tag{4.3}$$

Im Fall $\alpha_i = 0$ und $\beta_i = 1$ für $i = 0, 1$ spricht man von **Dirichlet-Randbedingungen** für die gilt:

$$u(0) \;=\; g_0, \qquad u(1) \;=\; g_1.$$

Im Fall $\alpha_i = 1$ und $\beta_i = 0$ für $i = 0, 1$ hingegen sprechen wir von **Neumann-Randbedingungen** für die gilt:

$$u'(0) \;=\; g_0, \qquad u'(1) \;=\; g_1.$$

In allen anderen Fällen erhalten wir gemischte Randwertbedingungen, die auch *Robin-Randbedingungen* genannt werden.

Solche linearen Randwertprobleme treten beispielsweise auf, wenn man die Auslenkung eines an zwei Seiten festgebundenen Seils beschreiben will. Hierbei wird klar, dass die Auslenkung des Seils am linken und rechten Rand verschwindet muss, d.h. wir haben Dirichlet-Randbedingungen vorliegen mit $u(0) = u(1) = 0$. Ein ähnliches Problem untersucht man bei der Berechnung der Temperaturverteilung eines Metallstabs, der an beiden Enden erhitzt wird.

Nehmen wir an $P(x)$ sei eine Stammfunktion der Funktion $p(x)$, so dass $P'(x) = p(x)$ gilt. Dann können wir beide Seiten der Gleichung (4.1) mit dem Term $a(x) := e^{P(x)}$ multiplizieren und erhalten somit:

$$-a(x) \cdot u''(x) - e^{P(x)} \cdot p(x) \cdot u'(x) + e^{P(x)} \cdot q(x) \cdot u(x) \;=\; e^{P(x)} \cdot g(x).$$

Wenn wir nun die Produktregel für Differentiation anwenden sehen wir, dass gilt

$$-(a(x) \cdot u'(x))' \;=\; -a(x) \cdot u''(x) - a'(x) \cdot u'(x) \;=\; -a(x) \cdot u''(x) - e^{P(x)} \cdot p(x) \cdot u'(x).$$

Zusammen mit den Hilfsfunktionen $c(x) := e^{P(x)} \cdot q(x)$ und $f(x) := e^{P(x)} \cdot g(x)$ können wir das Problem (4.1) schließlich in eine sogenannte **Divergenzform** bringen mit:

$$-(a(x) \cdot u'(x))' + c(x) \cdot u(x) \;=\; f(x). \tag{4.4}$$

Die Divergenzform (4.4) der ursprünglichen Differentialgleichung hat einige Vorteile für die Analyse der Eigenschaften der Differentialgleichung und ebenfalls bei ihrer numerischen Lösung, so dass wir diese im Folgenden weiter verwenden werden.

4.1 Existenz und Eindeutigkeit von Lösungen

Mit Hilfe der Divergenzform (4.4) können wir zunächst für den Spezialfall $c(x) \equiv 0$ die sogenannte *Greensche Funktion* zur analytischen Lösung der Differentialgleichung konstruieren. Es gilt nämlich in diesem Fall mit einer Integrationskonstante $c_1 \in \mathbb{R}$

$$a(x) \cdot u'(x) \;=\; c_1 - \int_0^x f(z)\,\mathrm{d}z.$$

Da $a(x) = e^{P(x)} > 0$ für alle $x \in \mathbb{R}$ ist, gilt damit auch für eine weitere Integrationskonstante $c_2 \in \mathbb{R}$

$$u(x) \;=\; c_2 + c_1 \cdot \int_0^x \frac{1}{a(y)}\,\mathrm{d}y - \int_0^x \frac{1}{a(y)} \int_0^y f(z)\,\mathrm{d}z\,\mathrm{d}y.$$

Definieren wir nun zwei Stammfunktionen

$$A(x) \;:=\; \int_0^x \frac{1}{a(y)}\,\mathrm{d}y, \qquad F(y) \;:=\; \int_0^y f(z)\,\mathrm{d}z$$

so erhalten wir durch partielle Integration

$$\begin{aligned}
u(x) \;&=\; c_2 + c_1 \cdot A(x) - \int_0^x \frac{1}{a(y)} \int_0^y f(z)\,\mathrm{d}z\,\mathrm{d}y \\
&=\; c_2 + c_1 \cdot A(x) - \int_0^x \frac{1}{a(y)} \cdot F(y)\,\mathrm{d}y \\
&=\; c_2 + c_1 \cdot A(x) - \big[A(y) \cdot F(y)\big]_0^x + \int_0^x A(y) \cdot f(y)\,\mathrm{d}y \\
&=\; c_2 + c_1 \cdot A(x) - \int_0^x A(x) \cdot f(y)\,\mathrm{d}y + \int_0^x A(y) \cdot f(y)\,\mathrm{d}y \\
&=\; c_2 + c_1 \cdot A(x) - \int_0^x (A(x) - A(y)) \cdot f(y)\,\mathrm{d}y.
\end{aligned} \tag{4.5}$$

Die unbekannten Integrationskonstanten $c_1, c_2 \in \mathbb{R}$ können wir aus den Randbedingungen bestimmen, die uns ein Gleichungssystem liefern, welches es zu lösen gilt. Hierzu betrachten wir im Folgenden die reinen Dirichlet- oder Neumann-Randbedingungen und vernachlässigen die gemischten Randbedingungen.

4.1.1 Dirichlet-Randbedingungen

Im Fall von Dirichlet-Randbedingungen haben wir $\alpha_i = 0$ und $\beta_i = 1$ für $i = 0, 1$ in (4.2) und (4.3) und somit haben wir die Randwerte

$$u(0) \;=\; g_0, \qquad u(1) \;=\; g_1.$$

Dann ist das Gleichungssystem zur Bestimmung der unbestimmten Integrationskonstanten $c_1, c_2 \in \mathbb{R}$ durch Einsetzen in (4.5) gegeben durch

$$u(0) \;=\; c_2 \;=\; g_0, \qquad u(1) \;=\; \underbrace{c_2}_{=g_0} + c_1 \cdot A(1) - \int_0^1 (A(1) - A(y)) \cdot f(y)\,\mathrm{d}y \;=\; g_1.$$

Lösen wir die zweite Randbedingung auf zur Konstanten $c_1 \in \mathbb{R}$ so erhalten wir:

$$c_1 \;=\; \frac{1}{A(1)} \cdot \left(g_1 - g_0 + \int_0^1 (A(1) - A(y)) \cdot f(y)\,\mathrm{d}y \right).$$

Hieraus können wir für die Lösung der Differentialgleichung in (4.5) folgern:

$$\begin{aligned} u(x) \;&=\; g_0 + \frac{A(x)}{A(1)} \cdot \left(g_1 - g_0 + \int_0^1 (A(1) - A(y)) \cdot f(y)\,\mathrm{d}y \right) - \int_0^x (A(x) - A(y)) \cdot f(y)\,\mathrm{d}y \\ &=\; \left(1 - \frac{A(x)}{A(1)}\right) \cdot g_0 + \frac{A(x)}{A(1)} \cdot g_1 \\ &\qquad + \frac{A(x)}{A(1)} \cdot \int_0^1 (A(1) - A(y)) \cdot f(y)\,\mathrm{d}y - \int_0^x (A(x) - A(y)) \cdot f(y)\,\mathrm{d}y. \end{aligned} \tag{4.6}$$

Wir wollen nun die sogenannte **Greensche Funktion** $G\colon [0,1]^2 \to \mathbb{R}$ einführen mit

$$G(x,y) \;:=\; \begin{cases} A(x) \cdot (1 - \frac{A(y)}{A(1)}) & \text{falls } 0 \le x \le y \le 1, \\ A(y) \cdot (1 - \frac{A(x)}{A(1)}) & \text{falls } 0 \le y < x \le 1. \end{cases}$$

Für jede Funktion $a(x) > 0$ gilt für die Stammfunktion $A(x) > A(y) \ge 0$ für $x > y$ und somit ist die Greensche Funktion nichtnegativ auf dem Intervall $[0,1]$.

