Diskretisierung und numerische Optimierung

Sommersemester 2022

Martin Burger
Department Mathematik, AMMN

Version vom **8.Juni 2022**

Friedrich-Alexander-Universität
Naturwissenschaftliche Fakultät

Department
MATHEMATIK

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

In dieser Vorlesung werden wir einige weiterführende Aspekte der numerischen Mathematik diskutieren, nämlich numerische Verfahren zur Lösung von Optimierungsproblemen und von (gewöhnlichen) Differentialgleichungen.

Der erste Teil der Vorlesung beschäftigt sich mit der Lösung von Differentialgleichungen, insbesondere mit gewöhnlichen Differentialgleichungen. Wir beginnen mit Anfangswertproblemen der Form

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

die nur für spezielle Formen von F gelöst werden können. Allgemeinere Funktionen $F : \mathbb{R}^n \to \mathbb{R}^n$ treten aber in einer Vielzahl von Anwendungen auf, z.B. bei den Newtonschen Gesetzen für die Dynamik von Teilchen. Die entstehenden Systeme sind dann auch beliebig groß, z.B. in der Molekulardynamik, wo $u(t)$ die Koordinaten M verschiedener Teilchen im Zeitverlauf beschreibt ($N = 3M$). Andere klassische Anwendungsgebiete gewöhnlicher Differentialgleichungen sind die Populationsdynamik oder auch die Modellierung von Aktienmärkten, wo meist noch eine zufällige Komponente hinzugefügt wird (stochastische Differentialgleichungen). Die numerischen Verfahren zur Lösung von Anfangswertproblemen sind einerseits ähnlich zu Iterationsverfahren, wenn die Ableitung durch Differenzenquotienten auf einem Gitter approximiert wird, andererseits zu numerischer Integration, wenn man die äquivalente Formulierung als Volterra-Integralgleichung

$$u(t) = u_0 + \int_0^t F(u(s), s)\ ds$$

benutzt und Quadraturformeln auf das Integral anwendet. Ein wichtiger Aspekt ist dabei 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 konvergiert.

Weiterhin werden wir auch Randwertprobleme betrachten, die zu partiellen Differentialgleichungen führen. Ein einfaches Beispiel ist die Lösung von

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

mit Randwerten $u(0) = u(1) = 0$. Hier müssen wir die Diskretisierung auf einmal im ganzen Intervall $(0, 1)$ durchführen und nicht von einem Schritt zum Nächsten wie bei Anfangswertproblemen. Die Diskretisierung liefert hier ein lineares Gleichungssystem, das wir anschließend lösen müssen. Die Abschätzung des Diskretisierungsfehlers erfordert dann weiterführende Methoden.

Abschließend beschäftigen wir uns mit Optimierungsproblemen, welche in vielen mathematischen Anwendungsbereichen auftreten, von klassischen ökonomischen Problemen über Materialoptimierung bis hin zu modernen Problemen in der Bildverarbeitung und im maschinellen Lernen. 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 anpassen um tatsächlich durch die Iteration die Funktionswerte zu verkleinern. Darüber hinaus werden wir geeignete Wahlen der Schrittweite kennenzulernen, um besser 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 Verfahren kennenlernen.

Kapitel 2

Numerische Lösung von Anfangswertproblemen

Im Folgenden wollen wir uns mit der numerischen Lösungen von Anfangswertproblemen für gewöhnliche Differentialgleichungen der Form

$$u'(t) = \frac{du}{dt} = F(t, u(t)), \quad u(0) = u_0 \tag{2.1}$$

beschäftigen, wobei $F : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^n$ eine gegebene stetige Funktion ist. Die Modellierung mit Differentialgleichungen ist oft kanonisch, da man nur verstehen muss wie sich eine Größe in hinreichend kleiner Zeit ändern wird, d.h. wir schreiben

$$u(t + \Delta t) \approx u(t) + \Delta t F(t, u(t)),$$

wobei der Unterschied zur Exaktheit in dieser Relation dann von höherer Ordnung in Δt sein kann. Im Grenzwert $\Delta t \to 0$ erhalten wir dann die Differentialgleichungen. Einfache Heuristiken für Differentialgleichungen kennen wir beispielsweise aus der Physik,

- Geschwindigkeit ist Änderung des Orts pro Zeit,
- Beschleunigung ist Änderung der Geschwindigkeit pro Zeit,
- Kraft ist Masse mal Beschleunigung.

Beschreibt $x(t)$ den Ort eines Teilchens zur Zeit t, $v(t)$ seine Geschwindigkeit und $a(t)$ die Beschleunigung, dann gilt

$$v(t) = \frac{dx}{dt}(t), \quad a(t) = \frac{dv}{dt}(t), \quad ma(t) = F(x(t), v(t), t),$$

wobei F ein Kraftfeld ist. Diese Gesetze können wir auch als Differentialgleichungen für x und v (oder den Impuls $p(t) = mv(t)$) lesen, wenn wir a eliminieren, es gilt dann

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

Kennen wir den Anfangsort $x(0)$ und die Anfangsgeschwindigkeit $v(0)$, so haben wir ein Anfangswert für ein System von Differentialgleichungen (im $\mathbb{R}^3$ insgesamt sechs Gleichungen für sechs Unbekannte).

In großen Systemen von N Teilchen $x_1, \dots, x_N$ hat man dann Gleichungen für jede Position und die Geschwindigkeiten $v_1, \dots, v_N$, mit

$$\frac{d}{dt}(x_i(t), v_i(t))_{i=1,\dots,N} = (v_i(t), F(x_1(t), v_1(t), \dots x_N(t), v_N(t), t)).$$

Zur Beschreibung realer Vorgänge mit vielen Teilchen (Moleküle, Zellen, Tiere in Herden, Fußgänger, Autos ...) erhält man also schnell beliebig komplexe Systeme von Differentialgleichungen. Unser Ziel in diesem Kapitel ist es die numerische Lösung solcher Differentialgleichungen zu untersuchen, zuvor klären wir aber noch einige Grundlagen der gewöhnlichen Differentialgleichungen um diese auch vernünftig verstehen zu können.

2.1 Theorie von Anfangswertproblemen für gewöhnliche Differentialgleichungen

Wir diskutieren im Folgenden kurz einige theoretische Aspekte gewöhnlicher Differentialgleichungen. Wir beginnen mit einer allgemeinen Theorie der Existenz und Eindeutigkeit. Grundlage dafür ist eine Umformulierung in eine Fixpunktform, sodass wir dann einfach einen passenden Fixpunktsatz anwenden können. Ist $u \in C^1([0,T])$ eine Lösung des Anfangswertproblems (2.1), dann gilt aus dem Hauptsatz der Integralrechnung auch

$$u(t) = u_0 + \int_0^t F(s, u(s))\ ds, \qquad 0 \le s \le T.$$

Dies können wir als Fixpunktgleichung $u = \mathcal{F}(u)$ in einem Banachraum interpretieren. Dafür können wir zwei Arten von Fixpunktsätzen anwenden: die erste Art (Satz von Browder, Schauder oder andere Varianten) basiert auf Kompaktheit, d.h. wenn man Funktionen u in den Operator reinsteckt sind diese topologisch danach in einer schöneren Menge. Insbesondere hat diese Menge dann einen Fixpunkt. In unserem Fall entsteht die Kompaktheit aus dem Satz von Arzela–Ascoli, man kann zeigen, dass für u_n beschränkt die Folge $\mathcal{F}(u_n)$ immer eine konvergente Teilfolge hat. Dies liefert den sogenannten Satz von Peano, der die Existenz einer Lösung für stetiges F garantiert. Da man hier recht abstrakt und über Teilfolgen argumentiert, hat man keine Chance die Eindeutigkeit eines Fixpunkts nachzuweisen.

Die zweite Art an Beweisen basiert eigentlich immer auf dem Banach'schen Fixpunktsatz, den wir hier näher diskutieren wollen. Dazu beachten wir, dass falls F im zweiten Argument Lipschitz-stetig ist (mit Modul L), folgendes gilt

$$\begin{aligned}\|\mathcal{F}(u_1) - \mathcal{F}(u_2)\|_\infty &= \max_{0\le t\le T} |\int_0^t F(s,u_1(s)) - F(s,u_2(s))\ ds| \\ &\le \max_{0\le t\le T} \int_0^t |F(s,u_1(s)) - F(s,u_2(s))|\ ds \\ &\le LT \max_{0\le s\le T} |u_1(s) - u_2(s)| = LT\|u_1 - u_2\|_\infty.\end{aligned}$$

Wir erkennen daraus, dass die Abbildung $\mathcal{F} : C([0,T]) \to C([0,T])$ kontraktiv ist, wenn T klein genug ist, da dann immer $LT < 1$ ist. Damit liefert der Banach'sche Fixpunktsatz die Existenz und Eindeutigkeit einer Lösung in $C([0,T])$ für T hinreichend klein. Da u dann die Stammfunktion der stetigen Funktion $t \mapsto F(t,u(t))$ ist, gilt auch $u \in C^1([0,T])$. Dies ist die erste Version des Satzes von Picard-Lindelöf, die uns die Existenz und Eindeutigkeit für kleine Zeiten liefert. Wir sehen dabei schon einige typische Techniken bei der Behandlung von gewöhnlichen und partiellen Differentialgleichungen: erst formulieren wir das Problme um, sodass weniger oder keine Ableitungen mehr vorkommen und zeigen die Existenz / Eindeutigkeit einer Lösung in einem größeren Raum (hier die stetigen Funktionen). Danach beweisen wir zusätzliche Regularität der Lösung, in unseerem Fall stetige Differenzierbarkeit. Wir beachten, dass wenn F k-mal stetig differenzierbar in beiden Variablen ist, eine Iteration des obigen Arguments sogar liefert, dass u $k+1$-mal stetig differenzierbar ist. Auch die Analyse in kleinen Zeiten ist ein typisches Vorgehen für Anfangswertprobleme.

In unserem Fall können wir aber ein besseres Resultat erreichen, in dem wir einfach die Norm passend wählen. Die Idee dazu liefert zunächst die einfache Gleichung

$$u'(t) = Lu(t)$$

mit $L > 0$. Ist u eine positive Lösung, dann folgt mit der Kettenregel

$$\frac{d}{dt} \log u(t) = L.$$

Dies können wir integrieren zu

$$\log u(t) - \log u(0) = Lt$$

und auflösen als

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

In diesem Fall ist trivialerweise L die Lipschitzkonstante von F und wir sehen, dass wir dann ein exponentielles Wachstum mit e^{Lt} erwarten müssen. Dies ist auch allgemein der Fall, wie das folgende Lemma von Gronwall zeigt:

LEMMA 2.1.

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

$$v(t) \leq a + \int_0^t bv(s)\ dx, \qquad \forall 0 \leq t \leq T$$

mit $a > 0$ und $b \in \mathbb{R}$ erfüllt. Dann gilt

$$v(t) \leq ae^{bt}, \forall 0 \leq t \leq T.$$

Beweis. Wir definieren $w(t) = e^{-bt}v(t) - a$, dann gilt

$$w(t) \leq a(e^{-bt} - 1) + \int_0^t e^{b(s-t)}b(w(s) + a)\ ds = \int_0^t bw(s)\ ds.$$

Aus der Ungleichung bei $t = 0$ folgt $w(0) \leq 0$. Sei T_0 die maximale Zeit zu der $w(t) \leq 0$ für alle $t \leq T_0$ gilt. Ist $T_0 = T$, so sind wir fertig. Ist $T_0 < T$, so gibt es ein hinreichend kleines Zeitintervall $(T_0, T_0 + \delta)$, in dem w positiv ist. Dann folgt für t in diesem Intervall

$$w(t) \leq \int_{T_0}^t bw(s)\ ds \leq \delta b \max_{T_0 \leq s \leq T_0+\delta} w(s)$$

und damit auch

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

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 t gelten. □

Wenden wir das Ergebnis auf eine Differentialgleichung mit Lipschitz-stetigem F an, so folgt

$$|u(t) - u_0| \leq \int_0^t |F(t, u(t)) - F(t, u_0)|\ dt + T \max_{0 \leq t \leq T} |F(t, u_0)| \leq L \int_0^t |u(t) - u_0|\ dt + a.$$

Aus dem Lemma von Gronwall angewandt auf $v(t) = |u(t) - u_0|$ und der Dreiecksungleichung sehen wir, dass $u(t)$ höchstens wie e^{Lt} wächst. Dies legt nahe, die folgende gewichtete Norm zu wählen:

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

Da e^{-Lt} nach oben durch eins und nach unten durch e^{-LT} beschränkt ist, ist dies eine äquivalente Norm im Raum der stetigen Funktionen. Wir wiederholen also unsere Abschätzung an den Fixpunktoperator in dieser Norm

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

Der Operator ist nun also kontraktiv für beliebiges T, da $1 - e^{-LT}$ gilt. Wir haben mit dem Banach'schen Fixpunktsatz also die folgende Version des Satzes von Picard-Lindelöf bewiesen:

THEOREM 2.2.

Sei F stetig und Lipschitzstetig bezüglich der zweiten Variable, dann besitzt das Anfangswertproblem (2.1) genau eine Lösung in $C^1([0,T])$.

Wir werden sehen, dass wir auch bei numerischen Verfahren ähnliche Aussagen und insbesondere ein Version des Lemma von Gronwall benötigen werden, um die Stabilität der Verfahren garantieren zu können. Abstrakt gesehen liefert das Lemma von Gronwall eine Stabilitätsaussage für die Differentialgleichung, in endlicher Zeit kann die Norm der Lösung nicht beliebig schnell wachsen.

Wir betrachten zum Abschluss noch einige spezielle Fälle von gewöhnlichen Differentialgleichungen in denen wir eine explizitere Form der Lösung berechnen können. Die Integration der einfachen linearen Gleichung oben ist ein Spezialfall sogenannter separabler Gleichungen der Form

$$u'(t) = G(u(t))H(t).$$

Für skalare Gleichungen, d.h. $u(t) \in \mathbb{R}$, können wir durch G dividieren (vorausgesetzt dieser Term ist ungleich null) und es gilt

$$\frac{u'}{G(u)} = H(t).$$

Ist g eine Stammfunktion von $\frac{1}{G}$ und h eine Stammfunktion von H, so schreiben wir das als $g'(u)u' = h'$ und integrieren zu $g(u) = h(t) + c$. Die Integrationskonstante c können wir aus dem Anfangswert mit $c = g(u_0) - h(0)$ berechnen. Damit gilt, vorausgesetzt g ist invertierbar

$$u(t) = g^{-1}(h(t) - h(0) + g(u_0)).$$

Ein wichtiger Fall, der uns auch kanonische Beispiele für numerische Verfahren liefert, sind lineare Gleichungen mit konstanten Koeffizienten, d.h.

$$u'(t) = Au(t) + b(t)$$

mit einer gegebenen Matrix $A \in \mathbb{R}^{n\times n}$. Wir betrachten zunächst den homogenen Fall $b = 0$. Ist A diagonalisierbar als $A = B-1DB$ mit Diagonalmatrix D, so können wir analog eine Gleichung für $v = Bu$ betrachten, die dann von der Form $v'(t) = Dv(t)$, d.h. jeder Eintrag erfüllt $v_i'(t) = D_{ii}v_i(t)$ und wir erhalten daraus $v_i(t) = v_i(0)e^{D_{ii}t}$. Die Lösung u erhalten wir wieder durch Multiplikation mit B^{-1}. Im Fall einer Diagonalmatrix ist es naheliegend die Exponentialfunktion einer Matrix e^{Dt} als die Diagonalmatrix mit den Einträgen $e^{D_{ii}t}$ zu definieren. Wir haben dann

$$u(t) = B^{-1}e^{Dt}v(0) = B^{-1}e^{Dt}Bu(0) =: e^{At}u(0).$$

Wir bekommen durch Diagonalisieren der Matrix also eine Definition des Matrixexponentials, die dann eine Lösung des Anfangswertproblems liefert. Im Fall einer nicht diagonalisierbaren Matrix ist ein ähnliches Vorgehen über die Jordan'sche Normalform möglich, was hier aber zu weit führen würde.

Ist $b \neq 0$, so können wir die sogenannte Variation der Konstanten benutzen um eine Lösung auszurechnen. Die Idee dabei ist ein Produktansatz $u(t) = e^{At}w(t)$, anstatt der Konstanten in der homogenen Lösung haben wir also jetzt eine Funktion. Dann gilt mit Produktregel

$$u'(t) = Ae^{At}w(t) + e^{At}w'(t) = Au(t) + e^{At}w'(t),$$

der Vergleich mit der rechten Seite der Differentialgleichung liefert $e^{At}w'(t) = b(t)$ bzw. $w'(t) = e^{-At}b(t)$. Um die Lösung zu erhalten müssen wir also nur $e^{-At}b(t)$ aufintegrieren.

Zum Abschluss betrachten wir noch eine Klasse von Gleichungen, die wir aus dem Gradientenverfahren der Optimierung erhalten, wenn die Schrittweite gegen Null geht. Interpretieren wir die Iterierten x^k als Wert einer Funktion u zur Zeit t, dann schreiben wir das Verfahren als

$$u(t + \alpha^k) = u(t) - \alpha^k \nabla F(u(t)),$$

die Differentialgleichung im Grenzwert ist der sogenannte Gradientenfluss

$$u'(t) = -\nabla G(u(t)).$$

Dieser hat natürlich immer noch eine Abstiegseigenschaft, es gilt

$$G(u)' = \nabla G(u)u' = -\|\nabla G(u)\|^2 = -\|u'\|^2.$$

Damit haben wir immer eine Funktion von u, die sogar gleichmässig in der Zeit beschränkt ist, und nicht exponentiell wächst wie die Norm im schlimmsten Fall. Ist G konvex, dann gilt für einen Minimierer u^* sogar

$$\frac{d}{dt}\|u - u^*\|^2 = -a\langle \nabla G(u) - \nabla G(u^*), u - u^*\rangle \leq 0,$$

d.h. die Norm von u ist gleichmäßig beschränkt.

2.2 Einschrittverfahren für Anfangswertprobleme

Im Folgenden betrachten wir eine einfache Möglichkeit zur Lösung von gewöhnlichen Differentialgleichungen, wir berechnen einfach sukzessive die Lösung zu verschiedenen Zeitschritten. Der Einfachheit halber wählen wir hier uniforme Zeitschritte $t_k = k\tau$, $k \geq 0$ zu einer Schrittweite $\tau > 0$, aber ein analoges Vorgehen ist auch im nicht-uniformen Fall möglich. Wir können dann sukzessive Approximationen $u_\tau(t_k)$ für die Lösung des Anfangswertproblems zu diesen diskreten Zeitschritten berechnen, indem wir die Ableitung durch Differenzenquotienten zu den diskreten Zeitpunkten approximieren oder eine Quadraturformel an diesen Zeitschritten für die zugehörige Integralgleichung ansetzen. Im Falle eines Einschrittverfahrens verwenden wir dabei eine Differenzenformel, die nur einen Schritt weit geht, d.h. zur Berechnung von $u_\tau(t_{k+1})$ wird nur $u_\tau(t_k)$ verwendet. Das einfachste Beispiel eines Einschrittverfahrens ist das Vorwärts-Euler-Verfahren oder explizite Euler-Verfahren

$$u_\tau(t_{k+1}) = u_\tau(t_k) + \tau\ F(t_k, u_\tau(t_k)). \tag{2.2}$$

Ein wenig komplizierter ist schon das Rückwärts-Euler-Verfahren oder implizite Euler-Verfahren

$$u_\tau(t_{k+1}) = u_\tau(t_k) + \tau F(t_{k+1}, u_\tau(t_{k+1})), \tag{2.3}$$

bei dem wir eine Gleichung für den neuen Zeitschritt lösen müssen. Insgesamt sind Einschrittverfahren von der Form

$$u_\tau(t_{k+1}) = u_\tau(t_k) + \tau f_\tau(t_k, u_\tau(t_k), u_\tau(t_{k+1})). \tag{2.4}$$

mit einer sogenannten Verfahrensfunktion f_τ. Hängt f_τ nur von $u_\tau(t_k)$ ab, so heißt das Verfahren **explizit**, da es eine explizite Vorschrift zur Berechnung des nächsten Zeitschritts liefert. Hängt f_τ von $u_\tau(t_{k+1})$ ab, so heißt das Verfahren **implizit**, da es nur eine implizite Bedingung (im Allgemeinen eine nichtlineare Gleichung) für den neuen Zeitschritt liefert. Wir betrachten zunächst einige Beispiele

BEISPIEL 2.3.

1. Das Vorwärts-Euler Verfahren hat die Verfahrensfunktion
$$f_\tau(t_k, u_\tau(t_k)) = F(t_k, u_\tau(t_k)),$$
es ist ein explizites Verfahren. In der Integralform bedeutet das, dass wir die Approximation
$$\int_{t_k}^{t_{k+1}} f(t, u(t))\ dt \approx \tau\ f(t_k, u(t_k))$$
benutzten, d.h. das Integral durch die Intervallänge mal dem Wert am linken Intervallrand annähern.

2. Das Rückwärts-Euler Verfahren hat die Verfahrensfunktion
$$f_\tau(t_k, u_\tau(t_{k+1})) = F(t_{k+1}, u_\tau(t_{k+1})),$$

es ist ein implizites Verfahren. In der Integralform bedeutet das, dass wir die Approximation

$$\int_{t_k}^{t_{k+1}} f(t, u(t))\ dt \approx \tau\ f(t_{k+1}, u(t_{k+1}))$$

benutzten, d.h. das Integral durch die Intervallänge mal dem Wert am rechten Intervallrand annähern.

3. Verwenden wir die summierte Trapezregel zur Approximation des Integrals, so erhalten wir das Crank-Nicholson Verfahren mit der Verfahrensfunktion

$$f_\tau(t_k, u_\tau(t_k), u_\tau(t_{k+1})) = \frac{1}{2}F(t_k, u_\tau(t_k)) + \frac{1}{2}F(t_{k+1}, u_\tau(t_{k+1})).$$

Auch dies ist ein implizites Verfahren.

Wir sehen, dass ein explizites Verfahren sofort wohldefiniert ist, falls 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.

LEMMA 2.4.

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 Modul L_2. Dann existiert für $\tau < \frac{1}{L_2}$ 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

$$E_\tau = \max_{k \in \mathbb{N}} \|u_\tau(t_k) - u(t_k)\| \tag{2.5}$$

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 2.5.

Sei E_τ definiert durch (2.5), 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.