Nun können wir die beiden Integrale in (4.6) geschickt vereinheitlichen durch

$$\begin{aligned}
&\frac{A(x)}{A(1)} \cdot \int_0^1 (A(1) - A(y)) \cdot f(y)\,\mathrm{d}y - \int_0^x (A(x) - A(y)) \cdot f(y)\,\mathrm{d}y \\
= &\int_0^1 A(x) \cdot \left(1 - \frac{A(y)}{A(1)}\right) \cdot f(y)\,\mathrm{d}y \\
&\quad - \int_0^x \left(A(x) - \frac{A(x)A(y)}{A(1)} - A(y) + \frac{A(y)A(x)}{A(1)}\right) \cdot f(y)\,\mathrm{d}y \\
= &\int_0^1 A(x) \cdot \left(1 - \frac{A(y)}{A(1)}\right) \cdot f(y)\,\mathrm{d}y - \int_0^x A(x) \cdot \left(1 - \frac{A(y)}{A(1)}\right) \cdot f(y)\,\mathrm{d}y \\
&\quad + \int_0^x A(y) \cdot \left(1 - \frac{A(x)}{A(1)}\right) \cdot f(y)\,\mathrm{d}y \\
= &\int_0^x A(y) \cdot (1 - \frac{A(x)}{A(1)}) \cdot f(y)\,\mathrm{d}y + \int_x^1 A(x) \cdot (1 - \frac{A(y)}{A(1)}) \cdot f(y)\,\mathrm{d}y \\
= &\int_0^x G(x,y) \cdot f(y)\,\mathrm{d}y + \int_x^1 G(x,y) \cdot f(y)\,\mathrm{d}y \;=\; \int_0^1 G(x,y) \cdot f(y)\,\mathrm{d}y.
\end{aligned}$$

Wir sehen also, dass wir mit Hilfe der Greenschen Funktion G die analytische Lösung u in (4.6) des Randwertproblems mit Dirichlet-Bedingungen und $c(x) \equiv 0$ kompakt schreiben können als:

$$u(x) \;=\; \left(1 - \frac{A(x)}{A(1)}\right) \cdot g_0 + \frac{A(x)}{A(1)} \cdot g_1 + \int_0^1 G(x,y) \cdot f(y)\,\mathrm{d}y. \tag{4.7}$$

Aus dieser Darstellung der analytischen Lösung können wir sofort folgendes Resultat ableiten.

THEOREM 4.1: Existenz und Eindeutigkeitssatz für $c \equiv 0$.

Sei $c(x) \equiv 0$ und $a \in C^1([0,1])$ mit $a(x) \geq a_0 > 0$ für alle $x \in [0,1]$. Sei außerdem $f \in C([0,1])$.
Dann existiert eine eindeutige Lösung $u \in C^2([0,1])$ der Differentialgleichung (4.4) unter den Dirichlet-Randwertbedingungen (4.2) und (4.3) mit $\alpha_i = 0$ und $\beta_i = 1$ für $i = 0, 1$. Ist $f(x) \geq 0$ für alle $x \in [0,1]$, so gilt darüber hinaus folgendes Maximumsprinzip:

$$u(x) \;\geq\; \min\{g_0, g_1\} \qquad \forall x \in [0,1].$$

Allgemeiner Fall

Wir sehen, dass der *lineare Operator*

$$\begin{aligned}
K\colon C([0,1]) \;&\to\; C([0,1]), \\
f \;&\mapsto\; \int_0^1 G(\cdot, y) \cdot f(y)\,\mathrm{d}y
\end{aligned}$$

offensichtlich auf dem Raum der stetigen Funktionen $C([0,1])$ wohldefiniert und beschränkt (äquivalent zu stetig) ist. Mit Hilfe dieser Einsicht können wir auch den allgemeinen Fall $c(x) \not\equiv 0$ behandeln, da in diesem Fall die Lösung u ein Fixpunkt der folgenden Gleichung ist:

$$u(x) \;=\; g_0 \cdot w(x) + g_1 \cdot (1 - w(x)) + K(f - c \cdot u)(x), \tag{4.8}$$

wobei wir die Notation $w(x) := 1 - \frac{A(x)}{A(1)}$ verwendet haben.

Um die Charakterisierung der Lösung $u(x)$ in Abhängigkeit der Funktion $c(x)$ besser zu verstehen, betrachten wir erst den Spezialfall einer **konstanten Funktion** $c(x) \equiv c \in \mathbb{R}$. Da der Operator K linear ist müssen wir die folgende Gleichung lösen:

$$(I + c \cdot K)(u)(x) \;=\; g_0 \cdot w(x) + g_1 \cdot (1 - w(x)) + K(f)(x).$$

Bei einer genaueren Analyse lässt sich nun feststellen, dass für die spezielle Wahl der Konstanten $c = -k^2\pi^2, k \in \mathbb{N}$ eine nichttriviale Lösung $u(x) = \sin(k\pi \cdot x)$ des homogenen Systems, d.h. für die Gleichung $(I - (k\pi)^2 \cdot K)(u)(x) = 0$ existiert. In diesem Fall ist der Operator $I + c \cdot K$ nicht invertierbar, da dessen Kern nicht nur die Nullfunktion enthält. Also können wir in diesen Fällen die inhomogene Gleichung nicht für beliebige rechten Seiten lösen.

Mit Hilfe der *Spektraltheorie kompakter Operatoren* lässt sich jedoch zeigen, dass es nur abzählbar viele Werte von c gibt, für die $I + c \cdot K$ nicht invertierbar ist. Mit dem *Satz von Arzela-Ascoli* kann man zunächst zeigen, dass $K : C([0,1]) \to C([0,1])$ kompakt ist, d.h. für jede beschränkte Folge $(u_n)_{n\in\mathbb{N}} \subset C([0,1])$ hat $K(u_n)$ eine konvergente Teilfolge. Die Spektraltheorie kompakter Operatoren garantiert nun, dass es nur eine abzählbare Menge von Eigenwerten $(\lambda_k)_{k\in\mathbb{N}}$ geben kann, so dass der Operator $\lambda_k \cdot I - K$ nicht invertierbar ist. Es lässt sich zeigen, dass in der Menge der Eigenwerte Null der einzige Häufungspunkt ist. Setzen wir in diesem Zusammenhang die Konstante $c := -\frac{1}{\lambda_k}$ für $k \in \mathbb{N}$, so sehen wir, dass es auch nur eine abzählbare Menge von Konstanten c geben kann, für die der Operator $I + c \cdot K$ nicht invertierbar ist. Diese Menge besitzt entsprechend die (uneigentlichen) Häufungspunkte $\pm\infty$.

Wenn wir nun eine **allgemeine Funktion** $c(x)$ betrachten, dann können wir immer noch zeigen, dass der Operator $u \mapsto u + K(c \cdot u)$ kompakt ist. Daraus können wir wieder folgern, dass der Operator genau dann invertierbar ist, wenn sein Nullraum trivial ist. Bevor wir uns hierfür eine hinreichende Bedingung erschließen, wollen wir zunächst im folgenden Lemma eine nützliche Eigenschaft des Integraloperators K nachweisen.

LEMMA 4.2: Eigenschaften des Integraloperators.

Für alle stetigen Funktionen $f \in C([0,1])$ gilt die folgende Abschätzung

$$\int_0^1 f(x) \cdot K(f)(x)\,\mathrm{d}x \;=\; \int_0^1 \int_0^1 f(x) \cdot G(x,y) \cdot f(y)\,\mathrm{d}x\,\mathrm{d}y \;\geq\; 0.$$

Das Integral wird genau dann Null, wenn $f(x) \equiv 0$ gilt.