2.2.1 Konsistenz von Einschrittverfahren

Gemäß der obigen Motivation definieren wir den Konsistenzfehler als

$$K_\tau = \max_{k\in\mathbb{N}} \|g_\tau(t_k)\| \tag{2.6}$$

mit

$$g_\tau(t_k) = f_\tau(t_k, u(t_k), u(t_{k+1})) - \frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t, u(t))\ dt, \tag{2.7}$$

wobei u eine Lösung des Anfangswertproblems (2.1).

DEFINITION 2.6.

Sei K_τ definiert durch (2.6), 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 2.7.

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 2.8.

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

$$\begin{aligned}
\|g_\tau(t_k)\| &= \|\frac{1}{2}(\varphi(t_k) + \varphi(t_{k+1})) - \frac{1}{\tau}\int_{t_k}^{t_{k+1}} \varphi(t)\ dt\| \\
&= \frac{1}{2\tau}\|\int_{t_k}^{t_{k+1}} (\varphi(t_k) + \varphi(t_{k+1}) - 2\varphi(t))\ dt\| \\
&\le \frac{1}{2\tau}\|\int_{t_k}^{t_{k+1}} \varphi'(t_k)(t_k + t_{k+1} - 2t) + r_k\ dt\|,
\end{aligned}$$

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.

2.2.2 Stabilität und Konvergenz

Wir widmen uns nun der Frage der Stabilität von Einschrittverfahren. Hierbei verwenden wir eine diskrete Version des Lemmas von Gronwall.

LEMMA 2.9: Diskretes Gronwall Lemma.

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

$$u_0 \leq \alpha \in \mathbb{R}_0^+$$
$$u_k \leq \alpha + \sum_{j=0}^{k-1} \beta_j u_j$$

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 2.10.

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

$$\max_{t_k \leq T} \|u_\tau(t_k)\| \leq M$$

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 $v_k = \|u_\tau(t_k) - u_0\|$

$$\begin{aligned} v_{k+1} &\leq v_k + \tau\|f_\tau(t_k, u_\tau(t_k), u_\tau(t_{k+1})) - f_\tau(t_k, u_0, u_0)\| + \tau\|f_\tau(t_k, u_0, u_0)\| \\ &\leq v_k + \tau L(v_k + v_{k+1}) + \tau C. \end{aligned}$$

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 $\tau \leq \frac{1}{2L}$, d.h. $1-\tau L \geq \frac{1}{2}$, dann folgt

$$v_{k+1} \leq 2(1 + \tau L)v_k + 2\tau C.$$

Das diskrete Lemma von Gronwall impliziert dann die Beschränktheit von v_k. □

Mit einem ähnlichen Beweis können wir auch die Stabilität zeigen.

THEOREM 2.11.

Seien u_τ die Lösung eines Einschrittverfahrens mit Lipschitz-stetiger Verfahrensfunktion f_τ und u die Lösung des Anfangswertproblems (2.1) mit gleichem Anfangswert u_0. Der lokale Konsistenzfehler $g_\tau(t_k)$ und der globale Konsistenzfehler K_τ seien definiert wie oben. Dann existiert eine Konstante C, sodass für τ hinreichend klein gilt:

$$E_\tau = \max_{t_k} \|u_\tau(t_k) - u(t_k)\| \leq C \max_{t_k} \|g_\tau(t_k)\| = CK_\tau.$$

Beweis. Wir definieren $v_k = \|u_\tau(t_k) - u(t_k))\|$, dann gilt wieder mit Dreiecksungleichung und Lipschitz-Stetigkeit von f_τ

$$\begin{aligned} v_{k+1} &= \|u_\tau(t_k) + \tau f_\tau(t_k, u_\tau(t_k), u_\tau(t_{k+1})) - u(t_k) - \int_{t_k}^{t_{k+1}} F(t, u(t))dt\| \\ &\leq v_k + \tau\|f_\tau(t_k, u_\tau(t_k), u_\tau(t_{k+1})) - f_\tau(t_k, u(t_k), u(t_{k+1}))\| \\ &\quad + \tau\|f_\tau(t_k, u(t_k), u(t_{k+1})) - \frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t, u(t))dt\| \\ &\leq v_k + \tau L(v_k + v_{k+1}) + \|g_\tau(t_k)\| \end{aligned}$$

also haben wir

$$\begin{aligned} v_{k+1} &\leq v_k + L\tau(v_k + v_{k+1}) + \|g_\tau(t_k)\| \\ \Rightarrow v_{k+1} &\leq \frac{1 + \tau L}{1 - \tau L} v_k + \max_{t_k} \|g_\tau(t_k)\| \end{aligned}$$

Für $\tau < \frac{1}{2L}$ erhalten wir die gewünschte Schranke wieder direkt aus dem diskreten Lemma von Gronwall. □

Eine direkte Folgerung ist die Äquivalenz von Konsistenz und Konvergenz für Einschrittverfahren.

KOROLLAR 2.12.

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).

Aus der Abschätzung des Konsistenzfehlers sehen wir nun sofort, dass Vorwärts- und Rückwärts-Euler Verfahren konvergent von der Ordnung eins sind, das Crank–Nicholson Verfahren ist konvergent von der Ordnung zwei. Wir widmen uns im Folgenden noch der Frage wie wir Einschrittverfahren höherer Ordnung konstruieren können. Wie wir gesehen haben reicht dazu die Analyse der Konsistenzordnung, wir müssen also Verfahrensfunktionen konstruieren, sodass die Taylorentwicklung einen Rest höhere Ordnung liefert. Dies ist bei denen sogenannten Runge–Kutta Verfahren der Fall, die wir im Folgenden diskutieren werden.

2.2.3 Runge–Kutta Verfahren

Bisher haben wir Verfahren kennengelernt, die auf einzelne Funktionsauswertungen an den t_k und t_{k+1} zurückgreifen. Damit haben wir meist die Konsistenzordnung eins erreicht, als maximale Konsistenzordnung zwei beim Crank–Nicholson Verfahren. Eine höhere Konsistenzordnung ist mit so einem Ansatz nicht möglich. Eine erste Möglichkeit zur Steigerung der Ordnung ist es Ableitungen von F bei t_k und t_{k+1} zu berücksichtigen, womit man offensichtlich die Taylor-Entwicklung besser approximieren und eine höhere Ordnung erreichen kann. Die Berechnung von Ableitungen von F ist jedoch potentiell numerisch aufwändig und instabil, deswegen geht man bei Runge–Kutta Verfahren einen anderen Weg und approximiert durch geschachtelte Funktionsauswertungen. Bei einem Runge–Kutta Verfahren der Stufe s berechnet man zunächst

$$f_i^k = F(t_k + c_i\tau, u_\tau(t_k) + \tau \sum_{j=1}^{s} a_{ij} f_j^k)$$

und die Verfahrensfunktion als

$$f_\tau = \sum_{i=1}^{s} b_i f_i^k.$$

Um sinnvoll in der Zeit vorwärts zu gehen wählt man c_i als aufsteigende Folge und die Matrix (a_{ij}) als untere Dreiecksmatrix, im Fall expliziter Verfahren mit Diagonaleinträgen $a_{ii} = 0$. Die grobe Idee ist die Approximation des Integrals $\int_{t_k}^{t_{k+1}}$ durch Quadratur an Zwischenpunkten im Intervall $[t_k, t_{k+1}]$. Die b_i sind dann die Gewichte der Quadraturformel und die f_i^k approximieren $F(t_k + c_i\tau, u(t_k + c_i\tau))$. Da wir $u_\tau(t_k + c_i\tau)$ ja nicht kennen, benötigen wir Approximationen dafür, die wir wieder durch eine numerische Approximation der Differentialgleichung im Intervall $[t_k, t_k + c_i\tau]$ erhalten - daher die geschachtelte Funktionsauswertung.

Wir beginnen wieder mit den einfachsten Fällen.

BEISPIEL 2.13.

Für $s = 1$ ist die Verfahrensfunktion $b_1 f_1^k$ und

$$f_1^k = F(t_k + c_1\tau, u_\tau(t_k) + \tau a_{11} f_1^k).$$

Wollen wir ein explizites Verfahren durchführen, so ist $a_{11} = 0$, das Verfahren ist also von der Form $f_\tau = b_1 F(t_k + c_1\tau, u_\tau(t_k))$. Für Konsistenz sehen wir sofort, dass $b_1 = 1$ gelten muss, die einzige sinnvolle Wahl für c_1 ist null, da wir sonst F zu einer anderen Zeit auswerten als u. Also erhalten wir das Vorwärts-Euler Verfahren. Im impliziten Fall können wir eine höhere Ordnung erreichen. Wir berechnen für die Lösung u des Anfangswertproblems

$$f_1^k = F(t_k, u(t_k)) + c_1\tau\partial_t F(t_k, u(t_k)) + \tau a_{11}\partial_u F(t_k, u(t_k)) f_1^k + \mathcal{O}(\tau^2).$$

Setzen wir auf der rechten Seite nochmal die führende Ordnung für f_1^k ein, so folgt

$$f_1^k = F(t_k, u(t_k)) + c_1\tau\partial_t F(t_k, u(t_k)) + \tau a_{11}\partial_u F(t_k, u(t_k))F(t_k, u(t_k)) + \mathcal{O}(\tau^2).$$

Andererseits ist

$$\frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t, u(t))\, dt = F(t_k, u(t_k)) + \frac{\tau}{2}\partial_t F(t_k, u(t_k)) + \frac{\tau}{2}\partial_u F(t_k, u(t_k))F(t_k, u(t_k)) + \mathcal{O}(\tau^2).$$

Ein Vergleich der beiden Formeln zeigt, dass wir Ordnung zwei erreichen, wenn $b_1 = 1$, $c_1 = \frac{1}{2}$ und $a_{11} = \frac{1}{2}$ gilt. Die Verfahrensfunktion ist also gegeben durch die Lösung von

$$f_1^k = F(t_k + \frac{\tau}{2}, u_\tau(t_k) + \frac{\tau}{2} f_1^k).$$

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 Rückwärts-Euler-Verfahren bestimmt wird.

BEISPIEL 2.14.

Für $s = 2$ ist die Verfahrensfunktion $b_1 f_1^k + b_2 f_2^k$, wobei im expliziten Fall

$$f_1^k = F(t_k + c_1\tau, u_\tau(t_k)), \quad f_2^k = F(t_k + c_2\tau, u_\tau(t_k) + \tau a_{21} f_1^k).$$

Wieder sehen wir, dass nur $c_1 = 0$ eine sinnvolle Wahl ist, und eine Taylor-Entwicklung liefert

$$b_1 f_1^k + b_2 f_2^k = (b_1 + b_2)F(t_k, u(t_k)) + b_2 c_2\tau\partial_t F(t_k, u(t_k)) + b_2 a_{21}\tau\partial_u F(t_k, u(t_k))F(t_k, u(t_k)) + \mathcal{O}(\tau^2).$$

Vergleichen wir wieder mit der Taylor-Entwicklung von $\frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t,u(t))\ dt$, so folgt

$$b_1 + b_2 = 1, \quad b_2 c_2 = \frac{1}{2}, \quad b_2 a_{21} = \frac{1}{2}.$$

Eine einfache Lösung ist $b_1 = 0$, $b_2 = 1$, $c_2 = a_{21} = \frac{1}{2}$. Dies liefert ein Verfahren der Konsistenzordnung zwei mit der Verfahrensfunktion

$$f_\tau = F(t_k + \frac{\tau}{2}, u_\tau(t_k) + \frac{\tau}{2} F(t_k, u(t_k))).$$

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.

Allgemein erhalten wir das Runge-Kutta Schema kodiert durch die Matrix A und die Vektoren b, c. Diese speichern wir im sogenannten Butcher-Schema

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}.$$

Die Einträge von A, b, c erhalten wir durch ein lineares Gleichungssystem, wenn wir einerseits das Runge-Kutta Verfahren und andererseits das Integral $\frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t,u(t))\ dt$ zur gewünschten Taylor-Entwicklung. Wie wir aus dem obigen Beispiel gesehen haben ist die Lösung meist nicht eindeutig, insbesondere bezüglich c benötigen wir vernünftige Kriterien. Am einfachsten ist ein Kriterium für die Konsistenz, da

$$\sum_{i=1}^{s} b_i f_i^j = \sum_{i=1}^{s} b_i F(t_k, u(t_k)) + \mathcal{O}(\tau)$$

und

$$\frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t,u(t))\ dt = F(t_k, u(t_k)) + \mathcal{O}(\tau).$$

Damit sehen wir:

LEMMA 2.15.

Ein Runge-Kutta Verfahren ist genau dann konsistent, wenn $\sum_{i=1}^{s} b_i = 1$ gilt.

Umgekehrt können wir auch die maximale Ordnung eines Runge-Kutta Verfahrens abschätzen.

LEMMA 2.16.

Ein s-stufiges explizites Runge-Kutta Verfahren sei für alle $F \in C^\infty$ von der Konsistenzordnung $p > 0$. Dann gilt $p \leq s$.

Beweis. Wir wählen $F(t,u) = u$ mit $u_0 = 1$, dann gilt wegen $u = e^t$

$$\frac{1}{\tau}\int_{t_k}^{t_{k+1}} F(t,u(t))\ dt = \frac{1}{\tau}\int_{t_k}^{t_{k+1}} u(t)\ dt = \frac{1}{\tau}\int_{t_k}^{t_{k+1}} u(t_k)\sum_{j=0}^{p}\frac{(t-t_k)^j}{j!}\ dt + \mathcal{O}(\tau^{p+1})$$
$$= \sum_{j=0}^{s}\frac{(t-t_k)^j}{(j+1)!} + \mathcal{O}(\tau^{p+1}).$$

Andererseits sehen wir

$$f_1^k = u(t_k), \quad f_2^k = u(t_k)+\tau a_{21}u(t_k), \quad f_3^k = u(t_k)+\tau a_{31}u(t_k)+\tau a_{32}u(t_k)+\tau^2 a_{32}a_{21}u(t_k), \ldots,$$

d.h. f_i^k ist ein Polynom vom Grad kleiner gleich i in τ, insgesamt ist f_τ ein Polynom vom Grad $s-1$. Damit können wir keine höhere Ordnung erreichen, da der Fehler ab der Ordnung τ^p nicht verschwindet. □

Wir sehen andererseits, dass das entstehende Gleichungssystem im Fall $s = p$ immer lösen können, wie in den Beispielen oben bleibt oft eine Nichteindeutigkeit bezüglich der c_i. Um diese sinnvoll wählen zu können, haben wir bisher mit Kausalität argumentiert, was bei höherer Ordnung auch nicht eindeutig ist. Ein systematischer Zugang ist die Invarianz des Verfahrens gegenüber sogenannter Autonomisierung zu fordern. Eine autonome Differentialgleichung ist von der Form $u'(t) = F(u(t))$, wobei die Funktion F nicht explizit von t abhängt. Unser allgemeines Anfangswertproblem (2.1) können wir autonomisieren, d.h. in ein äquivalentes System autonomer Differentialgleichungen für $\tilde{u} = (u,v)$ umschreiben, nämlich

$$u'(t) = F(v(t),u(t)), \quad v'(t) = 1,$$

mit den Anfangswerten $u(0) = u_0$, $v(0) = 1$. Wir sehen sofort, dass $v(t) = t$ gilt, was die Äquivalenz impliziert. Wir nennen diese Transformation Autonomisierung und fordern, dass das numerische Verfahren dagegen invariant ist, d.h. die Anwendung auf das neue System liefert die selbe Lösung $u_\tau(t)$ wie das ursprüngliche Verfahren, und zwar für jedes mögliche F. Schreiben wir das Verfahren in beiden Fällen hin, so gilt einmal

$$f_i^k = F(t_k + c_i\tau, u(t_k) + \tau\sum_j a_{ij}f_j^k)$$

und andererseits

$$f_i^k = F(v(t_k) + \tau\sum_j a_{ij}, u(t_k) + \tau\sum_j a_{ij}f_j^k), \quad g_i^k = 1.$$

Da wegen der exakten numerischen Integration der linearen Funktion $t \mapsto t$ immer $v(t_k) = t_k$ gilt, sehen wir, dass die Invarianz gegenüber Autonomisierung äquivalent zu

$$c_i = \sum_j a_{ij}$$

ist. Deshalb bestimmen wir die c_i immer aus dieser Gleichung und nur die Einträge a_{ij} aus der Konsistenzbedingung. Damit ist es auch immer möglich ein explizites s-stufiges Runge–Kutta Verfahren der Konsistenzordnung zu konstruieren, im impliziten Fall sogar von der Ordnung $s+1$.

2.3 Mehrschrittverfahren für Anfangswertprobleme

Im Folgenden betrachten wir Mehrschrittverfahren für Anfangswertprobleme, d.h. zur Berechnung von $u_\tau(t_{k+1})$ verwenden wir auch die Werte bei $t_k, \dots, t_{k-s+1}$. Wir nennen s die Stufe des Mehrschrittverfahrens, ein Einschrittverfahren wäre dementsprechend von der Stufe eins. Ein einfaches Beispiel erhalten wir aus der Mittelpunktsregel im Intervall (t_{k-1}, t_{k+1}), d.h.

$$u_\tau(t_{k+1}) = u_\tau(t_{k-1}) + 2\tau F(t_k, u_\tau(t_k)),$$

ein anderes die Simpson Regel

$$u_\tau(t_{k+1}) = u_\tau(t_{k-1}) + \frac{\tau}{3}(F(t_{k+1}, u_\tau(t_{k+1})) + 4F(t_k, u_\tau(t_k)) + F(t_{k-1}, u_\tau(t_{k-1}))).$$

Beide sind von der Form

$$\alpha_s u_\tau(t_{j+s}) + \alpha_{s-1} u_\tau(t_{j+s-1}) + \dots + \alpha_0 u_\tau(t_j) = \tau(\beta_s F_{j+s} + \beta_{s-1} F_{j+s-1} + \dots + \beta_0 F_j), \quad (2.8)$$

mit der Abkürzung $F_k = F(t_k, u_\tau(t_k))$. Ein solches Verfahren heißt lineares Mehrschrittverfahren, diese sind mit Abstand die Gebräuchlichsten und wir werden uns hier darauf einschränken. Damit wir wirklich ein s-schritt Verfahren haben, werden wir immer annehmen, dass $\alpha_s \neq 0$ und $|\alpha_0| + |\beta_0| > 0$ gilt.

Bei der numerischen Berechnung gehen wir analog wie bei Einschrittverfahren vor, wir müssen nur zusätzlich zu $u_\tau(t_k)$ auch die Werte $u_\tau(t_{k-1}), \dots, u_\tau(t_{k-s+1})$ speichern. Ein effektiver Unterschied zu Einschrittverfahren sind die Anfangswerte. Um ein Verfahren mit $s > 1$ Schritten durchzuführen benötigen wir nicht nur u_0, sondern auch $u_\tau(t_1), \dots, u_\tau(t_{s-1})$. Diese müssen wir durch ein anderes Verfahren, etwa ein Einschrittverfahren, erst berechnen. Dabei müssen wir natürlich drauf achten, dass dieses Verfahren von der selben Konvergenzordnung wie das Mehrschrittverfahren gewählt wird, um diese insgesamt nicht zu verkleinern.

Bei expliziten Verfahren, d.h. $\beta_s = 0$, ist die Wohldefiniertheit der einzelnen Schritte dann klar, im impliziten Fall benötigen wir das übliche Fixpunktargument um die Wohldefiniertheit zu gewährleisten:

LEMMA 2.17.

Sei F stetig und bezüglich dem zweiten Argument Lipschitz-stetig mit Modul L. Dann existiert eine eindeutige Lösung $u_\tau(t_{j+s})$ von (2.8), falls $\tau < \frac{|\beta_s|}{|\alpha_s|} L$ gilt.

Um die Notation zu vereinfachen, definieren wir den Shift-Operator E_τ für eine zeitabhängige Funktion als $E_\tau u(t) = u(t + \tau)$. Damit können wir das Mehrschrittverfahren als

$$\alpha_s E_\tau^s u_\tau(t_k) + \alpha_{s-1} E_\tau^{s-1} u_\tau(t_k) + \dots \alpha_0 E_\tau^0 u_\tau(t_k) = \tau(\beta_s E_\tau^s F(t_k, u_\tau(t_k)) + \dots + \beta_0 E_\tau^0 F(t_k, u_\tau(t_k)))$$

schreiben. Definieren wir die Polynome

$$\rho(x) = \sum_{j=0}^{s} \alpha_j x^j, \quad \sigma(x) = \sum_{j=0}^{s} \beta_j x^j,$$

so haben wir noch kompakter

$$\rho(E_\tau)u_\tau(t_k) = \tau\sigma(E_\tau)F(t_k, u_\tau(t_k))).$$

Wie wir sehen werden können die Eigenschaften des Verfahrens alleine über die Eigenschaften der Polynome ρ und σ charakterisiert werden. Wir beachten, dass ρ gemäß unserer Voraussetzungen immer Grad s hat, während σ genau dann Grad s hat, wenn das Verfahren implizit ist.

2.3.1 Konsistenz von Mehrschrittverfahren

Während wir die Konvergenzordnung analog zu Einschrittverfahren definieren können, benötigen wir eine passende Definition für Mehrschrittverfahren. Wir führen dazu den lokalen Konsistenzfehler

$$G_\tau(t) = |\frac{1}{\tau}\rho(E_\tau)u(t) - \sigma(E_\tau)F(t, u(t))|$$

ein und definieren:

DEFINITION 2.18.

Ein lineares Mehrschrittverfahren, heißt konsistent, falls

$$K_\tau := \max_{0\le t\le T-s\tau} G_\tau(t)$$

gegen Null konvergiert für $\tau \to 0$. Das Verfahren heißt konsistent von der Ordnung p, wenn $K_\tau = \mathcal{O}(\tau^p)$ für $\tau \to 0$.

Wir können zunächst ein einfaches Resultat zur Charakterisierung der Konvergenz herleiten:

LEMMA 2.19.

Ein lineares Mehrschrittverfahren ist genau dann konsistent von der Ordnung p, wenn (mit der Konvention $0^0 = 1$)

$$\sum_i \alpha_i i^m = m\sum_i \beta_i i^{m-1}, \qquad m = 0, 1, \ldots, p.$$

Beweis. Wir haben

$$\rho(E_\tau)u(t) = \sum_i \alpha_i u(t) + \sum_i \alpha_i(i\tau)u'(t) + \frac{1}{2}\sum_i \alpha_i(i\tau)^2 u''(t) + \ldots$$

und

$$\tau\sigma(E_\tau)u(t) = \tau\sum_i \beta_i u'(t) + \sum_i \alpha_i i\tau^2 u''(t) + \ldots.$$

Ein Koeffizientenvergleich liefert das gewünschte Resultat. □

Wir sehen also, dass Konsistenz sehr einfach an den Koeffizienten ablesbar ist, bzw. wir auch Gleichungssysteme für die Koeffizienten lösen können. Im Fall $s = 1$ bleibt nur $\alpha_0 = -\alpha_1$, wir können diese ohne Beschränkung der Allgemeinheit auf $\alpha_1 = 1$ und $\alpha_0 = -1$ normieren. Um Konsistenzordnung eins zu erreichen, muss $1 = \alpha_1 = \beta_1 + \beta_0$ gelten. Dies beinhaltet u.a. das Vorwärts-Euler Verfahren ($\beta_1 = 0, \beta_0 = 1$), das Rückwärts-Euler Verfahren ($\beta_1 = 1, \beta_0 = 0$) und da Crank-Nicholson Verfahren ($\beta_1 = \frac{1}{2}, \beta_0 = \frac{1}{2}$). Die Konsistenzbedingungen können wir auch über die ersten p Ableitungen des Polynoms ρ and der Stelle $x = 1$ und die ersten $p - 1$ Ableitungen des Polynoms $\sigma = 1$ schreiben. Die Bedingungen für $m = 0$ und $m = 1$ sind

$$\rho(1) = 0, \quad \rho'(1) = \sigma(1).$$

Für $s = 2$ und Konsistenzordnung 2 haben wir

$$\alpha_0 + \alpha_1 + \alpha_2 = 0, \quad \alpha_1 + 2\alpha_2 = \beta_0 + \beta_1 + \beta_2, \quad \alpha_1 + 4\alpha_2 = 2\beta_1 + 4\beta_2.$$

Wir sehen, dass die Mittelpunktsregel ($\alpha_0 = -1, \alpha_1 = 0, \alpha_2 = 1, \beta_0 = \beta_2 = 0, \beta_2 = 2$) eine Lösung dieses Systems liefert. Selbst wenn wir $\alpha_2 = 1$ normieren, können wir noch zwei weitere Gleichung aufstellen um Konsistenzordnung vier zu erreichen, nämlich

$$\alpha_1 + 8\alpha_2 = 3\beta_1 + 12\beta_2, \quad \alpha_1 + 16\alpha_2 = 4\beta_1 + 32\beta_2.$$

Insgesamt können wir für ein s-Schritt Verfahren die Konsistenzordnung $2s+1$ erreichen. Natürlich stellt sich dann die Frage, ob wir dann auch Stabilität und damit die gleiche Konvergenzordnung erreichen können.

2.3.2 Stabilität von Mehrschrittverfahren

Wir widmen uns im Folgenden der Stabilitätsanalyse von Mehrschrittverfahren und werden sehen, dass diese vor allem von den Eigenschaften des Polynoms ρ abhängt. Dazu müssen wir verstehen, wie die Lösung einer linearen Differenzengleichung der Form

$$\sum_{j=0}^{s} \alpha_j v_{k+j} = g_k$$

mit gegebener rechte Seite g_k aussieht. Wir werden dies über dem Körper der komplexen Zahlen tun, da wir dort die Lösung explizit berechnen können, der reelle Fall ist dann ein Spezialfall. Zunächst sehen wir, dass für $v_0, \ldots, v_{s-1}$ gegeben eine eindeutige Lösung v_k für $k \geq s$ existiert. D.h. der Nullraum des linearen Gleichungssystems hat die Dimension s und wir können ihn durch s Basisvektoren darstellen. Wir suchen also zunächst diese s Basisvektoren als linear unabhängige Lösungen des homogenen Systems.

Zunächst sehen wir, dass wir für jede Nullstelle $\lambda \in \mathbb{C}$ des Polynoms ρ eine Lösung der homogenen Gleichung von der Form

$$v_k = \lambda^k$$

konstruieren können, da

$$\sum_i \alpha_i v_{k+i} = \sum_i \alpha_i \lambda^{k+i} = \lambda^k \rho(\lambda) = 0.$$

Hat ρ nur s einfache Nullstellen, so haben wir daraus bereits eine Basis gefunden. Ist λ eine doppelte Nullstelle, d.h. es gilt auch $\rho'(\lambda) = 0$, so sehen wir, dass auch $v_k = k\lambda^k$ eine homogene Lösung ist, da

$$\sum_i \alpha_i v_{k+i} = \sum_i \alpha_i (k+i)\lambda^{k+i} = k\lambda^k \sum_i \alpha_i \lambda^i + \lambda^{k+1} \sum_i \alpha_i i \lambda^{i-1} = k\lambda^k \rho(\lambda) + \lambda^{k+1} \rho'(\lambda) = 0.$$

Analog gilt für eine q-fache Nullstelle, d.h. $\rho^(i)(\lambda) = 0$ für $i = 0, \ldots, q-1$, dass $v_k = k^i \lambda^k$ eine Lösung ist (Übung). Sind also $\lambda_1, \ldots, \lambda_r$ die unterschiedlichen Nullstellen und $q_1, \ldots, q_r$ ihre Vielfachheiten mit $q_1 + \ldots + q_r = s$, dann haben wir einen vollständige Basis des Nullraums aus solchen Lösungen. D.h. die allgemeine Lösung des homogenen Systems ist von der Form

$$v_k = \sum_{i=1}^{r} \sum_{j=1}^{q_r} c_{ij} k^{j-1} \lambda_i^k,$$

mit Konstanten c_{ij} abhängig von den Anfangswerten $v_0, \ldots, v_{s-1}$. Daraus erhalten wir schon eine Idee für die Stabilität:

- Ist λ_i eine Nullstelle mit $|\lambda_i| > 1$, dann existiert eine Lösung, die mit $k \to \infty$ divergiert.
- Ist λ_i eine Nullstelle mit $|\lambda_i| = 1$ und Vielfachheit größer eins, dann existiert ebenfalls eine divergente Lösung $k\lambda_i^k$.

Diese beiden Bedingungen werden Stabilitätskriterium von Dahlquist genannt. Gilt die Umkehrung, d.h. jede Nullstelle ist betragsmäßig kleiner eins oder sie hat Betrag eins und einfache Vielfachheit, dann folgt

$$\begin{aligned} |v_k| &= \left| \sum_{i=1}^{r} \sum_{j=1}^{q_r} c_{ij} k^{j-1} \lambda_i^k \right| = \sum_{i=1}^{r} \sum_{j=1}^{q_r} |c_{ij}|\ k^{j-1} |\lambda_i|^k \\ &\leq \sum_{i,|\lambda_i|<1} \sum_{j=1}^{q_r} |c_{ij}|\ k^{j-1} |\lambda_i|^k + \sum_{i,|\lambda_i|=1} |c_{i1}| \\ &\leq \gamma \sum_{i=1}^{r} \sum_{j=1}^{q_r} |c_{ij}|, \end{aligned}$$

wobei

$$\gamma = \max_i \max_j \max_k k^{j-1} |\lambda_i|^k < \infty.$$

Um die Stabilität des Verfahrens zu verstehen müssen wir noch das inhomogene Problem verstehen. Bei linearen Mehrschrittverfahren ist g_k eine Linearkombination von

Auswertungen von F, wenn F Lipschitz-stetig ist können wir wie beim Einschrittverfahren Differenzen abschätzen und damit g_k durch ein Vielfaches des Konsistenzfehlers. Da wir keine analogen Aussagen wie das diskrete Lemma von Gronwall für mehrstufige Differenzenverfahren haben, schreiben wir dieses in ein System von Differenzengleichungen erster Ordnung über. Dazu definieren wir $v_k^j = v_{k-j}$ für $j = 0, \ldots, s-1$. Dann gilt

$$\begin{aligned} v_{k+1}^0 &= -\frac{1}{\alpha_s}(\alpha_{s-1} v_k^0 + \ldots + \alpha_0 v_k^{s-1} - g_k) \\ v_{k+1}^j &= v_k^{j-1} \qquad \text{für } j > 1. \end{aligned}$$

In Matrixform ist dann

$$V_{k+1} = AV_k + G_k,$$

mit den Vektoren

$$V_k = (v_k^0, \ldots, v_k^{s-1})^T, \qquad G_k = (g_k, 0, \ldots, 0)^T$$

und der Matrix

$$A = \begin{pmatrix} -\frac{\alpha_{s-1}}{\alpha_s} & -\frac{\alpha_{s-2}}{\alpha_s} & \ldots & -\frac{\alpha_1}{\alpha_s} & -\frac{\alpha_0}{\alpha_s} \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix}.$$

Durch die Entwicklung der Determinante sehen wir leicht, dass

$$\det(\lambda I - A) = \frac{1}{\alpha_s}\rho(\lambda)$$

gilt. Deshalb sind die Eigenwerte von A genau die Nullstellen von ρ mit der gleichen Vielfachheit. Ist A diagonalisierbar, so sehen wir sofort, dass einfache Eigenwerte mit Betrag kleiner gleich eins genau die Stabilität

$$\max_k \|V_k\| \leq C \max_k \|G_k\| = C \max_k |g_k|$$

und damit auch

$$\max_k |v_k| \leq C|g_k|.$$

Im nicht-diagonalisierbaren Fall kann man über die Eigenschaften der Jordanschen Normalform zeigen, dass wir unter den Bedingungen des Stabilitätskriteriums von Dahlquist immer die obige Abschätzung erhalten. Da wir nun wieder ein System von Differenzengleichungen erster Ordnung haben und analog auf der rechten Seite den Konsistenzfehler schreiben können, ist die weitere Abschätzung analog wie bei Einschrittverfahren und wir erhalten folgendes Resultat:

THEOREM 2.20.

Ein lineares Mehrschrittverfahren mit den Polynomen σ und ρ erfülle die Konsistenzbedigungen zur Ordnung p sowie das Stabilitätskriterium von Dahlquist. Dann ist das Verfahren konvergent von der Ordnung p.

Es bleibt aber immer noch die Frage offen welche Konvergenzordnung wir wirklich erreichen können, d.h. welche Konsistenzordnung ein stabiles Verfahren erreichen kann. Es gilt tatsächlich folgende Einschränkung:

THEOREM 2.21.

Ein stabiles s-Schritt Verfahren sei konsistent von der Ordnung p . Dann gilt $p \leq s+2$, wenn s gerade ist und $p \leq s+1$ für s ungerade. Ist $\frac{\beta_s}{\alpha_s} \leq 0$, also insbesondere bei impliziten Verfahren, dann gilt $p \leq s$.

Als Beispiel für Verfahren, die im impliziten Fall die Ordnung $s+1$ und im expliziten die Ordnung s erreichen, betrachten wir die sogenannten Adams-Verfahren. Deren ist es ρ so zu wählen, dass maximale Stabilität erreicht wird und dann σ so zu wählen, dass die Konsistenzordnung maximiert wird. Eine offensichtliche Wahl ist $\rho(x) = (x-1)x^{s-1}$, dann haben wir die Nullstelle 1 mit Vielfachheit eins und sonst die Nullstelle null mit Vielfachheit $s-1$. Bei einem impliziten Verfahren können wir die Parameter $\beta_0, \ldots, \beta_s$ aus einem linearen Gleichungssystem so bestimmen, dass wir die Konsistenzordnung $s+1$ (da die Matrix dieses Systemes eine Vandermonde Matrix ist, ist dieses System auch lösbar). Im expliziten Fall haben wir s Parameter $\beta_0, \ldots, \beta_{s-1}$ zur Verfügung und können damit Konsistenzordnung s erreichen. Wir betrachten die einfachsten Beispiele.

- Für $s = 1$ führt das auf die Konsistenzbedingungen

  $$1 = \alpha_1 + \alpha_0 = \beta_1 + \beta_0, \qquad 1 = 2\beta_1,$$

  damit erhalten wir das Crank-Nicholson Verfahren $\beta_0 = \beta_1 = \frac{1}{2}$. Im expliziten Fall haben wir nur $\beta_0 = 1$, also das Vorwärts-Euler Verfahren.

- Für $s = 2$ haben wir im expliziten Fall

  $$1 = \beta_1 + \beta_0, \qquad 3 = 2\beta_1,$$

  daraus folgt $\beta_1 = \frac{3}{2}, \beta_0 = -\frac{1}{2}$. Im impliziten Fall erhalten wir

  $$1 = \beta_2 + \beta_1 + \beta_0, \qquad 3 = 4\beta_2 + 2\beta_1, \qquad 5 = 12\beta_2 + 4\beta_1,$$

  daraus folgt $\beta_2 = -\frac{1}{4}, \beta_1 = 2, \beta_0 = -\frac{3}{4}$.

2.4 Einige weiterführende Themen

Im Folgenden diskutieren wir noch ein paar Aspekte im Umkreis der Numerik von Einschrittverfahren. Dabei beginnen wir mit einfachen partiellen Differentialgleichungen und gehen dann auch noch zur Verbindung von Optimierung und Differentialgleichungen über.

2.4.1 Transport

Wir betrachten eine einfache lineare Transportgleichung auf dem Gitter $\Omega^h = h\mathbb{Z}$. Die Zustandsvariable $u_j(t)$ beschreibt den Zustand im Punkt jh zur Zeit t, bei einem Transport mit konstanter Geschwindigkeit $v = \frac{h}{\Delta t} > 0$ gilt

$$u_j(t + \Delta t) = u_{j-1}(t).$$

Dies können wir auch als

$$u_j(t + \Delta t) = u_j(t) - \Delta t \frac{v}{h}(u_j(t) - u_{j-1}(t))$$

schreiben, also als Vorwärts-Euler Diskretisierung von

$$u_j'(t) = -\frac{v}{h}(u_j(t) - u_{j-1}(t)).$$

Die rechte Seite hat eine Lipschitz-Konstante der Ordnung $\frac{1}{h}$, also beliebig groß für h klein. Dennoch ist das Verfahren stabil solange $\frac{v\Delta}{h} \leq 1$ gilt, denn dann ist

$$u_j(t + \Delta t) = (1 - \Delta t \frac{v}{h})u_j(t) + \Delta t \frac{v}{h} u_{j-1}(t)$$

und damit per Dreiecksungleichung

$$|u_j(t + \Delta t)| = (1 - \Delta t \frac{v}{h})|u_j(t)| + \Delta t \frac{v}{h}|u_{j-1}(t)| \leq \max\{|u_j(t)|, |u_{j-1}(t)|\}.$$

Daraus folgt sofort

$$\|u(t + \Delta t)\|_\infty \leq \|u(t)\|_\infty.$$

Das Rückwärts-Euler Verfahren liefert hier

$$u_j(t + \Delta t) = u_j(t) - \Delta t \frac{v}{h}(u_j(t + \Delta t) - u_{j-1}(t + \Delta t))$$

und wir wollen nun auch seine Stabilität verstehen. Der Einfachheit halber betrachten wir nur Lösungen mit $u_j(0) = 0$ für $j \leq 0$. Dann gilt dies auch für alle $t > 0$ und wir können ein gestaffeltes System lösen. Es gilt

$$u_1(t + \Delta) = \frac{h}{h + v\Delta t} u_1(t)$$

und für $j > 1$

$$u_j(t + \Delta) = \frac{h}{h + v\Delta t} u_j(t) + \frac{v\Delta t}{h + v\Delta t} u_{j-1}(t + \Delta t).$$

Damit zeigen wir leicht

$$|u_j(t + \Delta)| \leq \max\{|u_j(t)|, |u_{j-1}(t + \Delta t)|\}$$

und somit induktiv

$$|u_j(t + \Delta)| \leq \max_{k \leq j} |u_k(t)|.$$

Die impliziert wieder insbesondere die Stabilitätsabschätzung

$$\|u(t+\Delta t)\|_\infty \leq \|u(t)\|_\infty,$$

in diesem Fall aber ohne jede Beschränkung an den Zeitschritt Δt.

Wir sehen auch, dass wir im Fall $h \to 0$ eigentlich eine partielle Differentialgleichung approximiert haben, nämlich die lineare Transportgleichung

$$\partial_t u(x,t) = -v\partial_x u(x,t),$$

da ja für $h \to 0$ auch

$$\frac{v}{h}(u(jh,t) - u((j-1)h,t)) \to \partial_x u$$

gilt. Im obigen Fall haben wir also die partielle Ableitung in x durch einen Rückwärtsdifferenzenquotienten approximiert. Analog könnten wir auch ein Verfahren mit Vorwärtsdifferenzenquotienten in x aufschreiben, d.h. die gewöhnlichen Differentialgleichungen

$$u_j'(t) = \frac{v}{h}(u_j(t) - u_{j+1}(t)).$$

Hier ist allerdings, unabhängig von der Zeitdiskretisierung, das Differentialgleichungssystem schon instabil. Dies sehen wir mit Lösungen der Form

$$u_j(t) = e^{\alpha t + i\beta j},$$

die man mit $\alpha = i\frac{v}{h}(e^{i\beta} - 1)$ erhält. Falls β keine ganze Zahl ist, erhalten wir immer einen positiven Realteil von α, d.h. die Differentialgleichung ist instabil. Der Grund dafür liegt auch in der ursprünglichen Motivation der Transportgleichung. Mit positiver Geschwindigkeit v beschreibt die Transportgleichung die Ausbreitung in steigender x-Richtung, dies wird durch den Rückwärtsdifferenzenquotienten korrekt abgebildet, während der Vorwärtsdifferenzenquotient genau die entgegengesetzte Richtung verwendet.

2.4.2 Diffusion

Wir betrachten im Folgenden einen einfachen Diffusionsprozess, ein Teilchen führt ein Sprungprozess auf dem Gitter $h\mathbb{Z} \cap [0,1]$ mit periodischen Randbedingungen aus aus, wobei es mit gleicher Wahrscheinlichkeit nach links und rechts springt. Die Wahrscheinlichkeit für einen Sprung in einem kleinen Zeitintervall $(t, t+\Delta t)$ ist gleich $2\alpha\Delta t + \mathcal{O}(\Delta t^2)$. Dann gilt für die Wahrscheinlichkeit $p_j(t)$, dass das Teilchen zur Zeit t im Gitterpunkt jh ist:

$$p_j(t+\Delta t) = p_{j-1}(t)\alpha\Delta t + p_{j+1}(t)\alpha\Delta t + p_j(t)(1 - 2\alpha\Delta t) + \mathcal{O}(\Delta t^2), j = 0, \ldots, N,$$

wobei $p_{-1} = p_N$, $p_{N+1} = p_0$. Mit $\Delta t \to 0$ erhalten wir

$$p_j'(t) = \alpha(p_{j-1}(t) + p_{j+1}(t) - 2p_j(t)).$$

Wir beachten, dass die obige Herleitung wieder auf das Vorwärts-Euler Verfahren

$$p_j(t+\tau) = p_{j-1}(t)\alpha\tau + p_{j+1}(t)\alpha\tau + p_j(t)(1 - 2\alpha\tau)$$

führt. Wieder sehen wir sofort, dass das Verfahren stabil ist für $\alpha\tau < 1$. Dies ist potentiell eine starke Einschränkung an τ, wenn α sehr groß ist. Insbesondere können wir, mit $D = \alpha h^2$ das Verfahren wieder als Ortsdiskretisierung einer partiellen Differentialgleichung

$$\partial_t p = D\partial_{xx} p$$

sehen. Damit gilt Stabilität für $\tau \sim h^2$, was für sehr kleines h problematisch ist.

Verwenden wir ein Rückwarts-Euler Verfahren zur Zeitdiskretisierung, so ergibt sich

$$p_j(t+\tau) = p_j(t) + p_{j-1}(t+\tau)\alpha\tau + p_{j+1}(t+\tau)\alpha\tau - 2p_j(t)\alpha\tau.$$

Dieses Verfahren ist wiederum stabil ohne Schranke an den Zeitschritt. Dies sehen wir folgendermaßen: Sei $P(t) = (p_0(t), \ldots, p_N(t))^T$, dann gilt

$$(I + \alpha\tau B)P(t+\tau) = P(t),$$

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 jedes $P \in \mathbb{R}^{N+1}$

$$P^T BP = \sum_j (p_{j+1} - p_j)^2 \geq 0.$$

Damit ist B positiv semidefinit und für jedes $\tau > 0$ ist $I + \tau B$ positiv definit und invertierbar. Insbesondere erhalten wir eine Stabilitätsabschätzung

$$\|P(t+\tau)\|_2^2 + \alpha\tau P(t+\tau)^T BP(t+\tau) = P(t+\tau)^T P(t) \leq \frac{1}{2}\|P(t+\tau)\|_2^2 + \frac{1}{2}\|P(t)\|_2^2.$$

Wegen der positiven Semidefinitheit von B folgt dann

$$\|P(t+\tau)\|_2 \leq \|P(t)\|_2 \leq \ldots \leq \|P(0)\|_2.$$

Tatsächlich gilt in diesem Fall sogar

$$\min_j p_j(0) \leq \min_j p_j(t) \leq \max_j p_j(t) \leq \max_j p_j(0),$$

ähnliche Argumente werden wir im nächsten Kapitel sehen.

Die numerische Lösung der Differentialgleichung für p_j im Intervall $t \in (0, T)$ ist effizient zur Berechnung der Wahrscheinlichkeit, dass das Teilchen zur Zeit T in verschiedenen Punkten x ist, wenn es zur Zeit $t = 0$ in einem Punkt $x_0 = kh$ ist. Dann benötigen wir die Lösung des Systems für einen Anfangswert mit $p_k(0) = 1$ und $p_j(0) = 0$ für $j \neq k$.

Schwieriger ist die Beantwortung der umgekehrten Frage, d.h der Wahrscheinlichkeit, dass das Teilchen zur Zeit T in einem gewissen Punkt x_0 ist, abhängig davon, wo das Teilchen zur Zeit $t = 0$ war. Dazu müssten wir das Differentialgleichungssystem mit vielen verschiedenen Anfangswerten (gleich eins in jeweils einem Punkt und gleich null sonst) lösen und dann im Punkt x_0 auswerten.