Beweis. Um die Eigenschaften des Integraloperators zu zeigen, wählen wir ein möglichst einfaches Randwertproblem der Form

$$-(a(x) \cdot u'(x))' \;=\; f(x),$$

mit den Dirichlet-Randbedingungen $u(0) = u(1) = 0$. Wir können also die analytische Lösung u des Randwertproblems entsprechend (4.8) für $g_0 = g_1 = 0$ und $c \equiv 0$ für $x \in [0, 1]$ schreiben als

$$u(x) \;=\; K(f)(x).$$

Somit können wir mit Hilfe von partieller Integration folgern, dass gilt:

$$\begin{aligned}
\int_0^1 f(x) \cdot K(f)(x)\,\mathrm{d}x \;&=\; -\int_0^1 (a(x) \cdot u'(x))' \cdot u(x)\,\mathrm{d}x \\
&=\; -\underbrace{\big[a(x) \cdot u'(x) \cdot u(x)\big]_0^1}_{=0} + \int_0^1 a(x) \cdot u'(x) \cdot u'(x)\,\mathrm{d}x \\
&=\; \int_0^1 a(x) \cdot (u'(x))^2\,\mathrm{d}x \;\geq\; 0.
\end{aligned}$$

Die Nichtnegativität des letzten Integrals folgt aus dem quadratischen Term $(u'(x))^2$ und der Eigenschaft, dass $a(x) = e^{P(x)} > 0$ für alle $x \in [0, 1]$ gilt. Damit folgt sofort, dass das Integral genau dann Null wird, wenn $u'(x) \equiv 0$ gilt. Aus $-(a(x) \cdot u'(x))' = f(x)$ folgt dann aber auch schon, dass $f(x) \equiv 0$ gelten muss. □

Aus dem obigen Spezialfall mit konstanter Funktion $c(x) \equiv c$ sehen wir, dass ein nichttrivialer Nullraum bei negativem c auftreten kann. Diese Beobachtung lässt sich nun mit Hilfe der Eigenschaft des Integraloperators aus Lemma 4.2 auf den allgemeinen Fall von $c(x) \not\equiv 0$ übertragen, so dass wir im folgenden ein allgemeines Resultat zur Existenz und Eindeutigkeit von Lösungen des Randwertproblems erhalten.

THEOREM 4.3: Existenz und Eindeutigkeitssatz für $c \geq 0$.

Sei $a \in C^1([0, 1])$ eine stetig differenzierbare Funktion mit $a(x) \geq a_0 > 0$ für alle $x \in [0, 1]$. Seien außerdem $c, f \in C([0, 1])$ stetige Funktionen und es gelte $c(x) \geq 0$ für alle $x \in [0, 1]$.
Dann existiert eine eindeutige Lösung $u \in C^2([0, 1])$ des Randwertproblems (4.4) mit Dirichlet-Randbedingungen (4.2) und (4.2) mit $\alpha_i = 0$ und $\beta_i = 1$ für $i = 0, 1$.

Beweis. Basierend auf den oben genannten Argumenten aus der Funktionalanalysis genügt es zu zeigen, dass der lineare Operator $u \mapsto u + K(c \cdot u)$ invertierbar ist, was äquivalent dazu ist, dass er injektiv ist und somit sein Kern nur die Nullfunktion $u(x) \equiv 0$ enthält.

Sei also $u(x) + K(c \cdot u)(x) = 0$, dann gilt ebenfalls nach Multiplikation beider Seiten mit $c(x) \cdot u(x)$:

$$c(x) \cdot u^2(x) \;=\; -c(x) \cdot u(x) \cdot K(c \cdot u)(x).$$

Integrieren wir beide Seiten der Gleichung und wenden nun Lemma 4.2 für die speziell gewählte stetige Funktion $f(x) := c(x) \cdot u(x) \in C([0,1])$ an, so können wir abschätzen:

$$\int_0^1 c(x) \cdot u^2(x)\,\mathrm{d}x \;=\; -\int_0^1 c(x) \cdot u(x) \cdot K(c \cdot u)(x)\,\mathrm{d}x \;\leq\; 0.$$

Da auf der linken Seite ein nichtnegativer Integrand steht muss hier schon die Gleichheit mit Null vorliegen. Ebenfalls mit Lemma 4.2 wissen wir, dass das Integral genau dann Null wird, wenn $f(x) = c(x) \cdot u(x) \equiv 0$ ist. Wenn jedoch $c(x) \cdot u(x) \equiv 0$, so folgt aus der Linearität von K, dass $K(c \cdot u)(x) \equiv 0$ gelten muss. Da wir $u(x) + K(c \cdot u)(x) = 0$ angenommen haben, muss damit auch $u(x) \equiv 0$ gelten. Also ist der Nullraum des Operators trivial, woraus die Invertierbarkeit und damit auch die eindeutige Lösbarkeit des Randwertproblems folgt. □

Interessanterweise können wir auch im allgemeinen Fall ein ähnliches **Maximumsprinzip** zeigen, auch wenn wir nun keine explizite Darstellung der Lösung als Integral mehr vorliegen haben.

KOROLLAR 4.4: Maximumsprinzip.

Sei $a \in C^1([0,1])$ eine stetig differenzierbare Funktion mit $a(x) \geq a_0 > 0$ für alle $x \in [0,1]$. Seien außerdem $c, f \in C([0,1])$ nichtnegative, stetige Funktionen, d.h. es gelte $c(x) \geq 0$ und $f(x) \geq 0$ für alle $x \in [0,1]$.
Dann gilt für die eindeutige Lösung $u \in C^2([0,1])$ des Randwertproblems (4.4) mit beliebigen Dirichlet-Randbedingungen $u(0) = g_0 \in \mathbb{R}$ und $u(1) = g_1 \in \mathbb{R}$, die folgende Abschätzung:

$$u(x) \;\geq\; \min\{g_0, g_1\} \qquad \forall x \in [0,1].$$

Beweis. Die Existenz und Eindeutigkeit einer Lösung $u \in C^2([0,1])$ des Randwertproblems ist durch Theorem 4.3 gegeben. Für das zu zeigende Maximumsprinzip führen wir einen einfachen Widerspruchsbeweis.

Seien im Folgenden $c, f \in C([0,1])$ nichtnegative Funktionen und wir nehmen an, dass die analytische Lösung $u(x)$ nicht am Rand des Intervalls ihr Minimum annimmt, sondern in einem Punkt $\overline{x} \in (0,1)$. Dann gilt die notwendige Optimalitätsbedingungen $u'(\overline{x}) = 0$ in $\overline{x}$. Setzen wir diese in die Differentialgleichung ein, so folgt für alle $x \in [0,1]$

$$0 \;\leq\; f(\overline{x}) \;=\; -(a(\overline{x}) \cdot \underbrace{u'(\overline{x})}_{=0})' + c(\overline{x}) \cdot u(\overline{x}) \;=\; c(\overline{x}) \cdot u(\overline{x}) \;\leq\; c(\overline{x}) \cdot u(x).$$

Wegen der Annahme, dass c und f nichtnegative Funktionen sind, folgt damit, dass $u(x)$ ebenfalls nichtnegativ für alle $x \in [0,1]$ ist. Da wir jedoch beliebige Randwerte $u(0) = g_0$ und $u(1) = g_1$ für das Randwertproblem erlauben, können wir ebenfalls Randwerte betrachten, so dass $\min\{g_0, g_1\} < 0$ gilt. Damit erzeugt die gefolgerte Nichtnegativität der Lösung u einen Widerspruch zu den Randwerten und somit nimmt die analytische Lösung $u \in C^2([0,1])$ ihr Minimum am Rand an und somit haben wir das Maximumsprinzip gezeigt. □

Wegen der Linearität des Problems kann man aus dem Maximumsprinzip in Korollar 4.4 auch Stabilitätsaussagen herleiten. Nehmen wir an $\tilde{u} \in C^2([0,1])$ sei eine bekannte Lösung für das Problem

$$-(a(x) \cdot \tilde{u}'(x))' + c(x) \cdot \tilde{u}(x) \;=\; \tilde{f}(x),$$

mit Dirichlet Randwerten $\tilde{u}(0) = \tilde{g}_0$ und $\tilde{u}(1) = \tilde{g}_1$. Betrachten wir nun eine weitere rechte Seite $f \in C([0,1])$ der Differentialgleichung mit $\tilde{f}(x) \geq f(x)$ für alle $x \in [0,1]$, so folgt

$$\tilde{u}(x) - u(x) \;\geq\; \min\{\tilde{g}_0 - g_0, \tilde{g}_1 - g_1\}.$$

und daraus eine Abschätzung für das Maximum von u. Ist $c > 0$, so können wir etwa sehr einfache konstante Lösungen $\tilde{u} = \gamma$ konstruieren, indem wir $\tilde{f} = \gamma c$ setzen. Ist γ hinreichend groß, dann ist $\tilde{f} > f$ und $\tilde{g}_i > g_i$ für $i = 0, 1$.

Damit erhalten wir $u(x) \leq \gamma$ für alle x. Umgekehrt können wir die Rollen von $\tilde{u}$ und u auch vertauschen und $\tilde{u} = -\gamma$ wählen, damit erhalten wir eine Abschätzung für $|u(x)|$.