Eine Alternative für die letzte Fragestellung ist die sogenannte adjungierte Methode. Dazu berechnen wir die Lösung des adjungierten Problems

$$q_j'(t) = 2\alpha q_j(t) - \alpha q_{j+1}(t) - \alpha q_{j-1}(t),$$

mit Endwert $q_k(T) = 1$ und $q_j(T) = 0$ für $j \neq k$. Die Lösung dieses Problems ist genauso zu berechnen wie für das ursprüngliche System, dies sehen wir einfach mit der Variablentransformation $s = T - t$, dann haben wir ein Anfangswertproblem mit den gleichen Vorzeichen wie das System für p_j zu lösen. Hat man die Lösung berechnet, dann gilt

$$\begin{aligned} p_k(T) &= \sum_j p_j(T) q_j(T) \\ &= \sum_j p_j(0) q_j(0) + \sum_j \int_0^T (p_j q_j)' \, dt \\ &= \sum_j p_j(0) q_j(0) + \int_0^T \sum_j (p_j q_j' + p_j' q_j) \, dt \\ &= \sum_j p_j(0) q_j(0), \end{aligned}$$

da man leicht sieht, dass $\sum_j (p_j q_j' + p_j' q_j) = 0$ gilt. Nun können wir die obige Frage beantworten, denn wenn wir die Lösung p_j mit Anfangswert $p_\ell(0) = 1$ einsetzen, ist $p_k(T) = q_\ell(0)$. Im Gegensatz zur direkten Berechnung müssen wir nun wieder nur ein Differentialgleichungssystem für die q_j lösen.

Das dahinter liegende Prinzip ist das Folgende: haben wir eine lineare Differentialgleichung

$$u'(t) = A(t)u(t)$$

und interessieren wir uns nicht für die Lösung $u(T)$, sondern nur für eine lineare Funktion $L^T u(T) \in \mathbb{R}$. Dann lösen wir das adjungierte Problem

$$v'(t) = -A(t)v(t)$$

mit dem Endwert $v(T) = L$. Dann gilt

$$\begin{aligned} L^T u(T) = v(T)^T u(T) &= v(0)^T u(0) + \int_0^T (v(t)^T u(t))' \, dt \\ &= v(0)^T u(0) + \int_0^T (v'(t)^T u(t) + v(t)^T u'(t)) \, dt \qquad = v(0)^T u(0), \end{aligned}$$

da $v'(t)^T u(t) + v(t)^T u'(t) = -(A(t)^T v(t))^T u(t) + v(t)^T A(t) u(t) = 0$ gilt. Haben wir die adjungierte Gleichung für v berechnet, können wir $L^T u(T)$ für jeden Anfangswert sofort durch ein Skalarprodukt $v(0)^T u(0)$ berechnen.

Bei einer numerischen Lösung müssen wir natürlich die Differentialgleichung mit einer geeigneten Methode (Ein- oder Mehrschrittverfahren) diskretisieren. Es empfiehlt sich wiederum die adjungierte Gleichung mit einem passenden Verfahren zu lösen um die Eigenschaft der Adjungierten auch im Diskreten zu erhalten. Haben wir für u z.B. ein Vorwärts-Euler Verfahren

$$u(t_{k+1}) = u(t_k) + \tau A u(t_k)$$

verwendet, so liefert das explizite 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 leicht, dass bei anderen Verfahren für die adjungierte Gleichung, z.B. einem impliziten Euler-Verfahren, eine solche Identität nicht erhalten ist. Die Diskretisierung und Adjungierung kommutieren dann nicht. Beim expliziten Euler-Verfahen erhalten wir genau die Adjungierte der Diskretisierung der Differentialgleichung.

2.4.3 Optimierung bei Differentialgleichungen

Ein häufiges Thema in der Praxis ist die Bestimmung von Parametern in gewöhnlichen Differentialgleichungen, d.h. wir haben ein Anfangswertproblem

$$u'(t) = F(t, u(t), w), \quad u(0) = u_0(w),$$

bei dem die rechte Seite und der Anfangswert von Parametern $w \in \mathbb{R}^M$ abhängen. Wir nehmen an, dass F Lipschitz stetig ist, damit existiert für gegebenes w eine eindeutige stetig differenzierbare Lösung, die wir mit u_w bezeichnen. Um die Parameter zu bestimmen, misst man $G(u) \in R^K$, typischerweise mit $K > M$ oder versucht man versucht die Parameter so zu optimieren um einen Zustand $G(u)$ zu erreichen. Häufig sind dies die Werte der Lösung zu verschiedenen Zeiten, also $G(u) = (u(s_1), \ldots, u(s_K))$. Nun kann man bei gegebenen Daten g ein Optimierungsproblem, etwa das Kleinstquadrate-Problem

$$f(w) = \frac{1}{2}\|H(w) - g\|^2 = \frac{1}{2}\|G(u_w) - g\|^2 \to \min_w$$

lösen um die Parameter zu bestimmen. Wir fragen uns wie wir in diesem Fall die Lösung des Optimierungsproblems durch eines der Verfahren in dieser Vorlesung, etwas des Gradientenverfahrens, bestimmen können. Dazu ist die effiziente Berechnung von ∇w essentiell.

BEISPIEL 2.22.

Als einfaches Beispiel betrachten wir einen Fall, wo wir zwar wissen, dass eine lineare Differentialgleichung erfüllt ist, aber nicht mit welcher Steigung der linearen Funktion und auch den Anfangswert nicht kennen. Dies führt auf

$$u_0(w) = w_1, \qquad F(t, u, w) = w_2 u.$$

Hier können wir explizit $u_w(t) = w_1 e^{w_2 t}$ berechnen. Für $G(u) = (u(s_1), \dots, u(s_K))$ erhalten wir dann

$$\partial_{w_1} G(u_w) = (e^{w_2 s_1}, \dots, e^{w_2 s_K}), \quad \partial_{w_2} G(u_w) = (w_1 s_1 e^{w_2 s_1}, \dots, w_1 s_K e^{w_2 s_K}).$$

Wie gehen wir aber vor, wenn wir die Gleichung nicht explizit lösen können ? Dazu betrachten wir zunächst

$$u_w^i := \lim_{\delta \to 0} \frac{u_{w+\delta e_i}(t) - u_w(t)}{\delta},$$

wobei e_i der i-te Einheitsvektor ist. u_w^i ist die partielle Ableitung von u_w nach w_i. Diese Funktion können wir nicht berechnen, aber wir können ein Anfangswertproblem herleiten, das von ihr gelöst wird. Unter der Annahme, dass $w \mapsto u_0(w)$ differenzierbar ist, sehen wir sofort

$$u_w^i(0) = \partial_{w_i} u_0(w),$$

und wenn F bezüglich u und w differenzierbar ist folgt mit der Kettenregel

$$(u_w^i)'(t) = \partial_u F(t, u_w(t), w) u_w^i(t) + \partial_w F(t, u_w(t), w).$$

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

$$\partial_{w_i} H(w) = G'(u_w) u_w^i,$$

daraus bekommen wir also die Jacobi Matrix von H bzw. dann den Gradienten von f per Kettenregel.

Wir rechnen dies nun noch an unserem obigen Beispiel 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 $\partial_{w_i} G(u) = (u_w^i(s_1), \dots, u_w^i(s_K))$. 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 überein.

Ist die Anzahl M der Paramete 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 $G(u) = (u_w(s_1), \dots, u_w(s_K))$. Wir definieren v als die Lösung von

$$v'(t) = -\partial_u F(t, u_w(t), w) v(t), \qquad t \in (0,T) \setminus \{s_1, \dots, s_K\},$$

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 $s_0 = 0$, $s_{K+1} = T$,

$$\begin{aligned}
\partial_{w_i} f(w) &= \sum_k (u_w(s_k) - g_k) u_w^i(s_k) \\
&= \sum_k (\lim_{t \uparrow s_k} v(t) - \lim_{t \downarrow s_k} v(t)) u_w^i(s_k) \\
&= v(0) u_w^i(0) + \sum_{k=0}^{K} \int_{s_k}^{s_{k+1}} (v(t) u_w^i(t))' \, dt \\
&= v(0) u_w^i(0) + \int_0^T (v(t) u_w^i(t))' \, dt \\
&= v(0) \partial_{w_i} u_0(w) + \int_0^T v(t) \partial_{w_i} F(t, u_w(t), w) \, dt.
\end{aligned}$$

Damit genügt zur Berechnung des Gradienten die Lösung einer adjungierten Differentialgleichung, sowie von M Skalarprodukten mit Anfangswerten und M Integralen mit der Lösung v.

2.4.4 Deep Learning

In modernen Anwendungen des machine Learning kommen ähnliche Techniken wie bei Differentialgleichungen und deren Optimierung zum Einsatz. Die Idee dabei ist einen parametrisierten Zusammenhang zwischen Input- und Output Daten zu konstruieren. Dieser Zusammenhang wird beim deep Learning durch ein neuronales Netz mit vielen Schichten (Layer) modelliert, das mathematisch als Hintereinanderausführung von linearen Abbildungen (Austausch von Impulsen zwischen Neuronen) und punktweisen

Nichtlinearitäten (Aktivierung eines Neurons durch die eingelangten Impulse) modelliert. Die Aktivierungsfunktion bezeichnen wir mit $\sigma : \mathbb{R} \to \mathbb{R}$, typische Beispiele sind die Sigmoid-Funktion

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

und die Rectified Linear Unit (ReLU)

$$\sigma(x) = \max\{x, 0\}.$$

Dazu verwenden wir für Vektoren die Notation $\sigma(x)$ als $\sigma(x) = (\sigma(x_i))_{i=1,\dots,n}$.

Ein neuronales Netzwerk mit L Layern modelliert die Relation $x \mapsto y$ dann durch $u_0 = x$,

$$u_{k+1} = \sigma(A_k u_k + b_k), \qquad k = 0, \dots, L-1,$$

mit $A_k \in \mathbb{R}^{n \times n}$ und $b_k \in \mathbb{R}^n$,

$$y = C u_N + d,$$

mit $C \in \mathbb{R}^{O \times n}$, $d \in \mathbb{R}^O$. Während ältere Ansätze von neuronalen Netzen in den 80er und 90er Jahren des zwanzigsten Jahrhunderts meist ein sehr kleines M verwendeten ($M = 1, 2, 3$), benutzt eine modernes tiefes Netz M sehr groß, also ein tiefes Netzwerk (daher der Name deep Learning). Wir sehen also eine gewisse analogie zu expliziten Euler-Verfahren für gewöhnliche Differentialgleichungen, dieser ist noch deutlicher bei sogenannten residualen Netzwerken von der Form

$$u_{k+1} = u_k + \tau\sigma(A_k u_k + b_k), \qquad k = 0, \dots, L-1,$$

die wir direkt als Diskretisierung von

$$u'(t) = \sigma(A(t)u(t) + b(t))$$

interpretieren können.

Das Training einens neuronalen Netzwerks ist nun die optimale Bestimmung der Gewichte $w = (A_k, b_k, C, d)$ aus einer großen Menge an Trainingsdaten $(x_i, y_i)_{i=1,\dots,M}$. Dazu wird ein Minimierungsproblem der Form

$$f((A_k, b_k), C, d) = \frac{1}{M} \sum_{i=1}^{M} \ell(C u_N(x_i; (A_k, b_k)) + d, y_i),$$

wobei ℓ eine Loss-Funktion ist, die den Abstand misst, z.b. einfach

$$\ell(C u_N(x_i; (A_k, b_k)) + d, y_i) = \frac{1}{2} \| C u_N(x_i; (A_k, b_k)) + d - y_i \|^2.$$

Dies ist für grosse $N/O/n$ 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

$$w_{j+1} = w_j - \alpha_j \nabla_w \ell(C u_N(x_{i(j)}; (A_k, b_k)) + d, y_{i(j)}),$$

wobei $i(j) \in \{1, \dots, M\}$ 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

$$v_{k-1} = -(A_k \star \sigma'(A_k u_k + b_k))^T v_k$$

mit $v_N = \nabla_u \ell(Cu_N(x_{i(j)}; (A_k, b_k)) + d, y_{i(j)})$ berechnen. Dies ist in diesem Zusammenhang als Backpropagation bekannt.

Kapitel 3

Numerische Lösung von Randwertproblemen

In diesem Abschnitt wollen wir uns mit der Lösung von Randwertproblemen für lineare gewöhnliche Differentialgleichungen zweiter Ordnung beschäftigen. Diese sind von der Gestalt

$$-u'' + pu' + qu = g \tag{3.1}$$

für $x \in (0,1)$ mit Randbedingungen

$$\alpha_0 u'(0) + \beta_0 u(0) = g_0, \tag{3.2}$$
$$\alpha_1 u'(1) + \beta_1 u(1) = g_1. \tag{3.3}$$

Im Fall $\alpha_i = 0$ spricht man von Dirichlet-Randbedingungen, im Fall $\beta_i = 0$ von Neumann-Randbedingungen.

Mit $a = e^P$, $c = e^P q$ und $f = e^P g$, wobei $P' = p$ ist, können wir dieses Problem in einer sogenannten Divergenzform

$$-(au')' + cu = f \tag{3.4}$$

bringen. Diese hat einige Vorteile bei der Analyse und numerischen Lösung. So könen wir zunächst im Fall $c = 0$ die Green-Funktion zur Lösung konstruieren. Es gilt

$$a(x)u'(x) = c_1 - \int_0^x f(y)\, dy$$

und damit

$$u(x) = c_2 + c_1 \int_0^x \frac{1}{a(z)}\, dz - \int_0^x \frac{1}{a(z)} \int_0^z f(y)\, dy\, dz.$$

Mit $A(x) = \int_0^x \frac{1}{a(z)}\, dz$ und einem Wechsel der Integrale im letzten Term erhalten wir

$$u(x) = c_2 A(x) - \int_0^x (A(x) - A(y)) f(y)\, dy.$$

Die Konstanten c_1 und c_2 können wir aus den Randbedingungen bestimmen, die ein 2×2 Gleichungssystem liefern. Wir betrachten hier nur die reinen Dirichlet- oder Neumann-Randbedingungen, die gemischten Randbedingungen (genannt Robin-Bedingungen) überlassen wir als Übung.

Dirichlet-Randbedingungen

Wir betrachten hier den Fall $\alpha_i = 0$ und $\beta_i = 1$. Dann ist das Gleichungssystem zur Bestimmung der Konstanten

$$c_2 = h_0, \qquad c_2 + c_1 A(1) - \int_0^1 (A(1) - A(y)) f(y)\, dy = h_1.$$

Daraus folgern wir

$$\begin{aligned} u(x) &= h_0 + \frac{A(x)}{A(1)} \left(h_1 - h_0 + \int_0^1 (A(1) - A(y)) f(y)\, dy - \int_0^x (A(x) - A(y)) f(y)\, dy \right) \\ &= \left(1 - \frac{A(x)}{A(1)}\right) h_0 + \frac{A(x)}{A(1)} h_1 + \int_0^1 G_D(x,y) f(y)\, dy, \end{aligned}$$

mit der sogenannten Green-Funktion

$$G_D(x,y) = \begin{cases} A(x)(1 - \frac{A(y)}{A(1)}) & x < y \\ A(y)(1 - \frac{A(x)}{A(1)}) & x > y. \end{cases}$$

Aus dieser Rechnung sehen wir sofort folgendes Resultat:

THEOREM 3.1.

Sei $a \in C^1([0,1])$ mit $a(x) \geq a_0 > 0$ für alle $x \in [0,1]$, $c \equiv 0$ und $f \in C([0,1])$, Dann existiert eine eindeutige Lösung $u \in C^2([0,1])$ des Dirichlet-Problems (3.4), (3.2), (3.3). Ist $f(x) \geq 0$ für alle $x \in [0,1]$, so gilt

$$\min_x u(x) \geq \min\{h_0, h_1\}.$$

Wir sehen, dass der lineare Operator

$$K_D : f \mapsto \int_0^1 G_D(x,y) f(y)\, dy$$

offensichtlich auf dem Raum der stetigen Funktionen wohldefiniert und beschränkt ist. Damit können wir auch den Fall $c \neq 0$ behandeln, da in diesem Fall u eine Lösung der Fixpunktgleichung

$$u = h_0 w + h_1(1 - w) + K_D(f - cu)$$

mit der Notation $w(x) = 1 - \frac{A(x)}{A(1)}$ ist. Um das Verhalten in c besser zu verstehen, betrachten wir zunächst den Fall von c konstant, also lösen wir

$$(I + cK_D)u = h_0 w + h_1(1 - w) + K_D f.$$

Für $c = -k^2\pi^2$, $k \in \mathbb{N}$ existiert eine nichttriviale Lösung $u = \sin(k\pi x)$ des homogenen Systems, in diesem Fall ist der Operator $I + cK_D$ nicht invertierbar. Also können wir in diesen Fällen die inhomogene Gleichung nicht für alle rechten Seiten lösen. Wir sehen, dass es nur abzählbar viele Werte von c gibt, für die $I + cK_D$ nicht invertierbar ist. Dies ist kein Zufall, sondern eine Konsequenz der Spektraltheorie kompakter Operatoren. Mit dem Satz von Arzela-Ascoli kann man zeigen, dass $K_D : C([0,1]) \to C([0,1])$ kompakt ist, d.h. für jede beschränkte Folge u_n hat $K_D u_n$ eine konvergente Teilfolge. Die Spektraltheorie kompakter Operatoren garantiert nun, dass es nur eine abzählbare Menge von λ geben kann, sodass $\lambda I - K_D$ nicht invertierbar ist, dazu ist Null der einzige Häufungspunkt. Wir sehen den Zusammenhang $c = -\frac{1}{\lambda}$, also erhalten wir für die Konstanten c eine abzählbare Menge mit möglichen Häufungspunkten $\pm\infty$.

Wenn wir eine allgemeine Funktion c betrachten, dann können wir immer noch zeigen, dass der Operator $u \mapsto K_D(cu)$ kompakt ist, daraus können wir insbesondere wieder folgern, dass $I + K_D(c\cdot)$ invertierbar ist, wenn sein Nullraum trivial ist. Aus dem Fall von konstantem c sehen wir, dass ein nichttrivialer Nullraum nur bei negativem c auftritt. Dies gilt auch im allgemeinen Fall, Grundlage dafür ist folgende Eigenschaft.

LEMMA 3.2.

Für alle $f \in C([0,1])$ gilt

$$\int_0^1 f(x)(K_D f)(x)\, dx = \int_0^1 \int_0^1 G_D(x,y) f(x) f(y)\, dx\, dy \geq 0,$$

mit Gleichheit genau dann, wenn $f \equiv 0$.

Beweis. Wir sehen, dass (mit $u = K_D f$) gilt

$$\int_0^1 f(x)(K_D f)(x)\, dx = -\int_0^1 (au')' u\, dx = \int_0^1 (u')^2\, dx.$$

Damit folgt sofort, dass die linke Seite immer nichtnegativ ist und Gleichheit genau für $u' \equiv 0$ gilt. Dies ist aber äquivalent zu $f \equiv 0$. □

Damit können wir nun ein Resultat zur Existenz und Eindeutigkeit zeigen.

THEOREM 3.3: S.

i $a \in C^1([0,1])$ mit $a(x) \geq a_0 > 0$ für alle $x \in [0,1]$, $c \in C([0,1])$ mit $c(x) \geq 0$ für alle $x \in [0,1]$ und $f \in C([0,1])$. Dann existiert eine eindeutige Lösung $u \in C^2([0,1])$ des Dirichlet-Problems (3.4), (3.2), (3.3).

Beweis. Nach der obigen Argumentation genügt es zu zeigen, dass der lineare Operator $u \mapsto u + K_D(cu)$ invertierbar ist bzw. dann sogar, dass er injektiv ist. Sei $u + K_D(cu) = 0$, dann gilt mit $cu = -cK_D(cu)$ und dem obigen Lemma angewandt auf $f = cu$

$$\int_0^1 cu^2 dx = -\int_0^1 cuK_D(cu)\, dx \leq 0,$$

mit Gleichheit genau dann wenn $cu = 0$ ist. Gilt $cu = 0$ so folgt aber auch $K_D(cu) = 0$ und damit $u = 0$. Also ist der Nullraum des Operators trivial, woraus die Invertierbarkeit und damit auch die eindeutige Lösbarkeit des Randwertproblems folgt. □

Interessanterweise können wir auch in diesem Fall wieder ein ähnliches Maximumsprinzip zeigen, auch wenn wir keine explizite Darstellung der Lösung als Integral mehr haben. Dies erhält man leicht aus einem Widerspruchsbeweis. Sind c und f positiv und nehmen wir an, dass u im Inneren von $[0, 1]$ sein Minimum annimmt. Dann gilt an einem solchen Punkt $\overline{x}$

$$u'(\overline{x}) = 0, \qquad u''(\overline{x}) \geq 0.$$

Eingesetzt in die Differentialgleichung folgt dann

$$cu \geq -(au')' + cu = f$$

und damit ist u nicht negativ. Die einzige Möglichkeit, dass u negative Werte annimmt, besteht wenn das Minimum am Rand angenommen wird, d.h.

$$u(x) \geq \min\{g_0, g_1, 0\}.$$

Wegen der Linearität des Problems kann man aus solchen Maximumsprinzipien auch Stabilitätsaussagen herleiten. Ist etwa $\tilde{u}$ eine bekannte Lösung für das Problem

$$-(a\tilde{u}')' + c\tilde{u} = \tilde{f},$$

mit Dirichlet Randwerten $\tilde{u}(0) = \tilde{g}_0$ und $\tilde{u}(1) = \tilde{g}_1$. Ist $\tilde{f} \geq f$, so folgt

$$\tilde{u} - u \geq \min\{0, \tilde{h}_0 - h_0, \tilde{h}_1 - h_1\}$$

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 $\tilde{h}_i > h_i$. 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 $|u(x)|$.

Neumann-Randbedingungen

Während allgemeine Randbedingungen sehr ähnlich zu handhaben sind wie der Dirichlet-Fall, gibt es im Fall reiner Neumann-Randbedingungen einen speziellen Aspekt, den wir beachten müssen. Wir betrachten also $\alpha_i = 1$ und $\beta_i = 0$, und dazu den Fall $c \equiv 0$. Dann können wir die Differentialgleichung von links und rechts aufintegrieren zu

$$\begin{aligned} -a(x)u'(x) &= -a(0)g_0 + \int_0^x f(y)\ dy \\ -a(x)u'(x) &= -a(1)g_1 - \int_x^1 f(y)\ dy = -a(1)g_1 - \int_0^1 f(y)\ dy + \int_0^x f(y)\ dy. \end{aligned}$$

Daraus sehen wir sofort, dass wir die Bedingung

$$a(0)g_0 = a(1)g_1 + \int_0^1 f(y)\ dy$$

zur Konsistenz benötigen. Unter dieser Bedingung ist das System lösbar, aber eine der Konstanten ist unbestimmt. Dies wird üblicherweise durch eine zusätzliche Bedingung wie $\int_0^1 u(x)\ dx = 1$ erreicht. Der Grund für diese zusätzliche Konsistenzbedingung ist, dass der Nullraum des Neumann-Problems nichttrival ist, jede konstante Funktion löst das homogene Problem.

Interessanterweise ist die Theorie für $c > 0$ auch hier unterschiedlich. Sobald $c(x) \geq 0$ für alle x und $c \neq 0$ gilt, hat das Neumann-Problem nur noch trivialen Nullraum, wie wir aus Multiplikation der Gleichung mit u und Integration sehen, es folgt dann

$$0 = \int_0^1 (-(au')' + cu)u\ dx = \int_0^1 (a(u')^2 + cu^2)\ dx.$$

Daraus folgt sofort $u' = 0$, also u konstant. Ist c nicht identisch null, so kann $\int_0^1 cu^2\ dx = 0$ bei konstantem u nur für $u = 0$ gelten. Der Unterschied dieser Fälle und der potentielle Nullraum des Neumann-Problems ist natürlich auch bei der numerischen Lösung zu beachten, hier wird sich dieser in der Nichtinvertierbarkeit einer entsprechenden Matrix niederschlagen.

3.1 Differenzenverfahren für Randwertprobleme

Zur Diskretisierung von Randwertproblemen können wir zunächst genauso vorgehen wie bei der Lösung von Anfangswertproblemen. Wir konstruieren ein Gitter $0 = x_0 < x_1 < \ldots < x_N = 1$, im einfachsten Fall äquidistant als $x_k = kh$, $h = \frac{1}{N}$, $k = 0, \ldots, N$. Auf diesem Gitter approximieren wir Ableitungen durch finite Differenzen. Für die zweite Ableitung in x_k verwenden wir dann den (zentralen) Differenzenquotienten

$$u''(x_k) \approx \frac{u(x_k + h) - 2u(x_k) + u(x_k - h)}{h^2},$$

bzw. im nichtäquidistanten Fall

$$u''(x_k) \approx \frac{1}{h_k}\left(\frac{u(x_{k+1}) - u(x_k)}{x_{k+1} - x_k} - \frac{u(x_k) - u(x_{k-1})}{x_k - x_{k-1}}\right).$$

Hier ist zunächst die Frage wie wir h_k (abhängig von x_{k-1}, x_k, x_{k+1}) wählen sollen. Dies können wir als Frage an die maximale Konsistenzordnung formulieren. Für u hinreichend glatt gilt per Taylor-Entwicklung

$$\begin{aligned}\frac{u(x_{k+1}) - u(x_k)}{x_{k+1} - x_k} - \frac{u(x_k) - u(x_{k-1})}{x_k - x_{k-1}} =& u'(x_k) + \frac{1}{2}u''(x_k)(x_{k+1} - x_k) + \frac{1}{6}u'''(x_k)(x_{k+1} - x_k)^2 - \\ & u'(x_k) + \frac{1}{2}u''(x_k)(x_k - x_{k-1}) - \frac{1}{6}u'''(x_k)(x_k - x_{k-1})^2 + \mathcal{O}(h^3) \\ =& u''(x_k)\frac{x_{k+1} - x_{k-1}}{2} + \frac{1}{6}u'''(x_k)(x_{k+1} - 2x_k + x_{k-1})(x_{k+1} - x_{k-1})\end{aligned}$$

mit $h = \max_k |x_{k+1} - x_k|^2$. Wir sehen, dass wir Konsistenz mit der Wahl $h_k = \frac{x_{k+1} - x_{k-1}}{2}$ erreichen. Im äquidistanten Fall haben wir aus der Rechnung dann sogar Konsistenzordnung h^2, da $x_{k+1} - 2x_k + x_{k-1} = 0$ gilt. Die konsistente Wahl von h_k entspricht folgender Interpretation: eigentlich berechnen wir erste Ableitungen

$$\frac{u(x_{k+1}) - u(x_k)}{x_{k+1} - x_k} = u'(x_{k+1/2}) + \mathcal{O}\langle^{\in}$$
$$\frac{u(x_k) - u(x_{k-1})}{x_k - x_{k-1}} = u'(x_{k-1/2}) + \mathcal{O}\langle^{\in}$$

mit den Mittelpunkten

$$x_{k\pm1/2} = \frac{1}{2}(x_{k\pm1} + x_k),$$

also auf einem versetzten neuen Gitter. Nun können wir die zweite Ableitung auch diskret als Ableitung der ersten Ableitung betrachten, da eine weitere Anwendung der Differenzen auf die Werte an den versetzten Gitterpunkten $x_{k\pm1/2}$ genau eine Approximation der zweiten Ableitung bei x_k liefert. Die Schrittweite für diesen zweiten Schritt ist genau das von uns errechnete

$$h_k = \frac{x_{k+1} - x_{k-1}}{2} = \frac{x_{k+1} + x_k}{2} - \frac{x_k + x_{k-1}}{2} = x_{k+1/2} - x_{k-1/2}.$$

Damit können wir nun zumindest Gleichungen der Form

$$-u'' + cu = f$$

direkt diskretisieren. Wir betrachten wieder zunächst das Dirichlet-Problem, das wir als lineares Gleichungssystem für den Vektor

$$U = (u(x_1), \ldots, u(x_{N-1}))$$

schreiben können. Die Werte $u(x_0) = g_0$ und $u(x_N) = g_1$ ergeben sich dann wieder aus den Randbedingungen und können direkt eingesetzt werden. Definieren wir für $k = 1, \ldots, N-1$ die Matrix-Einträge

$$\begin{aligned}
A_{k,k} &= \frac{1}{h_k}\left(\frac{1}{x_{k+1} - x_k} + \frac{1}{x_k + x_{k-1}}\right) + c(x_k), \\
A_{k,k-1} &= -\frac{1}{h_k(x_k - x_{k-1})} \\
A_{k,k-1} &= -\frac{1}{h_k(x_{k+1} - x_k)} \\
A_{k,j} &= 0 \qquad \text{sonst},
\end{aligned}$$

und die rechte Seite

$$\begin{aligned}
F_k &= f(x_k) \qquad k \in \{2, \ldots, N-2\} \\
F_1 &= f(x_1) + \frac{g_0}{h_0(h_0(x_1 - x_0)} \\
F_{N_1} &= f(x_{N_1} + \frac{g_1}{h_{N-1}(x_N - x_{N-1})}
\end{aligned}$$

so erhalten wir das $(N-1) \times (N-1)$ System

$$AU = F$$

für die diskrete Lösung. Im äquidistanten Fall gilt

$$A = \frac{1}{h^2} \begin{pmatrix} 2 + h^2 c(x_1) & -1 & 0 & \dots & 0 \\ -1 & 2 + h^2 c(x_2) & -1 & \dots & 0 \\ 0 & -1 & 2 + h^2 c(x_3) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 2 + h^2 c(x_{N-1}) \end{pmatrix}.$$

Wir erkennen einige Eigenschaften der Matrix, die auch allgemeiner bei der Lösung von Randwertproblemen gelten:

- A ist *dünnbesetzt* (im engl. *sparse*), d.h. die meisten Einträge von A sind gleich null. Hier haben wir von den $(N-1)^2$ Einträgen nur $3N-5$ Einträge, die nicht verschwinden. Dies hat einige Konsequenzen bei der effizienten Speicherung von A sowie bei der Lösung des Systems $AU = F$. Wir können die Matrix im Sparse-Format speichern, d.h. als eine Liste (i, j, A_{ij}) für alle Nichtnulleinträge. Statt $(n-1)^2$ reeller Zahlen speichern wir hier nur $6N-10$ ganze und $3N-5$ reelle Zahlen. Dies ist eine enorme Einsparung für großes N. Bei der numerischen Lösung des linearen Systems ist es auch vorteilhaft diese Struktur zu benutzen. Bei direkten Verfahren wie der LR-Zerlegung 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, dann integrieren wir von 1 bis x um $u(x)$ zu erhalten, dies entspricht einer rechten oberen Dreiecksmatrix. Beide Matrizen haben dann keine dünnbesetzte Struktur, da das diskrete Integral alle Gitterpunkte benutzt.

- Für $c \geq 0$ ist die Matrix A schwach diagonaldominant, d.h. $A_{kk} \geq \sum_{j \neq k} |A_{kj}$ (bzw. auch in den Spalten $A_{kk} \geq \sum_{j \neq k} |A_{jk}$). Für $k = 1$ und $k = N-1$ sind diese Ungleichungen sogar strikt. Wir haben weiter sogar nur positive Diagonaleinträge und nichtpositive Nebendiagonaleinträge. Man nennt so eine Matrix M-Matrix, wir werden später noch sehen, dass A dann invertierbar ist und die Inverse nur nichtnegative Einträge hat. Dies ist das diskrete Äquivalent zum Maximumsprinzip. Hat F nur nichtnegative Einträge, so gilt dann auch $U = A^{-1}F$.

- Die Matrix A ist symmetrisch, d.h. $A_{jk} = A_{kj}$ für alle k, j. Dies ist eine Konsequenz aus der Divergenzform, da der Divergenzform, da der Operator $L : u \mapsto -(au')' + cu$ auch formal selbstadjungiert ist. Es gilt mit partieller Integration für u und v mit

Nullrandwerten

$$\begin{aligned}\langle Lu, v\rangle_{L^2} &= \int_0^1 (Lu)(x)\ v(x)\ dx = \int_0^1 (-(a(x)u'(x))v(x) + c(x)u(x)v(x)\ dx \\ &= \int_0^1 a(x)u'(x)v'(x) + c(x)u(x)v(x)\ dx \\ &= \int_0^1 (-(a(x)v'(x))u(x) + c(x)v(x)u(x))\ dx = \langle v, Lu\rangle_{L^2}.\end{aligned}$$

Ist $c \geq 0$, dann ist A sogar symmetrisch positiv definit. Dies ist eine Konsequenz aus

$$\langle Lu, u\rangle = \int_0^1 a(x)|u'(x)|^2 + c(x)u(x)^2\ dx > 0$$

für $u \neq 0$.

Im allgemeinen Fall einer nichtkonstanten Funktion a können wir die selbe Einsicht über verschobene Gitter wie oben benutzen. Die erste Ableitung interpretieren wir diskret als Wert am Mittelpunkt der Gitterzelle $x_{k\pm 1/2}$, da wir diese mit a multiplizieren, verwenden wir auch den Wert von a an diesen Gitterpunkten. Dementsprechend approximieren wir

$$(au')'(x_k) \approx \frac{1}{h_k}\left(a(x_{k+1/2})\frac{u(x_{k+1}) - u(x_k)}{x_{k+1} - x_k} - a(x_{k-1/2})\frac{u(x_k) - u(x_{k-1})}{x_k - x_{k-1}}\right).$$

Auch in diesem Fall gelten alle Eigenschaften der Matrix A wie oben. Damit haben wir insbesondeere die Invertierbarkeit aus der positiven Definitheit. Also ist die Lösung $u(x_k) = U_k$ eindeutig definiert mit $U \in \mathbb{R}^{N-1}$ Lösung des Gleichungssystems $AU = F$.

3.1.1 Konvergenz von Differenzenverfahren

Um die Konvergenz(ordnung) zu vestehen, gehen wir wieder in der üblichen Weise vor. Sei $\tilde{u} \in C^2([0,1])$ die Lösung des Randwertproblems $\tilde{U} = (\tilde{u}(x_k)) \in \mathbb{R}^{N-1}$. Dann gilt

$$A(U - \tilde{U}) = F - A\tilde{U} = G,$$

und wir sehen, dass die rechte Seite G der Konsistenzfehler ist, d.h.

$$G(x_k) = (a\tilde{u}')'(x_k) - \frac{1}{h_k}\left(a(x_{k+1/2})\frac{u(x_{k+1}) - u(x_k)}{x_{k+1} - x_k} - a(x_{k-1/2})\frac{u(x_k) - u(x_{k-1})}{x_k - x_{k-1}}\right).$$

Wir beachten dabei, dass der Term cu exakt diskretisiert wird und nicht weiter zum Fehler beiträgt. Damit können wir zunächst wieder den Konsistenzfehler

$$\kappa_h = \|G\|_\infty = \max_k |G(x_k)| \tag{3.5}$$

abschätzen. Um Konvergenz zu erhalten benötigen wir eine Abschätzung an die Inverse von A, es gilt

$$E_h := \|U - \tilde{U}\|_\infty = \|A^{-1}(U - \tilde{U})\|_\infty \leq \|A^{-1}\|\ U - \tilde{U}\|_\infty.$$

Dabei müssen wir die Zeilensummennorm als verträgliche Norm zur Maximumsnorm verwenden. Wir sehen, dass aus Konsistenz(ordnung p) auch Konvergenz(ordnung p) folgt, wenn $\|A^{-1}(U-\tilde{U})\|_\infty$ beschränkt ist für $N \to \infty$ bzw. $h \to 0$. Dies bezeichnen wir folglich als Stabilität und halten dies entsprechend in einer Definition fort:

DEFINITION 3.4.

Ein Differenzenverfahren der obigen Form heißt konsistent von der Ordnung p, wenn $\kappa_h = \mathcal{O}(h^p)$ gilt. Das Verfahren heißt konvergent von der Ordnung p, wenn $E_h = \mathcal{O}(h^p)$ gilt.

Für die Stabilitätseigenschaft nutzen wir die Eigenschaften der Matrix A, die wir ebenfalls definieren:

DEFINITION 3.5.

Eine Matrix $B \in \mathbb{R}^{n\times n}$ heißt M-Matrix, wenn folgende Eigenschaften erfüllt sind:

- B hat nur positive Diagonaleinträge und nichtpositive Nebendiagonaleinträge.
- B ist schwach diagonaldominant, d.h. $B_{kk} \geq \sum_{j\neq k} |B_{jk}|$.
- Für mindestens ein $j \in \{1,\dots,n\}$ gilt $B_{kk} > \sum_{j\neq k} |B_{jk}|$.

Offensichtlich benötigen wir die letzte Bedingung für die Invertierbarkeit, da sonst auch

$$B = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

zulässig wäre. Wie wir leicht nachrechnen können erfüllt unsere Diskretisierung diese Eigenschaft:

LEMMA 3.6: S.

i A die Matrix aus dem obigen Differenzenverfahren für $c \geq 0$. Dannn ist A eine M-Matrix.

Nun können wir ein wichtiges Resultat über M-Matrizen beweisen:

LEMMA 3.7: S.

i B eine M-Matrix, dann ist B invertierbar und B^{-1} hat nur nicht-negative Einträge.

Beweis. Wir nehmen an, dass B irreduzibel ist, d.h. es existiert eine Permutation π von $\{1,\dots,n\}$, sodass $B_{\pi(i)\pi(i+1)} \neq 0$ gilt. Andernfalls können wir analog für alle irreduziblen Blöcke von B vorgehen. Zunächst zeigen wir, dass B invertierbar, d.h. der Nullraum

trivial ist. Sei $z \in \mathcal{N}(B)$ und z_m so, dass $z_m \leq z_k$ für alle k gilt. Nehmen wir nun an, dass $z_m < 0$ ist, dann folgt

$$B_{mm} z_m = -\sum_{j \neq m} B_{mj} z_j = \sum_{j \neq m} |B_{mj}| z_j \geq \sum_{j \neq m} |B_{mj}| z_m.$$

Ist $z_m < 0$, so folgt wegen der Diagonaldominanz $\sum_{j \neq m} |B_{mj}| z_m \geq B_{mm} z_m$. Dies ist aber nur möglich, wenn $z_j = z_m$ für alle j mit $B_{mj} \neq 0$ 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

$$B_{kk} > \sum_{j \neq k} |B_{jk}| = -\sum_{j \neq k} B_{jk},$$

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 B nur nicht-negative Einträge hat, können wir zeigen, dass die Lösung z von $Bz = y$, mit y gleich einem Einheitsvektor, nur nicht-negative Einträge hat. Dies ist natürlich äquivalent dazu, dass die Lösung von $Bz = y$ 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

$$B_{kk} z_k = -\sum_{j \neq k} B_{kj} z_j + y_k \geq -\sum_{j \neq k} B_{kj} z_k \geq B_{kk} z_k$$

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 $z_k < 0$. □

Die M-Matrix Eigenschaft ist eine wesentliche Voraussetzung für den Stabilitätsbeweis. Wie beim Maximumsprinzip für die Differentialgleichung werden wir Vergleichslösungen suchen, um Fehlerschranken zu erhalten. Dazu ist es wichtig, dass die Vergleichslösung unabhängig von h (bzw. N) ist, deshalb konstruieren wir diese für die Differentialgleichung. Im Folgenden sei $v \in C^2([0,1])$ die Lösung von

$$-(av')' + cv = f \qquad \text{in } (0,1)$$

mit homogenen Dirichlet-Randwerten $v(0) = v(1) = 0$. Dazu bezeichnen wir wieder mit $\tilde{u}$ die Lösung des Dirichlet-Problems mit rechter Seite f und Randwerten g_0 und g_1. Um die Abhängigkeit von h klar zu machen, bezeichnen wir die Matrix und rechte Seite der Diskretisierung miA_h bzw. F_h. Die Lösung von $A_h U = F_h$ bezeichnen wir mit $U^h = (u^h(x_k))_{k=1,\dots,N-1}$. Mit dem üblichen Konsistenzfehler G_h gilt

$$A_h(\tilde{U}^h - U^h) = G_h$$

wobei $\tilde{U}^h = (\tilde{u}^h(x_k))_{k=1,\dots,N-1}$ ist. Nun betrachten wir

$$\tilde{U}^h - U^h \pm 2\kappa_h V^h$$

mit dem globalen Konsistenfehler $\kappa_h = \|G_h\|_\infty$ und $V^h = (v^h(x_k))_{k=1,\dots,N-1}$. Die Idee dabei ist, dass

$$A_h(\tilde{U}^h - U^h + 2\kappa_h V^h) \geq 0 \tag{3.6}$$
$$A_h(\tilde{U}^h - U^h - 2\kappa_h V^h) \leq 0 \tag{3.7}$$

gelten sollte, zumindest für kleines h. Damit erhalten wir aus der Monotonie

$$-2\kappa_h V^h \leq \tilde{U}^h - U^h \leq 2\kappa_h V^h,$$

und da

$$\|V^h\|_\infty \leq \max_{x\in[0,1]} |v(x)|$$

gilt, erhalten wir daraus dann eine gleichmäßige Schranke für den Konvergenzfehler

$$E_h = \|\tilde{U}^h - U^h\|_\infty.$$

Etwas präziser machen wir die obige Eigenschaft in folgendem Lemma.

LEMMA 3.8.

Seien $U^h, \tilde{U}^h, V^h, A_h$ und G_h definiert wie oben, mit $\kappa_h = \|G_h\|_\infty$. Dann gelten für h hinreichend klein die Ungleichungen (3.6) und (3.7).

Beweis. Wegen der Konsistenz des Verfahrens wird für h hinreichend klein die Differenz

$$A_h V^h - (-(av')'(x_k) + c(x_k))_{k=1,\dots,N-1} = A_h V^h - (1)_{k=1,\dots,N-1}$$

beliebig klein, insbesondere gilt dann

$$A_h V^h \geq (\frac{1}{2})_{k=1,\dots,N-1}.$$

Setzen wir nun ein, so folgt

$$A_h(\tilde{U}^h - U^h + 2\kappa_h V^h) = G_h + 2\kappa_h A V^h \geq G_h + \kappa_h (1)_{k=1,\dots,N-1} \geq 0$$

und da $\kappa_h = \|G\|_\infty$ gilt. Analog folgt die zweite Ungleichung. □

Als direkte Folgerung erhalten wir ein Konvergenzresultat für Differenzenverfahren, denn die M-Matrix Eigenschaft impliziert, dass $\tilde{U}^h - U^h \pm 2\kappa_h V^h$ nur nichtnegative bzw. nichtpositive Einträge haben:

THEOREM 3.9.

Seien $U^h, \tilde{U}^h, V^h, A_h$ und G_h definiert wie oben, mit $\kappa_h = \|G_h\|_\infty$. Dann gilt für h hinreichend klein

$$E_h = \|U^h - \tilde{U}^h\|_\infty \leq 2\|v\|_\infty \kappa_h,$$

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

$$-\nabla \cdot (a\nabla u) + cu = f$$

auf einem rechteckigen Gebiet lösen wollen, mit $\nabla \cdot (a\nabla u) = \sum_{i=1}^{d} \partial_{x_i}(a\partial_{x_i} u)$. 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.

3.2 Finite Elemente Verfahren für Randwertprobleme

Finite Elemente Verfahren konstruiert man im Mehrdimensionalen basierend auf einer Triangulierung des Gebiets, die Idee ist dabei stückweise polynomiale Ansatzfunktionen zu verwenden (ähnlich den Splines bei der Interpolation). Damit kann man einfache Ansatzfunktionen auf dem Gebiet konstruieren und numerische Lösungen als Linearkombinationen davon berechnen.

In einer Dimension kann man wieder Ansatzfunktionen auf einem Gitter $0 = x_0 < x_1 < \ldots < x_N = 1$ konstruieren. Wir beschränken uns auf den Fall stückweise linearer Funktionen. Die finiten Elemente sind dabei Ansatzfunktionen mit lokalem Träger, nämlich

$$\phi_0(x) = \begin{cases} \frac{x_1 - x}{x_1 - x_0} & x < x_1 \\ 0 & \text{sonst,} \end{cases}$$

$$\phi_j(x) = \begin{cases} \frac{x - x_{j-1}}{x_j - x_{j-1}} & x_{j-1} < x < x_j \\ \frac{x_{j+1} - x}{x_{j+1} - x_j} & x_j < x < x_{j+1} \\ 0 & \text{sonst,} \end{cases}$$

$$\phi_N(x) = \begin{cases} \frac{x - x_{N-1}}{x_N - x_{N-1}} & x > x_{N-1} \\ 0 & \text{sonst,} \end{cases}$$

Dies ist eine Basis für die stückweise linearen Funktionen über diesem Gitter. Der Träger von ϕ_j und ϕ_k ist für $|j-k| > 1$ disjunkt, was noch wichtige Auswirkungen auf das später zu lösende lineare System haben wird.