4.1.2 Neumann-Randbedingungen

Während allgemeine Randbedingungen sehr ähnlich behandelt werden wie im oben diskutierten Fall von Dirichlet-Randwertbedingungen, gibt es im Fall reiner Neumann-Randbedingungen einen Aspekt, den wir besonders beachten müssen.

Betrachten wir also im Folgenden die Differentialgleichung (4.4) mit Randwerten (4.2) und (4.3) für $\alpha_i = 1$ und $\beta_i = 0$. Zur Vereinfachung nehmen wir wieder an, dass $c(x) \equiv 0$ gilt. Dann können wir die Differentialgleichung zunächst von links und rechts aufintegrieren zu

$$\begin{aligned} -\int_0^x (a(y) \cdot u'(y))' \,\mathrm{d}y \;&=\; -a(x) \cdot u'(x) + a(0) \cdot \underbrace{u'(0)}_{= g_0} \;=\; \int_0^x f(y) \,\mathrm{d}y \\ \Rightarrow \quad -a(x) \cdot u'(x) \;&=\; -a(0) \cdot g_0 + \int_0^x f(y) \,\mathrm{d}y. \end{aligned}$$

Integrieren wir hingegen mit dem rechten Rand auf so erhalten wir:

$$\begin{aligned} -\int_x^1 (a(y) \cdot u'(y))' \,\mathrm{d}y \;&=\; -a(1) \cdot \underbrace{u'(1)}_{= g_1} + a(x) \cdot u'(x) \;=\; \int_x^1 f(y) \,\mathrm{d}y \\ \Rightarrow \quad -a(x) \cdot u'(x) \;&=\; -a(1) \cdot g_1 - \int_x^1 f(y) \,\mathrm{d}y \;=\; -a(1) \cdot g_1 - \int_0^1 f(y) \,\mathrm{d}y + \int_0^x f(y) \,\mathrm{d}y. \end{aligned}$$

Ein direkter Vergleich der beiden Gleichungen zeigt uns, dass die folgende Bedingung erfüllt sein muss

$$a(0) \cdot g_0 \;=\; a(1) \cdot g_1 + \int_0^1 f(y) \,\mathrm{d}y \tag{4.9}$$

Unter dieser Bedingung ist die Gleichung prinzipiell lösbar, jedoch ist eine der Konstanten $a(0), a(1) \in \mathbb{R}$ dann weiterhin unbestimmt. Dies wird üblicherweise durch eine

zusätzliche Normalisierungsbedingung der folgenden Art erreicht

$$\int_0^1 u(x)\,\mathrm{d}x \;=\; 1.$$

Der Grund für diese zusätzliche Bedingung ist im Gegensatz zu den oben diskutierten Dirichlet-Randwertbedingungen ist, dass der Nullraum des Randwertproblems (4.4) mit $c(x) \equiv 0$ unter Neumann-Randwertbedingungen nichttrival ist, denn jede konstante Funktion löst bereits das homogene Problem

$$-(a(x)\cdot u'(x))' \;=\; 0$$

mit Randwerten $u'(0) = u'(1) = 0$.

Interessanterweise ist die Theorie für den Fall nichtnegativer Funktionen $c(x) \geq 0$ hier unterschiedlich. Sobald $c(x) \geq 0$ mit $c(x) \not\equiv 0$ für alle $x \in [0,1]$ gilt, hat das Randwertproblem mit Neumann-Randwertbedingungen nur noch einen trivialen Nullraum. Dies lässt sich durch Multiplikation der homogenen Differentialgleichung mit $u(x)$ und anschließender Integration sehen. Es folgt dann nämlich mit partieller Integration und den Randwertbedingungen $u'(0) = u'(1) = 0$, dass gilt

$$\begin{aligned} 0 \;&=\; \int_0^1 (-(a(x)\cdot u'(x))' + c(x)\cdot u(x))\cdot u(x)\,\mathrm{d}x \\ &=\; \underbrace{-\big[a(x)\cdot u'(x)\cdot u(x)\big]_0^1}_{=0} + \int_0^1 a(x)\cdot u'(x)\cdot u'(x)\,\mathrm{d}x + \int_0^1 c(x)\cdot u^2(x)\,\mathrm{d}x \\ &=\; \int_0^1 a(x)\cdot (u'(x))^2 + c(x)\cdot u^2(x)\,\mathrm{d}x. \end{aligned}$$

Da $a(x) = e^{P(x)} > 0$ und $c(x) \geq 0$ angenommen wurde, folgt sofort, dass $u'(x) \equiv 0$ für $x \in [0,1]$ gelten muss. Also muss u eine konstante Funktion sein. Da $c(x) \not\equiv 0$ angenommen wurde kann

$$\int_0^1 c(x)\cdot u^2(x)\,\mathrm{d}x \;=\; 0$$

für eine konstante Funktion u nur im Fall $u(x) \equiv 0$ gelten.

Der Unterschied dieser Fälle und der potentiell nichttriviale Nullraum des Randwertproblems mit Neumann-Randwertbedingungen ist natürlich auch bei der numerischen Lösung zu beachten, welche im nächsten Abschnitt diskutiert wird. Hier wird sich dieser Nullraum in der Nichtinvertierbarkeit einer zugehörigen Matrix niederschlagen.

4.2 Differenzenverfahren für Randwertprobleme

Zur Diskretisierung von Randwertproblemen können wir zunächst genauso vorgehen wie bei der numerischen Lösung von Anfangswertproblemen. Wir konstruieren also zunächst ein Gitter $0 = x_0 < x_1 < \ldots < x_N = 1$ als Ortsdiskretisierung des Intervalls $[0,1]$. Im

einfachsten Fall wählen wir die Gitterpunkte wieder äquidistant mit Schrittweite $h := \frac{1}{N}$ als

$$\Omega_h \;:=\; \{x_k \in [0,1] : x_k := k \cdot h, k = 0, \ldots, N\}.$$

Auf diesem Gitter approximieren wir nun Ableitungen durch finite Differenzen. Für die zweite Ableitung in x_k verwenden wir in der Regel ein **Differenzenverfahren zweiter Ordnung** mit

$$u''(x_k) \;\approx\; \frac{u(x_k+h) - 2u(x_k) + u(x_k-h)}{h^2} \;=\; \frac{u(x_{k+1}) - 2u(x_k) + u(x_{k-1})}{h^2}.$$

Im *nichtäquidistanten Fall* sieht eine entsprechende Approximation hingegen wie folgt aus:

$$u''(x_k) \;\approx\; \frac{1}{h_k} \cdot \left(\frac{u(x_{k+1}) - u(x_k)}{x_{k+1} - x_k} - \frac{u(x_k) - u(x_{k-1})}{x_k - x_{k-1}} \right). \tag{4.10}$$

Hier ist zunächst die Frage wie wir den Parameter $h_k > 0$ (abhängig von den Gitterpunkten x_{k-1}, x_k, x_{k+1}) wählen sollen. Dies können wir als Bedingung an die maximale Konsistenzordnung eines numerischen Verfahrens formulieren.

Für hinreichend glatte Funktionen u und

$$h \;:= \max_{k=0,\ldots,N-1} |x_{k+1} - x_k|$$

gilt per Taylor-Entwicklung der Funktionswerte $u(x_{k+1})$ und $u(x_{k-1})$ jeweils im Punkt $x_k \in [0,1]$ dann

$$\begin{aligned}
&\frac{u(x_{k+1}) - u(x_k)}{x_{k+1} - x_k} - \frac{u(x_k) - u(x_{k-1})}{x_k - x_{k-1}} \\
=\;& u'(x_k) + \frac{1}{2}u''(x_k)(x_{k+1} - x_k) + \frac{1}{6}u'''(x_k)(x_{k+1} - x_k)^2 - \\
& u'(x_k) + \frac{1}{2}u''(x_k)(x_k - x_{k-1}) - \frac{1}{6}u'''(x_k)(x_k - x_{k-1})^2 + \mathcal{O}(h^3) \\
=\;& u''(x_k)\frac{x_{k+1} - x_{k-1}}{2} + \frac{1}{6}u'''(x_k)(x_{k+1} - 2x_k + x_{k-1})(x_{k+1} - x_{k-1}) + \mathcal{O}(h^3).
\end{aligned} \tag{4.11}$$

Wir sehen, dass wir eine konsistente Approximation der zweiten Ableitung mit der Wahl von $h_k = \frac{x_{k+1} - x_{k-1}}{2}$ im Vorfaktor erreichen.

Im äquidistanten Fall erkennen wir an obiger Rechnung wieder, dass die Konsistenzordnung $\mathcal{O}(h^2)$ erreicht wird, da dann gilt

$$x_{k+1} - 2x_k + x_{k-1} \;=\; \underbrace{(x_{k+1} - x_k)}_{=h} - \underbrace{(x_k - x_{k-1})}_{=h} \;=\; 0.$$

Die konsistente Wahl des Vorfaktors $h_k = \frac{x_{k+1} - x_{k-1}}{2}$ lässt sich ebenfalls anders motivieren: eigentlich berechnen wir numerische Approximationen erster Ableitungen mit Hilfe des zentralen Differenzenquotienten als

$$\begin{aligned}
\frac{u(x_{k+1}) - u(x_k)}{x_{k+1} - x_k} \;&=\; u'(x_{k+1/2}) + \mathcal{O}(h^2), \\
\frac{u(x_k) - u(x_{k-1})}{x_k - x_{k-1}} \;&=\; u'(x_{k-1/2}) + \mathcal{O}(h^2)
\end{aligned}$$

mit den beiden (impliziten) Mittelpunkten

$$x_{k-1/2} \;:=\; \frac{1}{2}(x_{k-1}+x_k), \qquad x_{k+1/2} \;:=\; \frac{1}{2}(x_{k+1}+x_k).$$

Diese beiden Mittelpunkte liegen also auf einem versetzten Gitter.

Nun können wir die zweite Ableitung wiederum als numerische Approximation der Ableitung der Diskretisierungen der ersten Ableitungen betrachten. Eine weitere Anwendung des zentralen Differenzenquotienten auf die Werte der versetzten Gitterpunkte $x_{k\pm1/2}$ liefert nämlich genau eine Approximation der zweiten Ableitung im Punkt $x_k \in \Omega_h$. Die Schrittweite für diesen zweiten Diskretisierungsschritt ist genau der von uns berechnete Vorfaktor

$$h_k \;=\; \frac{x_{k+1}-x_{k-1}}{2} \;=\; \frac{x_{k+1}+x_k}{2} - \frac{x_k+x_{k-1}}{2} \;=\; x_{k+1/2} - x_{k-1/2}.$$

Dirichlet Randbedingungen

Mit den oben eingeführten Differenzenverfahren können wir nun bereits Gleichungen der Form

$$-u''(x) + c(x)\cdot u(x) \;=\; f(x) \tag{4.12}$$

für $x \in [0,1]$ direkt diskretisieren. In diesem Kontext betrachten wir im Folgenden das Randwertproblem mit Dirichlet-Randwertbedingungen, welches wir als lineares Gleichungssystem für den unbekannten Lösungsvektor

$$U_h \;=\; (u_1,\ldots,u_{N-1})^T \in \mathbb{R}^{N-1}$$

schreiben können. Die beiden Werte $u_0 = u(0) = g_0$ und $u_N = u(1) = g_1$ ergeben sich ganz natürlich aus den Dirichlet-Randbedingungen und können direkt in einen Lösungsvektor für das gesamte numerische Gitter Ω_h eingesetzt werden.

Verwenden wir nun das Differenzenschema zweiter Ordnung in (4.10) zur numerischen Diskretisierung der Differentialgleichung, so können wir die Koeffizienten des Differenzenschemas für jeden Punkt in eine Zeile der Systemmatrix $A \in \mathbb{R}^{N-1\times N-1}$ des Problems schreiben. Hierzu können wir genauer für die Matrix-Einträge von A definieren als:

$$\begin{aligned}
A_{k,k} \;&:=\; \frac{1}{h_k}\left(\frac{1}{x_{k+1}-x_k} + \frac{1}{x_k+x_{k-1}}\right) + c(x_k), \qquad k = 1,\ldots,N-1 \\
A_{k,k-1} \;:=\; -\frac{1}{h_k(x_k-x_{k-1})}, \qquad A_{k-1,k} \;&:=\; -\frac{1}{h_k(x_{k+1}-x_k)} \qquad k = 2,\ldots,N-1 \\
A_{k,j} \;&:=\; 0 \qquad \text{sonst},
\end{aligned}$$

Wir erhalten damit also eine Tridiagonalmatrix $A \in \mathbb{R}^{N-1\times N-1}$, die sich im äquidistanten Fall vereinfacht zu

$$A \;:=\; \frac{1}{h^2}\begin{pmatrix}
2+h^2c(x_1) & -1 & 0 & \ldots & 0 \\
-1 & 2+h^2c(x_2) & -1 & \ldots & 0 \\
0 & -1 & 2+h^2c(x_3) & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & 2+h^2c(x_{N-1})
\end{pmatrix}. \tag{4.13}$$

Für die rechte Seite des linearen Gleichungssystems müssen wir zum einen die rechte Seite $f(x)$ der Differentialgleichung abbilden, als auch die Randwertbedingungen $u(0) = g_0$ und $u(1) = g_1$ berücksichtigen. Hierzu definieren wir uns entsprechend den Vektor $F \in \mathbb{R}^{N-1}$ mit:

$$\begin{aligned} F_k &:= f(x_k), \qquad k \in \{2, \dots, N-2\} \\ F_1 &:= f(x_1) + \frac{g_0}{h_1(x_1 - x_0)}, \\ F_{N-1} &:= f(x_{N-1}) + \frac{g_1}{h_{N-1}(x_N - x_{N-1})}. \end{aligned}$$

Insgesamt müssen wir also das lineare $(N-1) \times (N-1)$ Gleichungssystem $AU_h = F$ numerisch lösen um eine diskrete Lösung des Randwertproblems (4.12) mit Dirichlet-Randbedingungen $u(0) = g_0$ und $u(1) = g_1$ zu erhalten.

Im allgemeinen Fall einer nichtkonstanten Funktion $a(x) \not\equiv c \in \mathbb{R}$ können wir die selbe Einsicht über verschobene Gitter wie oben benutzen. Die erste Ableitung $(a(x_k) \cdot u'(x_k)'$ interpretieren wir diskret als Wert am Mittelpunkt der Gitterzellen $x_{k\pm 1/2}$. Da wir diese mit der Funktion a multiplizieren, verwenden wir auch den Wert von a an diesen Gitterpunkten. Dementsprechend approximieren wir

$$(au')'(x_k) \approx \frac{1}{h_k}\left(a(x_{k+1/2})\frac{u_{k+1} - u_k}{x_{k+1} - x_k} - a(x_{k-1/2})\frac{u_k - u_{k-1}}{x_k - x_{k-1}}\right). \tag{4.14}$$

In der folgenden Bemerkung wollen wir auf Eigenschaften der Systemmatrix A eingehen, die allgemein bei der Lösung von Randwertproblemen eine wichtige Rolle spielen.

BEMERKUNG 4.5 (Eigenschaften der Systemmatrix A). Wir können folgende Beobachtungen zu den Eigenschaften der Matrix A in (4.13) machen.

- Die Matrix A ist **dünnbesetzt** (im Englischen: *sparse*), d.h. die meisten Einträge von A sind gleich null. Genauer noch existieren von den $(N-1)^2$ möglichen Matrixeinträgen von A nur $3 \cdot (N-1) - 2 = 3N - 5$ Einträge, die nicht verschwinden. Dies hat einige Konsequenzen für die effiziente Speicherung von A, sowie bei der Lösung des Systems $AU_h = F$. Wir können die Matrix im sogenannten **Sparse-Format** speichern. Hierbei handelt es sich in der Regel um eine lineare Liste für alle Nichtnulleinträge der Form (i, j, A_{ij}). Statt $(n-1)^2$ reeller Zahlen speichern wir hier nur $6N-10$ ganze und $3N-5$ reelle Zahlen. Gegenüber potentiell $(N-1)^2$ zu speichernden reellen Zahlen ist dies eine enorme Einsparung für großes $N \in \mathbb{N}$.