Die diskrete Lösung können wir nun als Funktion in der Form

$$u^h(x) = \sum_{j=0}^{N} U_j \phi_j(x)$$

schreiben, die unbekannten Koeffizienten U_j erfüllen dann $U_j = u^h(x_j)$. Damit können wir aus den Dirichlet Randwerten direkt $U_0 = g_0$ und $U_N = g_1$ bestimmen, es bleibt also noch Gleichungen für U_j, $j = 1, \ldots, N-1$ herzuleiten. Natürlich können wir die gesamte Differentialgleichung nicht als Grundlage nehmen, nicht mal bei den Stützstellen x_j, da die Funktion u^h dort nicht einmal einfach differenzierbar ist. Stattdessen verwendet man die Idee der Galerkin-Diskretisierung, man projiziert die Gleichung und rechte Seite einfach in den Raum, der durch die ϕ_j aufgespannt wird. D.h. wir fordern statt einer linearen Gleichung $Lu = f$ nur

$$\langle Lu, \phi \rangle = \langle f, \phi \rangle$$

für alle $\phi \in \mathcal{U}_h = \text{lin}\{\phi_j\}_{j=1,\ldots,N-1}$. Dies liefert ein äquivalentes $(N-1) \times (N-1)$ System

$$\sum_j U_j \langle L\phi_j, \phi_i \rangle = \langle f - g_0 U_0 - g_1 U_N, \phi_i \rangle \qquad i = 1, \ldots, N-1.$$

Wir schreiben im Folgenden einfach f statt $f - g_0 U_0 - g_1 U_N$, da das Dirichlet Problem mit Randdaten g_0, g_1 offensichtlich die selbe Lösung liefert wie das homogene Dirichlet Problem mit geänderter rechter Seite f. Wie können wir in unserem Fall von stückweise linearen Funktionen aber $\langle L\phi_j, \phi_i \rangle$ definieren, da offensichtlich $L\phi_j$ nicht wohldefiniert ist (die erste Ableitung von ϕ_j existiert fast überall, ist aber in x_j unstetig, deshalb macht die zweite Ableitungen keinen Sinn in dieser einfachen Form.) Die Antwort liefert die sogenannte schwache Formulierung des Randwertproblems. Diese erhalten wir folgendermaßen: es gilt für Funktionen mit homogenen Randwerten

$$\langle Lu, \phi \rangle = \int_0^1 (Lu)(x)\phi(x)\, dx = \int_0^1 (-(au')' + cu)\phi \, dx = \int_0^1 au'\phi' + cu\phi \, dx =: B(u, \phi).$$

B ist eine symmetrische Bilinearform, die wir auch für stückweise stetig differenzierbare Funktionen äquivalent als

$$B(u, \phi) = \sum_{j=0}^{N-1} \int_{x_j}^{x_{j+1}} au'\phi' + cu\phi \, dx$$

schreiben können. Damit haben wir automatische eine geeignete Diskretisierung, die Lösung u_h lässt sich aus

$$B(u, \phi) = \langle f, \phi \rangle$$

für alle $\phi \in \mathcal{U}_h$ schreiben. Dies ist äquivalent zum linearen Gleichungssystem

$$AU = F$$

mit $F_i = \langle f, \phi_i \rangle$ und $A_{ij} = B(\phi_i, \phi_j)$.

Die Matrix A ist offensichtlich symmetrisch, sie ist sogar positiv definit, da für alle $z \in \mathbb{R}^{N-1}$ gilt

$$z^T A z = \sum_{ij} z_i z_j B(\phi_i, \phi_j) = B(\sum_i z_i \phi_i, \sum_j z_j \phi_j) > 0.$$

Dazu ist die Matrix A wegen des lokalen Trägers der ϕ_i wieder dünnbesetzt, sogar tridiagonal, da wir sofort sehen, dass

$$A_{ij} = B(\phi_i, \phi_j) = 0 \qquad \text{für } |i-j| > 1$$

gilt.

BEISPIEL 3.10.

Als Beispiel berechnen wir A im Fall $a = 1$, c konstant und $x_j = jh$. Dann gilt wegen $\phi_i' = \frac{1}{h}$ in (x_{i-1}, x_i) und $\phi_i' = -\frac{1}{h}$ in (x_i, x_{i+1})

$$A_{ii} = \int_{x_i-h}^{x_i+h} \frac{1}{h^2}\, dx + c\int_{x_i-h}^{x_i} \left(\frac{x - x_i + h}{h}\right)^2 dx + c\int_{x_i}^{x_i+h} \left(\frac{x_i + h - x}{h}\right)^2 dx = \frac{2}{h} + \frac{2c}{3}h.$$

Genauso erhalten wir

$$A_{i,i+1} = A_{i-1i} = -\frac{1}{h} + c\frac{1}{6}h.$$

Für $h < \sqrt{\frac{c}{6}}$ ist A wieder eine M-Matrix und wir können die Theorie aus dem letzten Abschnitt anwenden. Allgemeiner können wir aber eine Theorie basierend auf der Bilinearform B konstruieren.

Da B symmetrisch, positiv definit und bilinear ist, definiert B auch ein Skalarprodukt und

$$\|u\|_B = \sqrt{B(u,u)}$$

ist eine Norm. Es gilt auf dem Intervall $(0,1)$ für Funktionen mit $u(0) = 0$ sogar

$$\|u\|_\infty \le \sqrt{\int_0^1 (u')^2\, dx} \le C\sqrt{B(u,u)}$$

mit einer Konstante $c > 0$. Ist $\tilde{u}$ eine Lösung des Dirichlet-Problems, dann gilt insbesondere

$$B(\tilde{u}, \phi) = \int_0^1 f\phi\, dx$$

für alle $\phi \in \mathcal{U}_h$. Damit folgt die sogenannte Galerkin-Orthogonalität

$$B(\tilde{u} - u_h, \phi) = 0 \qquad \forall \phi \in \mathcal{U}_h.$$

Sei nun $\tilde{u}_h$ eine gegebene Approximation von $\tilde{u}$ in $\mathcal{U}_h$, z.B. die Interpolierende an den Gitterpunkten, dann gilt mit $\phi = \tilde{u}_h - u_h$ auch

$$0 = B(\tilde{u} - u_h, \tilde{u}_h - u_h) = B(\tilde{u} - u_h, \tilde{u} - u_h) - B(\tilde{u} - u_h, \tilde{u} - \tilde{u}_h).$$

Da eine Bilinearform immer die Cauchy-Schwarz Ungleichung erfüllt, folgt weiter

$$\|\tilde{u} - u_h\|_B^2 = B(\tilde{u} - u_h, \tilde{u} - u_h) = B(\tilde{u} - u_h, \tilde{u} - \tilde{u}_h) \leq \|\tilde{u} - u_h\|_B \|\tilde{u} - \tilde{u}_h\|_B$$

und damit

$$\|\tilde{u} - u_h\|_B \leq \|\tilde{u} - \tilde{u}_h\|_B.$$

Die rechte Seite können wir durch den Interpolationsfehler abschätzen, es gilt

$$\|\tilde{u} - \tilde{u}_h\|_B^2 \leq \max\{\|a\|_\infty, \|c\|_\infty\} \int_0^1 (\tilde{u}' - \tilde{u}_h')^2 + (\tilde{u} - \tilde{u}_h)^2 \, dx$$

Durch eine einfache Taylorentwicklung und Ausnutzung von $\tilde{u} \in C^2$ erhalten wir dann

$$\int_0^1 (\tilde{u}' - \tilde{u}_h')^2 + (\tilde{u} - \tilde{u}_h)^2 \, dx \leq C_I h^2$$

für eine Konstante C_I und daraus

$$\|\tilde{u} - u_h\|_B \leq \sqrt{\max\{\|a\|_\infty, \|c\|_\infty\} C_I} h.$$

Damit haben wir also direkt eine Abschätzung des Konvergenzfehlers hergeleitet. Wir beachten, dass die Konsistenz hier dem letzten Schritt (Taylor-Entwicklung der Interpolierenden) entspricht und wir die Stabilität aus der Galerkin-Orthogonalität erhalten, wir haben also auch wieder die übliche Kombination für ein konvergentes Diskretisierungsverfahren.

Kapitel 4

Unrestringierte 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, seien es Probleme des Alltags oder grundlegende Gesetzmäßigkeiten bei der Ergründung des Kosmos. Im folgenden Abschnitt wollen wir uns speziell mit der unbeschränkten (oder: unrestringierten) Optimierung beschäftigen und uns nützliche Werkzeuge zum Lösen von numerischen Problemstellungen herleiten. Nach einer allgemeinen mathematischen Einführung in Kapitel 4.1 beginnen wir in Kapitel 4.2 mit einer Klasse von Algorithmen, die einer einfachen Idee folgen: die Abstiegsverfahren. In Kapitel 2.2 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 Kapitel 2.3 einen modernen primal-dualen Optimierungsalgorithmus zur Lösung von konvexen, nicht-glatten Problemen.

4.1 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 [**nocedal_1999**]. Wir beginnen mit der Definition des allgemeinen Optimierungsproblems, welches wir im Verlauf der Vorlesung nur für Spezialfälle näher untersuchen werden.

DEFINITION 4.1: 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.

$$\min_{x\in\Omega} F(x) \qquad \text{mit} \begin{cases} c_i(x) = 0, & i \in \mathcal{E}, \\ c_i(x) \geq 0, & i \in \mathcal{I}. \end{cases} \tag{4.1}$$

Die reellwertigen Funktionen $c_i\colon \Omega \to \mathbb{R}$ 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 4.2.

Betrachte das 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 (4.1) umschreiben als:

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

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 sieht man, dass das globale Minimum der quadratischen Funktion 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, nämlich $x = (x_1, x_2)^T = (x, y)^T$.

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

DEFINITION 4.3: Unbeschränkte Optimierung.

Liegt ein allgemeines Optimierungsproblem der Form (4.1) ohne Nebenbedingungen vor, d.h., für die Indexmengen gilt $\mathcal{E} = \mathcal{I} = \emptyset$, so sprechen wir von einem *unbeschränkten* oder *unrestringierten Optimierungsproblem*.

BEMERKUNG 4.4. 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*. △

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, Konvexität, Differenzierbarkeit
- **Charakteristik des Optimums**, z.B. Sattelpunkt, lokales oder globales Optimum
- **Modelleigenschaften**, z.B. stochastisch oder deterministisch

Für den weiteren Verlauf der Vorlesung gehen wir immer, wenn nicht anders beschrieben, von einem unrestringierten Optimierungsproblem aus. Da wir uns intensiv mit der Bestimmung und numerischen Approximation von Optima beschäftigen werden, macht es Sinn diese zuerst formal zu beschreiben.

DEFINITION 4.5: Stationärer Punkt.

Sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Funktion. Wir nennen einen Punkt $x^* \in \Omega$ *stationären Punkt* von F, falls er die Bedingung $\nabla F(x^*) = 0$ erfüllt.

DEFINITION 4.6: Lokales und globales Minimum.

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

$$F(x^*) \leq F(x), \quad \forall x \in U. \tag{4.2}$$

Wir nennen $x^* \in \Omega$ ein *globales Minimum* von F, falls die Ungleichung (4.2) für jede beliebige Umgebung $U \subset \Omega$ gilt und somit insbesondere für $U = \Omega$.

BEMERKUNG 4.7. In obiger Definition sprechen wir nur von Minima, jedoch ist klar, dass sich jedes Maximierungsproblem durch einen Vorzeichenwechsel leicht in ein Minimierungsproblem umschreiben lässt, d.h.,

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

△

Da wir nun eine Bedingung für das Vorliegen eines lokalen Minimums haben, können wir mit folgenden Satz die notwendigen Bedingungen für solch ein lokales Minimum angeben.

THEOREM 4.8: Notwendige Optimalitätsbedingungen erster Ordnung.

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

Beweis. Wir führen einen Beweis durch Widerspruch. Nehmen wir also an, dass $x^* \in \mathbb{R}$ ein lokales Minimum von F sei, jedoch aber $\nabla F(x^*) \neq 0$ gelte. Wir wählen den Richtungsvektor $\mathbf{p} := -\nabla F(x^*) \neq 0$. 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 $t \in [0, T]$:

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

Somit gilt also $F(x^* + t\mathbf{p}) < F(x^*)$ 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 lokales Minimum von F. Das ist aber ein Widerspruch zur Annahme und somit ist die Behauptung bewiesen. □

KOROLLAR 4.9.

Jedes lokale Minimum $x^* \in \Omega$ einer Funktion $F\colon \Omega \to \mathbb{R}$ ist ein stationärer Punkt.

BEMERKUNG 4.10. Die Umkehrung der Aussage in Satz 4.1 gilt im Allgemeinen nicht. Man betrachte zum Beispiel die Funktion $F(x) := -x^3$. Diese besitzt einen stationären Punkt in $x^* = 0$, d.h., es gilt $\nabla F(0) = 0$. Dennoch handelt es sich hierbei nicht um ein lokales Optimum, sondern lediglich um einen Sattelpunkt. Aus diesem Grund handelt es sich nur um notwendige Bedingungen. △

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

THEOREM 4.11: Notwendige Optimalitätsbedingungen zweiter Ordnung.

Sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Funktion. Sei $x^* \in \Omega$ ein lokales Minimum von F in Ω und die Hessematrix $\nabla^2 F$ von F sei stetig in einer 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 $\langle \mathbf{p}, \nabla^2 F(x^*)\mathbf{p}\rangle \geq 0$ für alle $\mathbf{p} \in \mathbb{R}^n$.

Beweis: Der erste Teil der Behauptung folgt bereits aus Satz 4.1, so dass wir uns nur auf den Beweis für die zweite Behauptung konzentrieren müssen. Wir führen wieder einen Beweis durch Widerspruch. Sei $x^* \in \Omega$ nach Vorraussetzung ein lokaler Minimierer von F, das heißt nach Satz 4.1 gilt $\nabla F(x^*) = 0$. Wir nehmen an, dass $\nabla^2 F(x^*)$ nicht positiv semidefinit ist. Dann können wir einen Vektor $\mathbf{p} \in \mathbb{R}^n/\{0\}$ finden, so dass

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

gilt. Da $\nabla^2 F$ nach Vorraussetzung stetig ist in einer lokalen Umgebung $U \subset \Omega$ von x^* existiert ein $T > 0$, so dass

$$\langle \mathbf{p}, \nabla^2 F(x^* + t\mathbf{p})\mathbf{p}\rangle \;=\; 0, \qquad \text{für alle } t \in [0, T].$$

Nach dem Satz von Taylor gilt jedoch für alle $t \in (0, T)$

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

Damit folgt also, dass $F(x^*+t\mathbf{p}) < F(x^*)$ gilt. Wir haben also eine Richtung $\mathbf{p} \in \mathbb{R}^n/\{0\}$ gefunden entlang der die Funktionswerte von F abnehmen. Damit folgt, dass x^* kein lokales Minimum von F ist, was aber im Widerspruch zur Annahme steht. Das beweist die Aussage des Satzes. □

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

THEOREM 4.12: Hinreichende Optimalitätsbedingungen zweiter Ordnung.

Sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Funktion. Sei $x^* \in \Omega$ ein lokales Minimum von F in Ω und 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 $x^* \in \Omega$ 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 Vorraussetzung 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 $x^* \in \Omega$ um ein striktes lokales Minimum der Funktion F. □

BEMERKUNG 4.13. Die in Satz **??** genannten Bedingungen sind nur hinreichend, jedoch nicht notwendig für das Vorliegen eines strikten lokalen Minimums. Dies sieht man ein, wenn man beispielsweise die Funktion $F(x) := x^4$ 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 4.14: Konvexität.

Sei $\Omega \subset \mathbb{R}^n$ ein offenes, zusammenhängendes Gebiet und sei $F\colon \Omega \to \mathbb{R}$ eine reellwertige Funktion. 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 $x, y \in \Omega$ oberhalb des Graphen der Funktion F durch die Punkte x und y verläuft.

4.2 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 [**numerik1**] 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 Funktion, die es in dem Gebiet zu minimieren gilt. Die allgemeine Idee der Abstiegsverfahren ist es nun ausgehend von einem aktuellen Punkt $x_k \in \Omega$ einen Schritt in eine Richtung $\mathbf{p}_k \in \mathbb{R}^n$ zu machen, so dass der Funktionswert von F in dem neuen Punkt x_{k+1} abnimmt. Das bedeutet wir versuchen ein Iterationsschema der Form

$$x_{k+1} \;=\; x_k + \alpha_k \mathbf{p}_k, \quad \alpha_k > 0, \tag{4.3}$$

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 $\mathbf{p}_k \in \mathbb{R}^n$ wird hierbei auch in jedem Schritt des Abstiegsverfahrens gewählt.

4.2.1 Gradientenabstiegsverfahren

Zuerst beschäftigen wir uns mit dem wohl bekanntesten Abstiegsverfahren, dem *Gradientenabstiegsverfahren* (im Englischen: *gradient descent*). 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. 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 $\mathbf{p}$ mit Schrittweitenlänge $\alpha > 0$ für den k-ten Schritt des Abstiegsverfahren näher anschauen.

$$F(x_k + \alpha \mathbf{p}) \;=\; F(x_k) + \alpha \langle \nabla F(x_k), \mathbf{p} \rangle + \mathcal{O}(\mathcal{H}[F(x_k)]),$$

wobei $\mathcal{H}[F]$ die Hesse-Matrix der Funktion F bezeichnet. Es wird also klar, dass die Änderung der Funktionswerte von F im Wesentlichen von der Größe des Terms $\langle \nabla F(x_k), \mathbf{p} \rangle$ bestimmt wird. Wir können also für eine maximale Verringerung der Funktion F nach derjenigen Richtung mit Einheitslänge suchen, die das folgende Minimierungsproblem löst:

$$\min_{\mathbf{p} \in \mathbb{R}^n} \langle \nabla F(x_k), \mathbf{p} \rangle, \quad \text{mit } ||\mathbf{p}|| = 1. \tag{4.4}$$

Da außerdem

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

gilt, wobei θ der Winkel zwischen den Vektoren $\nabla F(x_k)$ und $\mathbf{p}$ bildet, können wir das Problem (4.4) umschreiben zu

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

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

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

Daraus folgt aber auch schon, dass

$$\mathbf{p} \;=\; -\frac{\nabla F(x_k)}{||\nabla F(x_k)||}. \tag{4.5}$$

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

Da wir daran interessiert sind die Funktion F schnellstmöglich zu minimieren, macht es Sinn in 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 $x_1, \ldots$ bestimmen, deren Funktionswerte monoton fallen sollten:

$$x_{k+1} \;=\; x_k - \nabla F(x_k). \tag{4.6}$$

Intuitiv stoppt man das Iterationsschema (4.6) sobald die Folge der Funktionswerte $F(x_k)$ nicht mehr kleiner wird. Man sieht leicht ein, dass das simple Iterationsschema (4.6) ein Spezialfall des allgemeinen Abstiegsverfahrens (4.3) mit fester Schrittweite $\alpha_k = ||\nabla F(x_k)||$ und Richtung $\mathbf{p}_k = \nabla F(x_k)$ ist. Diese einfache Methode lässt sich durch folgenden Algorithmus implementieren.

ALGORITHMUS 4.15: 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 4.2.1 in dieser Form praktisch nicht anwendbar. Warum dies so ist sieht man leicht an folgenden Beispiel.

BEISPIEL 4.16.

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 4.2.1 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 4.2.1 zeigt, besteht bei dem Gradientenabstiegsverfahren in Algorithmus 4.2.1 die Gefahr das Minimum zu überspringen. Aus diesem Grund kommt man auf die Idee die Schrittweitengröße $\alpha_k > 0$ in (4.3) 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 (4.6) durch:

$$x_{k+1} = x_k - \tau \frac{\nabla F(x_k)}{||\nabla F(x_k)||}, \quad \tau > 0. \tag{4.7}$$

Das Iterationsschema (4.8) ist wiederum ein Spezialfall des allgemeinen Abstiegsverfahrens in (4.3) für eine feste Schritteweite $\alpha_k = \tau > 0$ und Richtung $\mathbf{p}_k = -\nabla F(x_k)/||\nabla F(x_k)||$. Es ist klar, dass durch eine kleinere Schrittweite $\tau > 0$ das lokale Minimum der Funktion F im Punkt $x^* \in \Omega$ 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 wie folgt angeben:

$$\begin{aligned} \alpha_{k+1} &= \begin{cases} \alpha_k, & \text{falls } F\left(x_k - \alpha_k \frac{\nabla F(x_k)}{||\nabla F(x_k)||}\right) < F(x_k), \\ \sigma\alpha_k, & \text{sonst für } 0 < \sigma < 1 \text{ fix.} \end{cases}, \\ x_{k+1} &= x_k - \alpha_{k+1} \frac{\nabla F(x_k)}{||\nabla F(x_k)||}. \end{aligned} \tag{4.8}$$

Das adaptive Gradientenabstiegsverfahren in (4.8) lässt sich mit folgendem Algorithmus umsetzen.

ALGORITHMUS 4.17: 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
$x^* = x_k$
$F(x^*) = F(x_k)$
}

In Abbildung 4.1 ist ein typischer Verlauf des adaptiven Gradientenverfahrens in Algorithmus 4.2.1 bei der Minimierung einer Funktion $F\colon \mathbb{R}^2 \to \mathbb{R}$ 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 4.18. Das in diesem Abschnitt beschriebene Gradientenabstiegsverfahren mit adaptiver Schrittweite $\tau_k > 0$ 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 Abbildung 4.1), welcher in den meisten Fällen offensichtlich suboptimal ist. Aus diesen

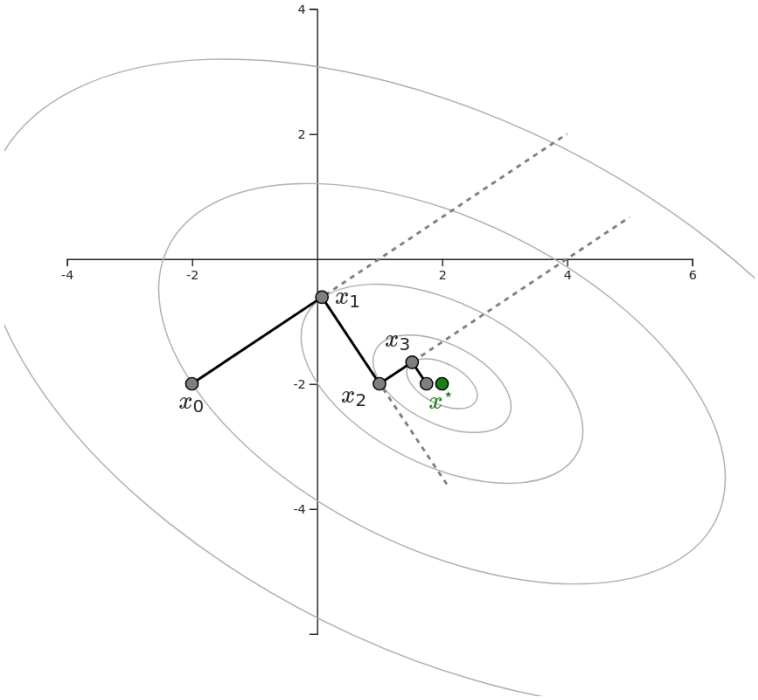

Abbildung 4.1: Approximation des Minimierers einer Funktion F in zwei Variablen mit Hilfe des adaptiven Gradientenverfahrens (4.8).