  Bei der numerischen Lösung des linearen Systems ist es ebenfalls vorteilhaft diese Struktur zu benutzen. Bei direkten Verfahren wie der LR-Zerlegung jedoch ist dies nicht der Fall. Dort kann es zum sogenannten **Fill-In** kommen, d.h. L und R sind nicht mehr dünnbesetzt. Grob können wir uns den Effekt bei der LR-Zerlegung wie bei der Integration von $-u'' = f$, $u(0) = g_0$, $u(1) = g_1$ vorstellen. Hier erhalten wir durch die Integration von 0 bis x eine Integraldarstellung für $u'(x)$, die im diskreten einer linken unteren Dreiecksmatrix enstpricht. Integrieren wir wiederum von 1 bis x um $u(x)$ zu erhalten, entspricht dies einer rechten oberen Dreiecksmatrix. Beide

Matrizen haben dann keine dünnbesetzte Struktur mehr, da das diskrete Integral alle vorhandenen Gitterpunkte benutzt.

- Für nichtnegative Funktionen $c(x) \geq 0$ ist die Matrix A **schwach diagonaldominant**, d.h. es gilt $A_{kk} \geq \sum_{j \neq k} |A_{kj}|$ (bzw. auch in den Spalten $A_{kk} \geq \sum_{j \neq k} |A_{jk}|$). In der ersten Zeile für $k = 1$ und in der letzten Zeile für $k = N - 1$ der Matrix A sind diese Ungleichungen sogar strikt.

  Darüber hinaus hat A nur positive Diagonaleinträge und nichtpositive Nebendiagonaleinträge. Wie wir später sehen werden ist die Matrix A dann invertierbar und die Inverse enthält nur nichtnegative Einträge. Dies ist das diskrete Äquivalent zum Maximumsprinzip, denn hat F nur nichtnegative Einträge, so gilt auch $U_h = A^{-1}F \geq 0$.

- Die Matrix A ist **symmetrisch**, d.h. es gilt $A_{jk} = A_{kj}$ für alle k, j. Dies ist eine Konsequenz aus der Divergenzform (4.4), da der Operator $L : u \mapsto -(au')' + cu$ formal selbstadjungiert ist. Es gilt mit partieller Integration für u und v und angenommenen Dirichlet-Nullrandwerten:

$$\begin{aligned}
\langle Lu, v\rangle_{L^2} &= \int_0^1 (Lu)(x)\ v(x)\,\mathrm{d}x \;=\; \int_0^1 (-(a(x)u'(x))v(x) + c(x)u(x)v(x)\,\mathrm{d}x \\
&= \int_0^1 a(x)u'(x)v'(x) + c(x)u(x)v(x)\,\mathrm{d}x \\
&= \int_0^1 (-(a(x)v'(x))u(x) + c(x)v(x)u(x))\,\mathrm{d}x \;=\; \langle v, Lu\rangle_{L^2}.
\end{aligned}$$

  Ist $c(x) \geq 0$, dann ist die Matrix A sogar symmetrisch positiv definit. Dies ist für $u(x) \not\equiv 0$ eine Konsequenz aus

$$\langle Lu, u\rangle \;=\; \int_0^1 a(x) \cdot |u'(x)|^2 + c(x) \cdot u(x)^2 \mathrm{d}x \;>\; 0.$$

△

Mit den oben diskutierten Eigenschaften der Systemmatrix A erhalten wir insbesondere ihre Invertierbarkeit aus der positiven Definitheit. Also existiert eine eindeutige Lösung $U_h \in \mathbb{R}^{N-1}$ des linearen Gleichungssystems $AU_h = F$.

4.2.1 Konvergenz von Differenzenverfahren

In diesem Abschnitt wollen wir uns mit der Frage beschäftigen welche Bedingungen gelten müssen, damit ein Differenzenverfahren zur numerischen Lösung eines Randwertproblems gegen die echte Lösung der Differentialgleichung konvergiert. Wie wir bereits gesehen haben führt das Differenzenverfahren zu einem linearen Gleichungssystem der Form $AU_h = F$, wobei $U_h \in \mathbb{R}^{N-1}$ die numerische Approximation der echten Lösung $u \in C^2([0,1])$ in den Gitterpunkten darstellt, d.h., es gilt $(U_h)_k = u_k \approx u(x_k)$ für $k = 1, \ldots, N-1$.

M-Matrizen

Wie wir im Folgenden feststellen werden hängt die Stabilität und folglich damit die Konvergenz eines Differenzenverfahrens maßgeblich von den Eigenschaften der Systemmatrix $A \in \mathbb{R}^{N-1\times N-1}$ ab. Die in Bemerkung 4.5 diskutierten Eigenschaften der speziellen Systemmatrix A motivieren die folgende Definition.

DEFINITION 4.6: M-Matrix.

Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt **M-Matrix**, wenn sie folgende Eigenschaften erfüllt:

- A hat nur positive Diagonaleinträge und nichtpositive Nebendiagonaleinträge.
- A ist schwach diagonaldominant, d.h. es gilt $A_{kk} \geq \sum_{j\neq k} |A_{jk}|$.
- Für mindestens ein $j \in \{1,\dots,n\}$ gilt $A_{kk} > \sum_{j\neq k} |A_{jk}|$.

Wie wir leicht nachrechnen können erfüllt unsere Diskretisierung diese Eigenschaft:

KOROLLAR 4.7.

Sei A die Matrix aus dem obigen Differenzenverfahren in (4.13) für eine nichtnegative Funktion $c(x) \geq 0$ für $x \in [0,1]$. Dann ist A eine M-Matrix.

Die letzte Eigenschaft einer M-Matrix in Definition 4.6 ist entscheidend für deren Invertierbarkeit, da sonst beispielsweise auch die folgende Matrix zulässig wäre:

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

Diese Eigenschaft können wir allgemein für M-Matrizen beweisen und darüber hinaus etwas über die Einträge der Inversen aussagen.

LEMMA 4.8: Inverse einer M-Matrix.

Sei $A \in \mathbb{R}^{n\times n}$ für $n \in \mathbb{N}$ eine M-Matrix. Dann ist A invertierbar und A^{-1} hat nur nicht-negative Einträge.

Beweis. Wir nehmen zunächst an, dass A irreduzibel ist, d.h. es existiert eine Permutation $\pi \in \Pi_n$ von $\{1,\dots,n\}$, sodass $A_{\pi(i)\pi(i+1)} \neq 0$ gilt. Andernfalls können wir analog wie im Folgenden für alle irreduziblen Blöcke von A vorgehen.

Zunächst zeigen wir, dass A invertierbar ist, d.h. wir müssen zeigen, dass der Nullraum trivial ist. Sei $z \in \mathcal{N}(A)$ ein Vektor des Nullraums und es sei ein Vektoreintrag z_m für $m \in \{1,\dots,n\}$ so, dass $z_m \leq z_k$ für alle k gilt. Nehmen wir nun an, dass $z_m < 0$ ist, dann folgt

$$A_{mm}z_m = -\sum_{j\neq m} A_{mj}z_j = \sum_{j\neq m} |A_{mj}|z_j \geq \sum_{j\neq m} |A_{mj}|z_m.$$

Ist $z_m < 0$, so folgt wegen der Diagonaldominanz $\sum_{j\neq m} |A_{mj}| z_m \geq A_{mm} z_m$. Dies ist aber nur möglich, wenn $z_j = z_m$ für alle j mit $A_{mj} \neq 0$ gilt. Damit ist auch z_j minimal und wir können das gleiche Argument auf z_j anwenden, wegen der Irreduzibilität erreichen wir dann schrittweise alle Einträge von z. Dies bedeutet der konstante Vektor $z_j = 1$ für alle j ist eine Lösung, was aber der Bedingung

$$A_{kk} \;>\; \sum_{j\neq k} |A_{jk}| \;=\; -\sum_{j\neq k} A_{jk},$$

für ein k gilt. Also muss $\min_k z_k \geq 0$ gelten. Analog können wir aber auch $\max_k z_k \leq 0$ zeigen, also bleibt nur $z \equiv 0$ im Nullraum.

Um zu zeigen, dass die Inverse von A nur nicht-negative Einträge hat, können wir zeigen, dass die Lösung z von $Az = y$, mit y gleich einem Einheitsvektor, nur nicht-negative Einträge hat. Dies ist natürlich äquivalent dazu, dass die Lösung von $Az = y$ mit nicht-negativem y nur nicht-negative Einträge hat. Dazu benutzen wir das selbe Argument wie oben: Sei $z_k = \min_j z_j < 0$. Dann ist

$$A_{kk} z_k \;=\; -\sum_{j\neq k} A_{kj} z_j + y_k \geq -\sum_{j\neq k} A_{kj} z_k \;\geq\; A_{kk} z_k$$

mit Gleichheit wenn $y_k = 0$ und $x_j = x_k$ für alle j. Dann folgt aber auch $y_j = 0$ für alle j und deshalb $z = 0$, ein Widerspruch zu $z_k < 0$. □

Konvergenz von Differenzenverfahren

Um die Konvergenz eines numerischen Differenzenverfahrens für ein Randwertproblem zu untersuchen, benötigen wir wieder die üblichen Begriffe der Konsistenz und Stabilität. Diese wollen wir im Folgenden einführen bevor wir hinreichende Bedingungen angeben. Sei $u \in C^2([0,1])$ die Lösung des Randwertproblems und dementsprechend sei $U := (u(x_k)) \in \mathbb{R}^{N-1}$ die Auswertung der Lösung auf dem Diskretisierungsgitter. Sei außerdem $U_h \in \mathbb{R}^{N-1}$ die numerische Lösung des linearen Gleichungssystems $AU_h = F$ des Randwertproblems. Dann können wir den Unterschied zwischen der numerischen Lösung und der echten Lösung der Differentialgleichung bezüglich der rechten Seite F betrachten mit

$$A(U_h - U) \;=\; F - AU \;=:\; G \in \mathbb{R}^{N-1}. \tag{4.15}$$

Die rechte Seite G misst also die Auswirkung des Diskretisierungsfehlers, der durch das Aufstellen der Systemmatrix A entsteht, bei Anwendung auf die echte Lösung der Differentialgleichung. Für die in (4.14) betrachtete Diskretisierung ergibt sich somit für die rechte Seite G unter Verwendung von $F_k = f(x_k) = (a \cdot u')'(x_k)$ für $k = 1, \ldots, N-1$

$$G_k \;=\; (au')'(x_k) - \frac{1}{h_k}\left(a(x_{k+1/2}) \frac{u_{k+1} - u_k}{x_{k+1} - x_k} - a(x_{k-1/2}) \frac{u_k - u_{k-1}}{x_k - x_{k-1}} \right).$$

Wir beachten dabei, dass der Term $c(x) \cdot u(x)$ als Funktionsauswertung exakt diskretisiert wird und daher nicht weiter zum Fehler beiträgt. Hierauf aufbauend können wir also den Konsistenzfehler definieren.

DEFINITION 4.9: Konsistenzfehler und -ordnung.

Sei $A \in \mathbb{R}^{N-1\times N-1}$ eine Systemmatrix, die ein Differenzenverfahren für ein Randwertproblem in Divergenzform (4.4) realisiert. Dann definieren wir den **globalen Konsistenzfehler** des Verfahrens als

$$K_h \;:=\; \|G\|_\infty \;=\; \max_{k=1,\dots,N-1} |G_k| \tag{4.16}$$

Ein Differenzenverfahren heißt **konsistent von der Ordnung p** $\in \mathbb{N}$, wenn für den Konsistenzfehler $K_h = \mathcal{O}(h^p)$ gilt.

Wie wir bereits in (4.11) nachgerechnet haben für den Spezialfall einer konstanten Funktion $a(x) \equiv 1$ für $x \in [0,1]$ ist das diskutierte Differenzenverfahren konsistent von Ordnung 2. Dieses Ergebnis lässt sich leicht mittels Taylorentwicklung für allgemeine Funktionen $a(x)$ erweitern.

Um eine Konvergenz des Differenzenverfahrens zu erhalten benötigen wir eine Abschätzung an die Inverse von A, denn ist A invertierbar so gilt für den Konvergenzfehler E_h mit der Submultiplikativität der Matrixnorm:

$$E_h \;:=\; \|U_h - U\|_\infty \;=\; \|A^{-1} \cdot \underbrace{A(U_h - U)}_{=G}\|_\infty \;\le\; \|A^{-1}\|_\infty \cdot \|G\|_\infty \;=\; \|A^{-1}\|_\infty \cdot K_h.$$

Wir sehen aus dieser Abschätzung sofort, dass aus Konsistenzordnung $p \in \mathbb{N}$ auch schon direkt Konvergenzordnung p folgt, wenn $\|A^{-1}\|_\infty$ beschränkt ist für $N \to \infty$ bzw. $h \to 0$ und somit $E_h \in \mathcal{O}(h^p)$ gilt. Dies bezeichnen wir folglich als Stabilität und halten dies entsprechend in einer Definition fest.

DEFINITION 4.10: Stabilität von Differenzenverfahren.

Wir nennen ein Differenzenverfahren eines Randwertproblems **stabil**, falls für den Konvergenzfehler folgende Abschätzung gilt:

$$E_h \;=\; \|U_h - U\|_\infty \;\le\; C \cdot K_h,$$

wobei K_h der globale Konsistenzfehler aus Definition 4.9 ist und $C > 0$ eine von der Diskretisierungsschrittweite $h := \max_{k=1,\dots,N-1} |x_{k-1} - x_k|$ unabhängige Konstante.

Wie wir sehen werden ist die M-Matrix Eigenschaft eine hinreichende Bedingung für die Stabilität eines Differenzenverfahrens. Wie beim Maximumsprinzip für die Differentialgleichung in Abschnitt 4.1 werden wir Vergleichslösungen suchen, um Fehlerschranken zu erhalten. Dazu ist es wichtig, dass die Vergleichslösung unabhängig von der Schrittweite $h > 0$ (bzw. der Anzahl der Gitterpunkte $N \in \mathbb{N}$) ist.

Im Folgenden sei $v \in C^2([0,1])$ die Lösung des Randwertproblems mit *homogenen Dirichlet-Randwertbedingungen*

$$\begin{aligned} -(a(x)\cdot v'(x))' + c(x)\cdot v(x) &= f(x), \qquad x \in (0,1), \\ v(0) = v(1) &= 0. \end{aligned}$$

Außerdem sei $u \in C^2([0,1])$ die Lösung des Randwertproblems mit *allgemeinen Dirichlet-Randwertbedingungen*

$$\begin{aligned} -(a(x)\cdot u'(x))' + c(x)\cdot u(x) &= f(x), \qquad x \in (0,1), \\ u(0) = g_0, \quad u(1) &= g_1. \end{aligned}$$

Die numerische Lösung von $AU_h = F$ bezeichnen wir mit $U_h = (u_k)_{k=1,\dots,N-1}$. Mit dem üblichen Konsistenzfehler G gilt

$$A(U - U_h) = G,$$

wobei $U = (u(x_k))_{k=1,\dots,N-1}$ die Auswertung der Lösung des allgemeinen Dirichletproblems an den Gitterpunkten ist. Nun betrachten wir die beiden Vektoren

$$H_\pm := U - U_h \pm 2\cdot K_h \cdot V$$

mit dem globalen Konsistenzfehler $K_h = \|G\|_\infty$ und der Lösung des homogenen Dirichletproblems ausgewertet an den Gitterpunkten $V = (v(x_k))_{k=1,\dots,N-1}$. Das folgende Lemma liefert uns praktische Abschätzungen.

LEMMA 4.11: Abschätzungen durch Vergleichslösungen.

Mit den obigen Definitionen von $U, U_h, V \in \mathbb{R}^{N-1}$ und dem globalen Konsistenzfehler K_h betrachten wir die beiden Vektoren $H_\pm := U - U_h \pm 2\cdot K_h \cdot V \in \mathbb{R}^{N-1}$. Sei außerdem $A \in \mathbb{R}^{N-1\times N-1}$ eine M-Matrix.
Dann gelten für hinreichend kleine Schrittweiten $h > 0$ die Ungleichungen

$$\begin{aligned} A\cdot H_+ &= A\cdot(U - U_h + 2\cdot K_h\cdot V) \geq 0, \\ A\cdot H_- &= A\cdot(U - U_h - 2\cdot K_h\cdot V) \leq 0. \end{aligned} \tag{4.17}$$

Die Ungleichungen (4.17) sind hierbei komponentenweise gemeint.