Gründen wollen wir uns in den nächsten Abschnitten mit alternativen Minimierungsmethoden beschäftigen. △

4.2.2 Koordinatenabstiegsverfahren

Eine weitere Variante des in Kapitel 4.2.1 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 $\mathbf{e}_j \in \mathbb{R}^n$ lässt sich das Koordinatenabstiegsverfahren schreiben als:

$$x_{k+1} \;=\; x_k - \alpha_k \langle \nabla F(x_k), \mathbf{e_j} \rangle \mathbf{e_j} \;=\; x_k - \frac{\partial F}{\partial x^i}(x_k)\mathbf{e_j}. \tag{4.9}$$

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 (4.9) kann hierbei ähnlich wie

in Kapitel 4.2.1 fest oder adaptiv gewählt werden. Das Koordinatenabstiegsverfahren in (4.9) mit adaptiver Schrittweite lässt sich mit folgendem Algorithmus umsetzen.

ALGORITHMUS 4.19: Koordinatenabstiegsverfahren.

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

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**
  # Wähle zufällige Koordinatenrichtung
  $i =$ `randomDraw`$([1:n])$
  # Berechne Richtungsableitung
  $\mathbf{p}_k = \frac{\partial}{\partial x^i} F(x_k)$
  while $F(x_k - \alpha_k \mathbf{p}_k / ||\mathbf{p}_k||) > F(x_k)$ **do**
    # Verringere Schrittweite um Faktor σ
    $\alpha_k = \sigma \alpha_k$
  end while
  # Update in Richtung des größten Gradientenabstiegs
  $x_{k+1} = x_k - \alpha_k \mathbf{p}_k / ||\mathbf{p}_k||$
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, da jedes eindimensionale Optimierungsproblem wesentlich leichter zu lösen ist als die Berechnung des gesamten Gradienten in jedem Schritt.

BEMERKUNG 4.20. 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 *Blockkoordinatenabtiegsverfahren* (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:

$$x_{k+1} \;=\; x_k - \alpha_k \langle \nabla F(x_k), \sum_{i=1}^{s} \mathbf{e}_{\sigma(i)} \rangle \sum_{i=1}^{s} \mathbf{e}_{\sigma(i)}. \tag{4.10}$$

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 (4.9) und das normale Gradientenabstiegsverfahren in (4.8) Spezialfälle des Blockkoordinatenabstiegsverfahrens in (4.10) für Blockgrößen $s = 1$ und $s = n$ sind. △

4.2.3 Stochastisches Gradientenabstiegsverfahren

Eine aktuell weit verbreitete Variante des Gradientenabstiegsverfahrens in (4.8) 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 schlechte, 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:

$$F(x) \;=\; \sum_{i=1}^{m} F_i(x), \qquad \text{für alle } x \in \Omega. \tag{4.11}$$

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 (4.8) umschreiben zu:

$$x_{k+1} \;=\; x_k - \alpha_k \frac{\nabla F(x_k)}{||\nabla F(x_k)||} \;=\; x_k - \alpha_k \frac{\sum_{i=1}^{m} \nabla F_i(x_k)}{||\sum_{i=1}^{m} \nabla F_i(x_k)||}, \quad \alpha_k > 0. \tag{4.12}$$

Die Idee des stochastischen Gradientenverfahrens ist es nun einen zufälligen Summanden aus (4.11) zu wählen und nur den Gradienten bezüglich dieses Summanden zu

betrachten. Durch diese starke Vereinfachung von (4.12) führt man mit einem zufällig ausgewählten Index $j \in \{1, \dots, m\}$ nun einen Gradientenabstieg der Form

$$x_{k+1} \;=\; x_k - \alpha_k \frac{\nabla F_j(x_k)}{||\nabla F_j(x_k)||}, \quad \alpha_k > 0 \tag{4.13}$$

durch.

BEMERKUNG 4.21. Ähnlich wie im Fall des Koordinatenabstiegsverfahrens in Kapitel 4.2.2, gibt es auch beim stochastischen Gradientenverfahren die Möglichkeit einen Kompromiss zwischen dem normalen Gradientenabstieg in (4.7) und dem auf einen Summanden beschränkten Gradientenabstieg (4.13) 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 $1, \dots, m$. △

4.2.4 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 Kapitel 5.1 der Vorlesung „Einführung in die Numerik“haben wir das Newton Verfahren zur Approximation von Nullstellen nichtlinearer Gleichungssysteme hergeleitet. Aus der Taylorentwicklung einer nichtlinearen Nullstellengleichung $F(x^*) = 0$ von der Form

$$0 \;=\; F(x^*) \;=\; F(x) + (x^* - x)F'(x) + \mathcal{O}(F'').$$

haben wir die folgende Fixpunktfunktion als Approximation erster Ordnung angegeben:

$$G(x) \;=\; x - (F'(x))^{-1} F(x), \qquad \text{für } F'(x) \neq 0. \tag{4.14}$$

Hierbei haben wir die Fixpunktgleichung als erfüllt gesehen, wenn wir ein $x^* \in \Omega$ gefunden haben, so dass die linke und rechte Seite von (4.14) übereinstimmen. Unter dieser Beobachtung haben wir das *Newton-Verfahren* als iteratives Schema zur Bestimmung eines solchen Fixpunktes $x^* \in \Omega$ hergeleitet:

$$x_{k+1} \;=\; x_k - \frac{F(x_k)}{F'(x_k)}. \tag{4.15}$$

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 Bedingungen für die lokale Konvergenz des Newton-Verfahrens.

THEOREM 4.22.

Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ in einer Umgebung von $\overline{x}$ stetig differenzierbar, F' lokal Lipschitz stetig , $F(\overline{x}) = 0$ und $F'(\overline{x})$ regulär. Dann existiert eine Umgebung $B_R(\overline{x})$, sodass das Newton-Verfahren für jeden Startwert $x_0 \in B_R(\overline{x})$ konvergiert, d.h. $x_k \to \overline{x}$.

Beweis. Siehe [**numerik1**]. □

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 zu approximieren und damit die notwendigen Optimalitätsbedingungen 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 $x_k + \mathbf{p} \in \Omega$ des allgemeinen Iterationsschemas (4.3) aber berücksichtigen diesmal auch Terme von zweiter Ordnung:

$$F(x_k + \mathbf{p}) \;\approx\; F(x_k) + \langle \mathbf{p}, \nabla F(x_k)\rangle + \frac{1}{2}\langle \mathbf{p}, \nabla^2 F(x_k)\mathbf{p}\rangle \;=:\; m_k(\mathbf{p}). \tag{4.16}$$

Wenn wir für den Moment davon ausgehen, dass die Hessematrix $\nabla^2 F(x_k)$ positiv definit ist, so können wir ein eindeutiges Minimum der Funktion $m_k(\mathbf{p})$ wie folgt bestimmen:

$$\mathbf{p} \;=\; -(\nabla^2 F(x_k))^{-1}\nabla F(x_k). \tag{4.17}$$

Die Richtung $\mathbf{p} \in \mathbb{R}^n/\{0\}$ in (4.17) wird auch *Newton-Richtung* genannt. Mit ihr 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:

$$x_{k+1} \;=\; x_k + \mathbf{p} \;=\; x_k - (\nabla^2 F(x_k))^{-1}\nabla F(x_k). \tag{4.18}$$

Damit das Newton-Abstiegsverfahren in (4.18) überhaupt sinnvoll ist, müssen wir fordern, dass die Hessematrix in jedem Punkt $x_k \in \Omega$ der Iterationsfolge 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:

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

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 [**nocedal_1999**]. Auf diese werden wir jedoch im weiteren Verlauf der Vorlesung nicht näher eingehen.

BEMERKUNG 4.23. Das Newton-Abstiegsverfahren in (4.18) ist ein Abstiegsverfahren der Art (4.3) dessen Schrittweite α_k 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^*. △

4.2.5 Quasi-Newton Verfahren

Im Kapitel 4.2.4 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 Funktion F verwenden. Dennoch ist das Newton Verfahren aus numerischer Sicht noch nicht ideal, da 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 (4.17) zuerst die Hessematrix bestimmt und anschließend invertiert werden. Dies ist aus Effizienzgründen ungewünscht, da die Inversion einer $n \times n$-Matrix in $\mathcal{O}(n^3)$ Rechenoperationen liegt. Da die Bestimmung und die Inversion der Hessematrix in jedem Iterationsschritt passieren müssen, ist das Newton Verfahren nicht 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

$$B_k \approx \nabla^2 F(x_k). \tag{4.19}$$

Damit können wir die Modellfunktion $m_k(\mathbf{p})$ in (4.16) schreiben als:

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

das heißt, wir approximieren die Zielfunktion F im k-ten Iterationsschritt entlang der Richtung $\mathbf{p} \in \mathbb{R}^n$ 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 $\mathbf{p} = 0$ die Approximation exakt ist, da

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

Wenn wir fordern, dass B_k in (4.19) eine positiv definite Matrix ist, so lässt sich ein Abstiegsschritt des Iterationsverfahrens (4.3) analog zur Herleitung des Newton Abstiegsverfahrens in Kapitel 4.2.4 angeben als:

$$x_{k+1} \;=\; x_k + \alpha_k \mathbf{p}_k, \qquad \mathbf{p}_k \;=\; -B_k^{-1} \nabla F(x_k). \tag{4.20}$$

Die sogenannten *Quasi-Newton Verfahren* verfolgen diesen Ansatz. 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 Kapitel 4.2.4.

Die entscheidende Frage bei der Konstruktion eines Quasi-Newton Abstiegsverfahrens der Form (4.20) 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 in jedem Schritt des Iterationsverfahrens zu aktualisieren. 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 $x_{k+1} = x_k + \alpha_k \mathbf{p}$ erhalten. Unsere quadratische Approximation in diesem neuen Punkt für eine neue Richtung $\mathbf{p} \in \mathbb{R}^n$ sieht dementsprechend wie folgt aus:

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

Eine Forderung, die man nun die Modellfunktion m_{k+1} stellen kann ist, dass ihre Ableitung mit der Ableitung der Zielfunktion F in den letzten beiden Punkten x_k und x_{k+1} übereinstimmt. Da

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

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

$$\nabla m_{k+1}(-\alpha_k \mathbf{p}_k) \;=\; \nabla F(x_{k+1}) - \alpha_k B_{k+1} \mathbf{p}_k \;\overset{!}{=}\; \nabla F(x_k). \tag{4.21}$$

Durch Umstellen von (4.21) erhalten wir die Bedingung

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

Eine vernünftige Wahl der Matrix B_{k+1} in (4.19) 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:

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

Für unser allgemeines Quasi-Newton Verfahren in (4.20) 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} aktualisieren, so dass der folgende Zusammenhang für den nächsten Punkt $x_{k+1} \in \Omega$ erfüllt wird:

$$B_{k+1} s_k = y_k, \tag{4.22}$$

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. Um die positive Definitheit der Matrix B_{k+1} in Schrittrichtung $\alpha \mathbf{p}_k$ zu gewährleisten müssen wir fordern, dass die Vektoren y_k und s_k die sogenannte **Krümmungsbedingung** erfüllen:

$$\langle s_k, y_k \rangle > 0. \tag{4.23}$$

Dies ist eine hinreichende Bedingung für die positive Definitheit von B_{k+1} bezüglich der Richtung $\alpha_k \mathbf{p}_k$, da wir einfach die Sekantengleichung (4.22) von links mit dem Vektor s_k^T multiplizieren können und so erhalten wir mit der Forderung (4.23) schon:

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

BEMERKUNG 4.24. Falls die Funktion F strikt konvex ist, so ist die Krümmungsbedingung (4.23) 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 (4.23) erfüllt ist, so existiert mindestens eine Lösung B_{k+1} der Sekantengleichung (4.22). 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 (4.22) 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:

$$\begin{aligned} &\min_B ||B - B_k||, \quad \text{unter den Nebenbedingungen:} \\ B = B^T, \qquad &Bs_k = y_k, \qquad \langle \mathbf{p}, B\mathbf{p} \rangle > 0, \forall \mathbf{p} \in \mathbb{R}^n / \{0\}, \end{aligned} \tag{4.24}$$

wobei s_k und y_k definiert sind wie in der Sekantengleichung (4.22). Man beachte, dass man eine unterschiedliche Lösung B_{k+1} des Optimierungsproblems (4.24) 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 [**davidon_1959**], das im Übrigen bei der Erstbegutachtung abgelehnt wurde, wählt man für die Norm im Optimierungsproblem (4.24) 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 (4.24) verhält. Ein konkretes Beispiel 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

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

Mit der konkreten Wahl dieser Gewichtungsmatrix $W = G_k^{-1}$ wird die gewichtete Frobeniusnorm dimensionslos und man erhält eine eindeutige Lösung des Optimierungsproblems (4.24) wie folgt:

$$B_{k+1} \;=\; (I - \gamma_k y_k s_k^T) B_k (I - \gamma_k s_k y_k^T) + \gamma_k y_k y_k^T, \quad \text{mit } \gamma_k \;:=\; \frac{1}{\langle y_k, s_k \rangle}. \tag{4.25}$$

Die Gleichung (4.25) 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 (4.25) noch zu hoch. Das liegt daran, dass wir für einen Schritt des Quasi-Newton Verfahrens in (4.20) 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 auf Gleichung (4.25) anwenden um die neue Inverse H_{k+1} durch eine Aktualisierung der Matrix H_k zu berechnen:

$$H_{k+1} \;=\; H_k - \frac{H_k y_k y_k^T H_k}{\langle y_k, H_k y_k \rangle} + \frac{s_k s_k^T}{\langle y_k, s_k \rangle}. \tag{4.26}$$

Wie man einssieht liegt der numerische Rechenaufwand für das Update von H_{k+1} in (4.26) 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 abgeläst, dass noch besser war und bis heute zu den effizientesten Quasi-Newton Verfahren gehört: das Broyden–Fletcher–Goldfard–Shanno (BFGS) Verfahren. Die Idee des BFGS Verfahrens leitet sich unmittelbar aus der Idee des DFP Verfahren ab. Anstatt das Optimierungsproblem (4.24) 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 (4.24) von der Form:

$$\begin{aligned} &\min_H ||H - H_k||, \qquad \text{unter den Nebenbedingungen:} \\ H \;=\; H^T, \qquad &H y_k \;=\; s_k, \qquad \langle \mathbf{p}, H\mathbf{p}\rangle > 0, \forall \mathbf{p} \in \mathbb{R}^n/\{0\}. \end{aligned} \tag{4.27}$$

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 (4.27) als:

$$H_{k+1} \;=\; (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T, \qquad \text{mit } \rho_k := \frac{1}{\langle y_k, s_k\rangle}. \tag{4.28}$$

Das Update der Matrix H_k in (4.28) 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 $\mathcal{O}(\backslash)$, Matrix-Vektor Multiplikationen in $\mathcal{O}(n^2)$ und dyadische Produkte in $\mathcal{O}(\backslash^{\in})$ berechnen muss. Im Gegensatz hierzu würde eine naive Implementierung des BFGS-Updates in (4.28) 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 (4.28) 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.

4.3 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 [**hestenes_1952**] 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 $n \times n$ Matrix A 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 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 [**shewchuk_1994**]. Für eine ansprechende, interaktive Visualisierung des Verfahren der konjugierten Gradienten empfehlen wir den Mathematik Blog von Philipp Wacker [**wacker**].

4.3.1 Problemstellung

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

$$Ax \;=\; b \tag{4.29}$$

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 [**numerik1**] ist bekannt, dass es genau dann eine eindeutige Lösung $x \in \mathbb{R}^n$ für die Gleichung (4.29) 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 also im Folgenden immer davon aus, dass A eine symmetrische und positiv definite $n \times n$-Matrix ist. In diesem Fall ist das Bestimmen einer Lösung von (4.29) ein gut-gestelltes Problem und die Lösung ließe sich direkt angeben als:

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

Wie wir jedoch ebenfalls aus [**numerik1**] wissen ist die Inversion einer $n \times n$-Matrix $A \in \mathbb{R}^{n \times n}$ numerisch sehr aufwending 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 (4.29) mittels Inversion nicht durchführbar. Glücklicherweise liefert uns das Verfahren der konjugierten Gradienten (neben anderen iterativen Lösungsalgorithmen) eine Möglichkeit das lineare Gleichungssystem numerisch zu lösen.

Hierzu betrachten wir zunächst ein quadratisches Optimierungsproblem der Form

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle + c, \tag{4.30}$$

wobei A und b wie im Fall des linearen Gleichungssystems in (4.29) 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 4.25.

Das quadratische Minimierungsproblem in (4.30) ist äquivalent zum ursprünglichen linearen Gleichungssystem in (4.29), d.h., jede Lösung von (4.30) ist schon Lösung von (4.29) 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 Satz **??** für das Minimierungsproblem (4.30). 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 bestimmen als

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

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 Funktion F ist. Daraus können wir folgern, dass x_* ein stationärer Punkt von F ist und somit gelten muss:

$$\nabla F(x_*) \;=\; Ax_* - b \;=\; 0.$$

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

Der Satz **??** erlaubt es uns also ein quadratisches Optimierungsproblem der Form (4.30) numerisch zu lösen anstatt einen unbekannten Lösungsvektor für ursprüngliche lineare Gleichungssystem (4.29) 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 (4.30) und somit gegen die eindeutige Lösung des linearen Gleichungssystems (4.29) konvergiert. Hierfür benötigen wir noch zusätzliche Notation, um das angestrebte Verfahren vernünftig zu beschreiben.

DEFINITION 4.26: 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 (4.30) 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 4.27. In Bezug auf Definition **??** 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 bereits im Beweis von Satz **??** gesehen, 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 4.2 illustriert anschaulich die geometrische Bedeutung der beiden in Definition **??** eingeführten Vektoren. Wie man unschwer erkennt zeigen Fehler und Residuum im Allgemeinen nicht in die selbe Richtung. Das erklärt auch warum das Gradienten-

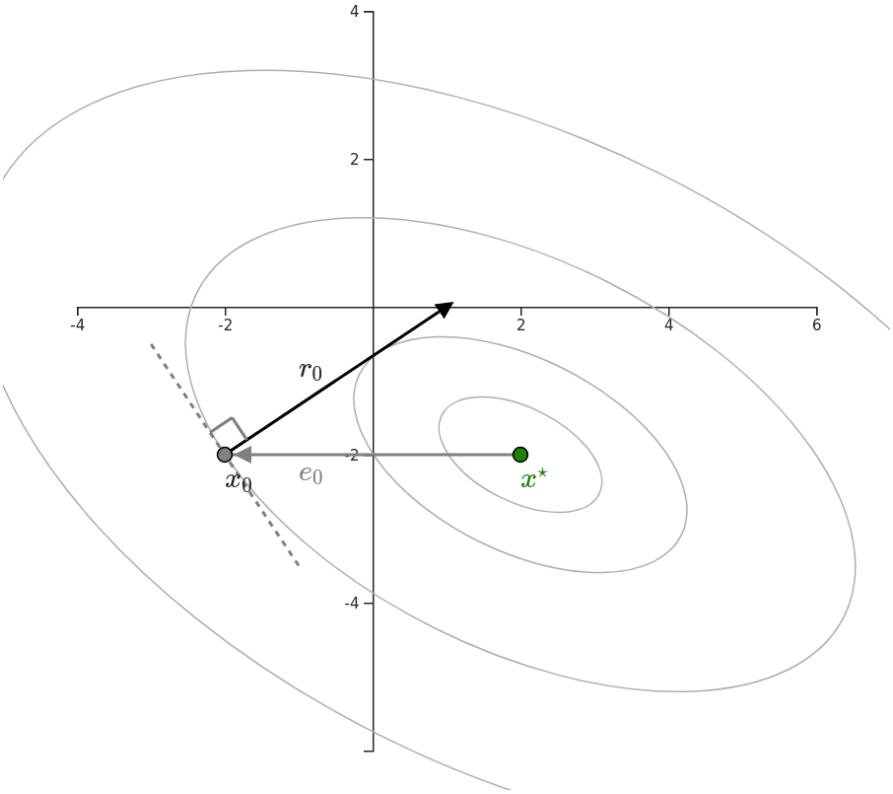

Abbildung 4.2: Visualisierung des Fehlers $e_0 \in \mathbb{R}^n$ und des Residuums $r_0 \in \mathbb{R}^n$ für einen Startpunkt $x_0 \in \mathbb{R}^n$.

abstiegsverfahren in Kapitel 4.2.1 selbst bei optimaler Schrittweite $alpha_k > 0$ nicht in einem Schritt die gesuchte Lösung $x_* \in \mathbb{R}^n$ erreicht.

4.3.2 Motivation

Um das Vorgehen beim Verfahren der konjugierten Gradienten zu motivieren rufen wir uns noch einmal das Gradientenabstiegsverfahren aus Kapitel 4.2.1 in Erinnerung. Nehmen wir an wir befinden uns im k-Schritt des Gradientenabstiegsverfahrens in Algorithmus 4.2.1 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 springen 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. Abbildung 4.1). Um dieses typische Verhalten besser zu verstehen können wir eine Vorüberlegung zur Schrittweitenwahl für das quadratische Optimierungsproblem in

(4.30) machen. Hierzu gehen wir analog zur Bestimmung der optimalen Schrittrichtung in Kapitel 4.2.1 vor, nur dass wir diesmal die Schrittrichtung $\mathbf{p}_k := -\nabla F(x_k)$ 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 Satz 4.1, um folgenden Zusammenhang herzustellen:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\alpha}F(x_{k+1}) &= \left\langle \nabla F(x_{k+1}), \frac{\mathrm{d}x_{k+1}}{\mathrm{d}\alpha} \right\rangle = \left\langle \nabla F(x_{k+1}), \frac{\mathrm{d}(x_k - \alpha\nabla F(x_k))}{\mathrm{d}\alpha} \right\rangle \\ &= \langle \nabla F(x_{k+1}), -\nabla F(x_k) \rangle = \langle \nabla F(x_{k+1}), \mathbf{p}_k \rangle \overset{!}{=} 0. \end{aligned} \tag{4.31}$$

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 Abbildung 4.1 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 (4.30).

LEMMA 4.28.

Sei $F\colon \mathbb{R}^n \to \mathbb{R}$ die quadratische Funktion aus (4.30). 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:

$$\alpha_k = \frac{\langle r_k, r_k \rangle}{\langle r_k, Ar_k \rangle}. \tag{4.32}$$

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 (4.32). □

Obwohl wir die optimale Schrittweite α_k in (4.5) für das quadratische Optimierungsproblem (4.30) 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. 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 $N \in \mathbb{N}$ Schritt des Gradientenabstiegsverfahren im selben Punkt $x_N \in \mathbb{R}^N$ mit nur zwei Iterationen landen.