Beweis. Wegen der Konsistenz des Verfahrens wird für hinreichend kleine Schrittweiten $h > 0$ die Differenz

$$A\cdot V - (-(av')'(x_k) + c(x_k))_{k=1,\dots,N-1} = A\cdot V - (1)_{k=1,\dots,N-1}$$

beliebig klein, insbesondere gilt dann

$$A\cdot V \geq (\frac{1}{2})_{k=1,\dots,N-1}.$$

Setzen wir nun ein, so folgt

$$A \cdot (U - U_h + 2 \cdot K_h \cdot V) \;=\; G_h + 2 \cdot K_h \cdot A \cdot V \;\geq\; G + K_h(1)_{k=1,\ldots,N-1} \;\geq\; 0$$

und da $K_h = \|G\|_\infty$ gilt. Analog folgt die zweite Ungleichung. □

Nun können wir eine hinreichende Bedingung für die Stabilität eines Differenzenverfahrens formulieren.

THEOREM 4.12: Stabilität eines Differenzenverfahrens.

Sei $A \in \mathbb{R}^{N-1\times N-1}$ eine Systemmatrix, die ein Differenzenverfahren für ein Randwertproblem in Divergenzform (4.4) realisiert. Seien $U, U_h, V \in \mathbb{R}^{N-1}$ definiert wie oben und K_h sei der globale Konsistenzfehler. Außerdem sei A eine M-Matrix.
Dann ist das Differenzenverfahren stabil.

Beweis. Um die Stabilität des Differenzenverfahrens zu zeigen betrachten wir zunächst wieder die Vektoren

$$H_\pm \;:=\; U - U_h \;\pm\; 2 \cdot K_h \cdot V.$$

Da A eine M-Matrix ist nach Voraussetzung können wir Lemma 4.11 anwenden und es gelten somit die Abschätzungen (4.17). Wegen der Nichtnegativität der Matrixeinträge der Inversen A^{-1} können wir dann komponentenweise folgern

$$\begin{aligned} H_+ \;&=\; U - U_h + 2 \cdot K_h \cdot V \;\geq\; A^{-1} \cdot 0 \;=\; 0, \\ H_- \;&=\; U - U_h - 2 \cdot K_h \cdot V \;\leq\; A^{-1} \cdot 0 \;=\; 0, \end{aligned}$$

Somit können wir also den Fehler zwischen der Lösung der Differentialgleichung U und der numerischen Lösung des Differenzenverfahren U_h nach oben und unten abschätzen durch:

$$-2 \cdot K_h \cdot V \;\leq\; U - U_h \;\leq\; 2 \cdot K_h \cdot V.$$

Da die Lösung $v \in C^2([0,1])$ insbesondere stetig ist, nimmt sie ihr betragliches Maximum auf dem kompakten Intervall $[0,1] \subset \mathbb{R}$ an und wir können somit ebenfalls unabhängig von der Schrittweite $h > 0$ des Gitters abschätzen:

$$\|V\|_\infty \;\leq\; \max_{x\in[0,1]} |v(x)|.$$

Mit diesen Abschätzungen erhalten wir schließlich eine gleichmäßige Schranke für den Konvergenzfehler E_h durch:

$$E_h \;=\; ||U_h - U||_\infty \;\leq\; 2 \cdot ||V||_\infty \cdot K_h \;\leq\; \underbrace{2 \cdot \max_{x\in[0,1]} |v(x)|}_{=:\, C > 0} \cdot K_h.$$

□

Mit der Stabilität eines Differenzenverfahrens erhalten wir als direkte Folgerung das folgende Konvergenzresultat.

KOROLLAR 4.13: Konvergenz eines Differenzenverfahrens.

Seien $U, U_h, V \in \mathbb{R}^{N-1}$ definiert wie oben und K_h sei der globale Konsistenzfehler. Dann gilt für h hinreichend klein

$$E_h \;=\; \|U_h - U\|_\infty \;\leq\; 2 \cdot \|v\|_\infty \cdot K_h.$$

Insbesondere stimmen die Konsistenz- und Konvergenzordnung überein.

Wir sehen, dass die obige Theorie auch leicht auf andere Differenzenverfahren anwendbar ist, solange Konsistenz vorliegt und die entsprechende M-Matrix Eigenschaft erfüllt ist. So kann man die Aussagen auch auf partielle Differentialgleichungen erweitern, etwa wenn wir

$$-\nabla \cdot (a(x)\nabla u(x)) + c(x) \cdot u(x) \;=\; f(x), \qquad x \in [0,1]^2$$

auf einem rechteckigen Gebiet lösen wollen, mit $\nabla \cdot (a(x)\nabla u(x)) = \sum_{i=1}^{n} \partial_{x_i}(a(x)\partial_{x_i} u(x))$. Legen wir darüber ein Gitter und diskretisieren die zweiten Ableitungen in jeder Richtung analog, so erhalten wir wieder ein konsistentes Verfahren, das durch ein lineares System mit einer M-Matrix beschrieben wird. Damit können wir eine völlig analoge Theorie durchführen und Konvergenz beweisen. Der einzige kleine Unterschied ist, dass wir keine Tridiagonalmatrix erhalten, sondern eine etwas allgemeinere dünnbesetzte Matrix. Dies ändert aber wenig an der Struktur, nur die numerische Lösung des Gleichungssystems $AU = F$ wird deutlich aufwändiger, da wir viel mehr Gitterpunkte benötigen. Wir werden uns deshalb später noch mit der numerischen Lösung dünnbesetzter linearer Systeme beschäftigen. Der einzige Nachteil in mehreren Dimensionen ist, dass finite Differenzen kanonisch für rechteckige Gitter verwendbar sind. Hat man andere Gebiete auf denen man partielle Differentialgleichungen lösen will, kommen eher Verfahren wie finite Elemente zum Einsatz, deren Idee wir im Folgenden noch kurz diskutieren wollen.

Literatur

Books

[NW99] J. Nocedal und S. J. Wright. *Numerical Optimization*. 1. Aufl. Springer Verlag, New York, 1999.

Articles

[Dav59] W. C. Davidon. „Variable Metric Method for Minimization“. In: (1959).

[BFGS70] C. G. Broyden. „The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations“. In: *IMA Journal of Applied Mathematics* 6.1 (1970), S. 76–90. eprint: https://academic.oup.com/imamat/article-pdf/6/1/76/2233756/6-1-76.pdf.

[HS52] M. R. Hestenes und E. Stiefel. „Methods of Conjugate Gradients for Solving Linear Systems“. In: *Journal of Research of the National Bureau of Standards* 49.6 (1952).

[She94] J. R. Shewchuk. „An Introduction to the Conjugate Gradient Method Without the Agonizing Pain“. In: (1994).

Online

[N-Körper] Wikipedia. *Das N-Körper-Problem*. URL: https://de.wikipedia.org/wiki/N-K%C3%B6rper-Problem (besucht am 18. 04. 2023).

[Numerik 1] D. Tenbrinck und T. Roith. *Vorlesungsskript zur Einführung in die Numerik (WS 22/23) an der FAU Erlangen-Nürnberg*. URL: https://www.math.fau.de/wp-content/uploads/2023/05/tenbrinck_script_numerik.pdf (besucht am 23. 05. 2023).

[SMW Formel] Wikipedia. *Sherman -Morrison-Woodbury Formel*. URL: https://de.wikipedia.org/wiki/Sherman-Morrison-Woodbury-Formel (besucht am 02. 05. 2023).

[Wacker] P. Wacker. *Mathematik Blog „All About That Bayes“*. URL: https://de.wikipedia.org/wiki/N-K%C3%B6rper-Problem (besucht am 03. 05. 2023).

[Data Science 2] D. Tenbrinck. *Vorlesungsskript zur Mathematik für Data Science 2 (SS 21) an der FAU Erlangen-Nürnberg*. URL: https://www.math.fau.de/wp-content/uploads/2023/05/tenbrinck_script_MfDS2.pdf (besucht am 23. 05. 2023).

[NumAna] F. Wübbeling. *Vorlesungsskript zur Numerischen Analysis (SS 19) an der WWU Münster*. URL: https://www.uni-muenster.de/AMM/num/Vorlesungen/wuebbeling/NumAna2019.pdf (besucht am 28. 06. 2023).

[NumDGL] F. Wübbeling. *Vorlesungsskript zur Analysis und Numerik von gewöhnlichen Differentialgleichungen (SS 22) an der WWU Münster*. URL: https://www.uni-muenster.de/AMM/num/Vorlesungen/wuebbeling/AnaNumDglSS22.pdf (besucht am 28. 06. 2023).