Leider können wir nicht alle Schrittweiten aufaddieren, da wir für die optimalen Schrittlängen $alpha_k$ bereits alle Schritte $k = 0, \ldots, k-1$ berechnen müssten. Außerdem würde ein großer Schritt in die erste der beiden Richtungen dazu führen, dass man keinen Abstieg der Funktionswerte von F mehr realisiert sondern einen Aufstieg. Diese Beobachtung ist in Abbildung 4.3 illustriert.

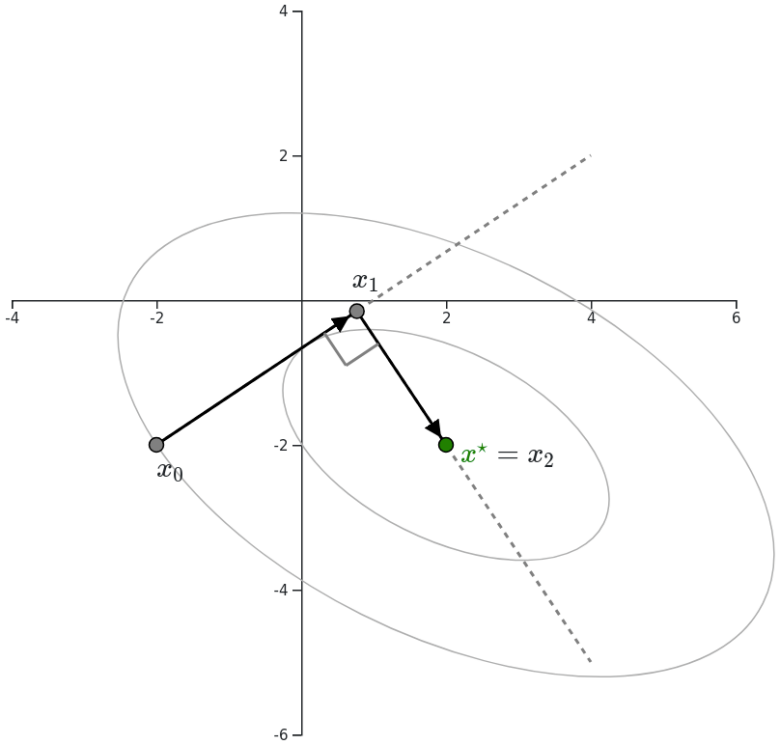

Abbildung 4.3: Illustration eines Abstiegsverfahrens mit zwei orthogonalen Richtungen. Mach beachte, dass die Schrittweite $\alpha_0 > 0$ so gewählt werden muss, dass man im ersten Schritt nicht in einem Punkt $x_1 \in \mathbb{R}^2$ mit minimalen Funktionswert $F(x_1)$ entlang der Richtung $x_0 - \alpha_0 \nabla F(x_0)$ endet.

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 Gradienten-

abstiegsverfahren mit der optimalen Schrittweite α_k in (4.32) im Fall des quadratischen Optimierungsproblems (4.30) in genau einem Schritt zum lokalen Minimum $x_0 \in \mathbb{R}^n$ führt, wenn der Fehler $e_0 = x_0 - x_*$ ein Eigenvektor von A ist. Man müsste bei der Wahl des Startpunktes x_0 jedoch viel Glück haben, um diese Forderung zu erfüllen. Darum wollen wir uns mit alternativen Ideen beschäftigen.

4.3.3 Orthogonale Abstiegsrichtungen

Wir wünschen uns einen Algorithmus, der ähnlich dem Gradientenabstiegsverfahren nur orthogonale Richtungen $\{\mathbf{d}_0, \dots, \mathbf{d}_{n-1}\}$ mit $\mathbf{d}_i \in \mathbb{R}^n$ 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 $\{\mathbf{d}_0, \dots, \mathbf{d}_{n-1}\}$ das lokale Minimum der Funktion zu erreichen. Damit hätten wir ein Abstiegsverfahren der Form

$$x_{k+1} \;=\; x_k + \alpha_k \mathbf{d}_k, \quad \alpha_k > 0, k = 0, \dots, n-1 \tag{4.33}$$

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

$$\langle e_{k+1}, \mathbf{d}_k \rangle \;\overset{!}{=}\; 0. \tag{4.34}$$

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

$$e_{k+1} \;=\; x_{k+1} - x_* \;=\; x_k + \alpha_k \mathbf{d}_k - x_* \;=\; e_k + \alpha_k \mathbf{d}_k,$$

können wir die Forderung (4.34) umformulieren zu:

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

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

$$\alpha_k \;=\; -\frac{\langle e_k, \mathbf{d}_k \rangle}{\langle \mathbf{d}_k, \mathbf{d}_k \rangle}. \tag{4.35}$$

Obwohl wir in (4.35) eine optimale Schrittweite α_k für das Verfahren mit orthogonalen Abstiegsrichtungen in (4.33) 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.

4.3.4 Konjugierte Abstiegsrichtungen

Obwohl unsere Idee von orthogonalen Abstiegsrichtungen in Kapitel 4.3.3 nicht zum Ziel geführt hat, so war die Idee gar nicht schlecht. Das Problem liegt begründet in der Forderung (4.34), nämlich dass der Fehlervektor e_{k+1} orthogonal zur aktuellen Richtung $\mathbf{d}_k$ stehen soll. Diese Forderung führt nämlich dazu, dass man orthogonale Vektoren erhält, die nicht an die Geometrie des quadratischen Minimierungsproblem (4.30) angepasst sind. Wenn man sich die Niveaulinien der Funktion F genauer anschaut (siehe zum Beispiel Abbildung 4.3), 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 dessen Eigenvektoren bestimmt. Daher wollen wir diese Eigenschaften bei der Konstruktion eines iterativen Abstiegsverfahren berücksichtigen. Hierzu führen wir folgendes hilfreiche Konzept ein.

DEFINITION 4.29: 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 $\{\mathbf{d}_0, \dots, \mathbf{d}_{n-1}\}$ zu erzwingen wie in Kapitel 4.3.3, fordern wir nun, dass diese konjugiert bezüglich der Matrix A sind und damit besser an das Problem angepasst.

BEMERKUNG 4.30. 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 (4.30) 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 schiebt, diese Vektoren anschließend orthogonal zueinander stehen. Diese Idee ist in Abbildung 4.4 dargestellt.

Anstatt also ein Abstiegsverfahren der Form (4.33) mit orthogonalen Vektoren $\{\mathbf{d}_0, \dots, \mathbf{d}_{n-1}\}$ zu verwenden wollen wir ein Abstiegsverfahren mit A-orthogonalen Vektoren $\{\mathbf{d}_0, \dots, \mathbf{d}_{n-1}\}$ konstruieren, d.h., wir verwenden das Iterationsschema

$$x_{k+1} \;=\; x_k + \alpha_k \mathbf{d}_k, \quad \alpha_k > 0,\; k = 0, \dots, n-1 \tag{4.36}$$

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

$$\langle \mathbf{d}_i, A\mathbf{d}_j\rangle = 0 \text{ für alle } i \neq j.$$

Wir nehmen für den Moment an, dass wir einen numerischen Algorithmus kennen mit dem wir eine Menge von A-orthogonalen Vektoren $\{\mathbf{d}_0, \dots, \mathbf{d}_{n-1}\}$ konstruieren können.

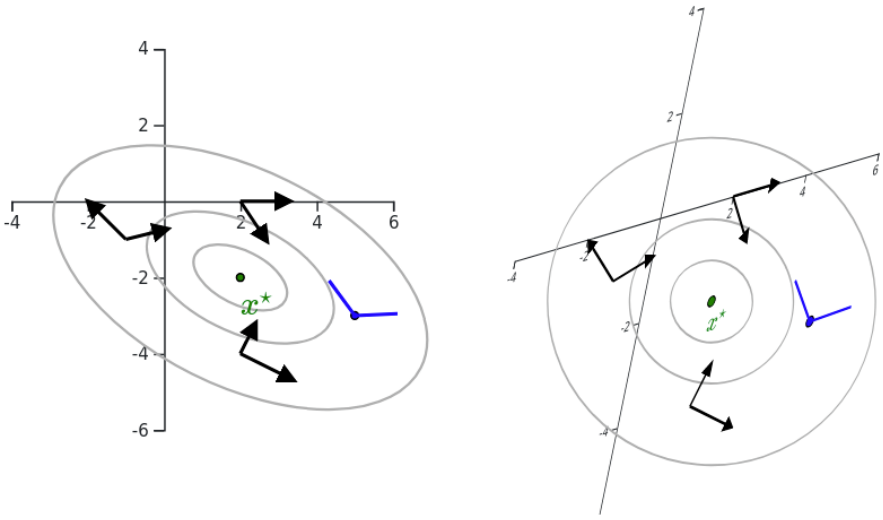

Abbildung 4.4: Illustration der Geometrie von konjugierten Vektoren im Referenzsystem $\mathbb{R}^2$ (links) und der selben Vektoren in einem symmetrisierten System bezüglich der Matrix A (rechts).

Wie man diese Menge konkret erhält werden wir uns im Anschluss erschließen. Sei also nun im Folgenden $\{\mathbf{d}_0, \ldots, \mathbf{d}_{n-1}\}$ eine gegebene Menge von A-orthogonalen Vektoren. Dann stellen wir uns die Frage, wie die optimalen Schrittweiten $\alpha_k > 0$ in (4.36) 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 $x_n = x_*$ an. Um das lokale Minimum wirklich in n Schritten zu erreichen müssen wir fordern, dass wir in jede Richtung $\mathbf{d}_k$ nur einmal gehen und der entstehende Fehler e_{k+1} konjugiert dazu ist bezüglich der Matrix A. Das entspricht der Forderung, dass man im entzerrten Problem auf der rechten Seite von Abbildung 4.4 nur orthogonale Richtungen verwendet. Wir wollen also folgende Eigenschaft erzwingen:

$$\langle Ae_{k+1}, \mathbf{d}_k \rangle \;=\; 0. \tag{4.37}$$

Analog zur Idee der orthogonalen Richtungen in Kapitel 4.3.3 können wir den Fehler e_{k+1} in (4.37) wieder entwickeln, um die optimale Schrittweitenlänge $\alpha_k > 0$ zu bestimmen

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

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

$$\alpha_k \;=\; \frac{\langle \mathbf{d}_k, r_k \rangle}{\langle \mathbf{d}_k, A\mathbf{d}_k \rangle}. \tag{4.38}$$

Im Gegensatz zur Idee der orthogonalen Richtungen in (4.35) lässt sich der Ausdruck in (4.38) berechnen und hängt nicht von dem unbekannten lokalen Minimum $x_* \in \mathbb{R}^n$ ab. Das heißt aus der Bedingung, dass die Abstiegsrichtung $\mathbf{d}_k \in \mathbb{R}^n$ A-orthogonal zum Fehlervektor $e_{k+1} \in \mathbb{R}^n$ stehen soll, konnten wir eine Schrittlänge $\alpha_k > 0$ finden, 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

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

Wir erhalten also für die Optimalität der unbekannten Schrittlänge $\alpha > 0$, dass die Abstiegsrichtung $\mathbf{d}_k \in \mathbb{R}^n$ und der Fehlervektor e_{k+1} konjugiert bezüglich der Matrix A sein müssen. Das ist aber genau die Eigenschaft, die wir bereits in (4.37) gefordert hatten. Wir erhalten also für eine gegebene Menge von A-orthogonalen Vektoren $\{\mathbf{d}_0, \dots, \mathbf{d}_{n-1}\}$ ein Abstiegsverfahren mit optimalen Schrittlangen $\alpha_k > 0$, die wir in (4.38) angeben können und die uns einen Abstieg garantieren.

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 4.31: Konvergenz des Abstiegsverfahrens.

Gegeben sei eine Menge von A-konjugierten Vektoren $\{\mathbf{d}_0, \dots, \mathbf{d}_{n-1}\}$ mit $\mathbf{d_k} \in \mathbb{R}^n / \{0\}$. Dann konvergiert das Abstiegsverfahren in konjugierte Richtungen

$$x_{k+1} = x_k + \alpha_k \mathbf{d}_k, \quad \alpha_k = \frac{\langle r_k, \mathbf{d}_k \rangle}{\langle \mathbf{d}_k, A\mathbf{d}_k \rangle} \tag{4.39}$$

in n Schritten gegen die Lösung $x_* \in \mathbb{R}^n$ des quadratischen Optimierungsproblems (4.30).

Beweis. Für den Beweis der Konvergenz des Iterationsverfahrens (4.39) betrachten wir zunächst den initialen Fehler e_0 durch die Wahl eines Startpunktes $x_0 \in \mathbb{R}^n$. Es ist klar, dass die Menge $\{\mathbf{d}_k\}_{k=0,\dots,n-1}$ 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:

$$e_0 = \sum_{k=0}^{n-1} \delta_k \mathbf{d}_k. \tag{4.40}$$

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

$$\langle \mathbf{d}_i^T A, e0 \rangle = \langle \mathbf{d}_i^T A, \sum_{k=0}^{n-1} \delta_k \mathbf{d}_k \rangle = \sum_{k=0}^{n-1} \delta_k \langle \mathbf{d}_i^T A, \mathbf{d}_k \rangle = \delta_i \langle \mathbf{d}_i^T A, \mathbf{d}_i \rangle.$$

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

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

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 $\mathbf{d}_i$ sind:

$$\delta_i \;=\; \frac{\langle \mathbf{d}_i^T A, e_0 + \sum_{k=0}^{i-1} \alpha_k \mathbf{d}_k\rangle}{\langle \mathbf{d}_i^T A, \mathbf{d}_i\rangle}. \tag{4.41}$$

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

$$e_i \;=\; e_0 + \sum_{k=0}^{i-1} \alpha_k \mathbf{d}_k. \tag{4.42}$$

Nun können wir die Gleichung (4.42) in die Bestimmung der Koeffizienten δ_i in (4.41) einsetzen und erhalten:

$$\delta_i \;=\; \frac{\langle \mathbf{d}_i^T A, e_0 + \sum_{k=0}^{i-1} \alpha_k \mathbf{d}_k\rangle}{\langle \mathbf{d}_i^T A, \mathbf{d}_i\rangle} \;=\; \frac{\langle \mathbf{d}_i^T A, e_i\rangle}{\langle \mathbf{d}_i^T A, \mathbf{d}_i\rangle} \;=\; \frac{\langle \mathbf{d}_i, Ae_i\rangle}{\langle \mathbf{d}_i^T A, \mathbf{d}_i\rangle} \;=\; -\frac{\langle \mathbf{d}_i, r_i\rangle}{\langle \mathbf{d}_i^T A, \mathbf{d}_i\rangle}.$$

Das bedeutet, dass die Koeffizienten δ_i in (4.41) gerade den negativen optimalen Schrittweiten α_i in (4.38) entsprechen, d.h., $\delta_i = -\alpha_i$. Aus der Basisdarstellung des initialen Fehlers $e_0 = x_0 - x_*$ in (4.40) können wir somit die Behauptung des Satzes folgern:

$$x_* \;=\; x_0 - e_0 \;=\; x_0 - \sum_{k=0}^{n-1} \delta_k \mathbf{d}_k \;=\; x_0 + \sum_{k=0}^{n-1} \alpha_k \mathbf{d}_k \;=\; x_n.$$

□

BEMERKUNG 4.32. Anstatt im Beweis von Theorem 4.31 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 (4.42):

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

Man sieht also das 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 $e_i = \sum_{k=i}^{n-1} \delta_k \mathbf{d}_k$ keine Anteile der Richtungen

$\{\mathbf{d}_j\}_{j=0,\dots,k-1}$ 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 (4.39) 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 sehr großen $n >> 1$ eine wichtige Rolle. △

BEISPIEL 4.33.

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 $\mathbf{d}_0, \mathbf{d}_1 \in \mathbb{R}^2/\{0\}$ als gegeben an mit:

$$\mathbf{d}_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \mathbf{d}_1 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}.$$

Als Startwert für unser Iterationsverfahren wählen wir $x_0 = (-2, 2)^T$. Wir sehen ein, dass die Vektoren $\mathbf{d}_0, \mathbf{d}_1$ konjugiert bezüglich der Matrix A sind, denn es gilt:

$$\langle \mathbf{d}_0, A\mathbf{d}_1 \rangle = (0,1) \cdot \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 3 \\ -1 \end{pmatrix} = (0,1) \cdot \begin{pmatrix} 7 \\ 0 \end{pmatrix} = 0.$$

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 α_0 für den ersten Schritt bestimmen mit:

$$\alpha_0 = \frac{\langle \mathbf{d}_0, r_0 \rangle}{\langle \mathbf{d}_0, A\mathbf{d}_0 \rangle} = \frac{4}{3}.$$

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

$$x_1 = x_0 + \alpha_0 \mathbf{d}_0 = \begin{pmatrix} -2 \\ -2/3 \end{pmatrix}.$$

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:

$$\alpha_1 \;=\; \frac{\langle \mathbf{d}_1, r_1\rangle}{\langle \mathbf{d}_1, A\mathbf{d}_1\rangle} \;=\; \frac{28}{21}.$$

Mit dieser können wir den nächsten und letzten Abstiegsschritt gehen und erhalten somit:

$$x_2 \;=\; x_1 + \alpha_1 \mathbf{d}_1 \;=\; \begin{pmatrix} 2 \\ -2 \end{pmatrix} \;=\; x_*.$$

Der folgende Satz hilft uns zu verstehen, warum ein Abstiegsverfahren mit konjugierten Richtungen besser funktioniert als das Gradientenabstiegsverfahren in Kapitel (4.2.1).

THEOREM 4.34.

Das Residuum $r_{i+1} = b - Ax_{i+1}$ des Abstiegsverfahren mit konjugierten Richtungen in (4.39) ist orthogonal zu allen bisherigen Abstiegsrichtungen $\mathbf{d}_j, j = 0, \ldots, i$, d.h.

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

Beweis. Aus Bemerkung 4.3.4 wissen wir, dass wir den Fehler e_{i+1} nach i Iterationen des Abstiegsverfahrens angeben können als

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

Wir können beide Seiten der Gleichung multiplizieren mit einem Vektor $-\mathbf{d}_j^T A$ für $j \leq i$ und erhalten damit:

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

□

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 Kapitel 4.3.2 gesehen haben. Hier war man nur optimal bezüglich der **letzten Abstiegsrichtung** und nicht bezüglich aller vorherigen Richtungen.

4.3.5 Konjugierte Gradienten

Wir haben im Kapitel 4.3.4 gesehen, dass wir ein iteratives Abstiegsverfahren mit konjugierten Abstiegsrichtungen $\{\mathbf{d}_j\}_{j=0,\ldots,n-1}$ verwenden können, um in n Iterationen die eindeutige Lösung des quadratischen Minimierungsproblems (4.30) und somit die Lösung

des linearen Gleichungssystems $Ax = b$ zu erhalten. Bisher sind wir jedoch davon ausgegangen, dass wir die Menge der konjugierten Vektoren $\{\mathbf{d}_0, \ldots, \mathbf{d}_{n-1}\}$ 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 $\{u_0, \ldots, u_{n-1}\}$ konjugieren bezüglich der Matrix A. Hierzu würde man die erste Abstiegsrichtung $\mathbf{d}_0$ des Abstiegsverfahrens mit konjugierten Richtungen wählen als den ersten Vektor der Menge, d.h., wir wählen $\mathbf{d}_0 = u_0$. Anschließend konstruieren wir die nächste Abstiegsrichtung $\mathbf{d}_1$ indem wir alle Komponenten von u_1 entfernen, die nicht A-orthogonal zu $\mathbf{d}_0$ sind. Für die nächste Abstiegsrichtung $\mathbf{d}_2$ gehen wir analog vor, nur müssen wir darauf achten alle Komponenten von u_2 zu entfernen, die nicht A-orthogonal zu $\mathbf{d}_0$ und $\mathbf{d}_1$ sind. Dieses Vorgehen lässt sich iterativ bis zum Vektor $\mathbf{d}_{n-1}$ fortführen und man erhält eine Menge von konjugierten Vektoren $\{\mathbf{d}_0, \ldots, \mathbf{d}_{n-1}\}$. Diese lassen sich in geschlossener Form angeben als:

$$\mathbf{d}_i \;=\; u_i + \sum_{k=0}^{i-1} \beta_{i,k}\mathbf{d}_k, \qquad \text{für } i = 1, \ldots, n-1. \tag{4.43}$$

Wir müssen jedoch die Koeffizienten $\beta_{i,k}$ so bestimmen, dass die Vektoren $\mathbf{d}_i$ konjugiert zu allen vorherigen Richtungsvektoren $\mathbf{d}_j, j < i$ sind. Um diese Koeffizienten zu bestimmen multiplizieren wir (4.43) von links mit einem Vektor $\mathbf{d}_j^T A$ für ein $j \in \{0, \ldots, i-1\}$ und erhalten

$$\begin{aligned}
\langle \mathbf{d}_j, A\mathbf{d}_i \rangle \;&=\; \langle \mathbf{d}_j, Au_i \rangle + \sum_{k=0}^{i-1} \beta_{i,k} \langle \mathbf{d}_j, A\mathbf{d_k} \rangle \\
\Rightarrow\; 0 \;&=\; \langle \mathbf{d}_j, Au_i \rangle + \beta_{i,j} \langle \mathbf{d}_j, A\mathbf{d}_j \rangle \\
\Rightarrow\; \beta_{i,j} \;&=\; -\frac{\langle \mathbf{d}_j, Au_i \rangle}{\langle \mathbf{d}_j, A\mathbf{d}_j \rangle}.
\end{aligned}$$

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 in $\mathcal{O}(n^3)$ liegt, was genau so teuer ist wie eine Invertierung der Matrix A, zum Beispiel mit dem Eliminationsverfahren von Gauss [**numerik1**].

Glücklicherweise gibt es eine Möglichkeit eine Menge von konjugierten Abstiegsrichtungen im Laufe des Iterationsverfahren (4.39) zu konstruieren ohne den numerischen Rechenaufwand des Gram-Schmidtschen Orthogonalisierungsverfahren zu benötigen. In der Tat gibt es in jeder Iteration einen Vektor, der A-orthogonal zu allen vorherigen Abstiegsrichtungen ist mit Ausnahme der letzten. Diese Aussage wird durch folgendes Lemma präzisiert.

LEMMA 4.35.

Wir befinden uns im i-ten Schritt des Abstiegsverfahrens für $i \in \{1, \dots, n-1\}$ mit konjugierten Richtungen in (4.39) und $\{\mathbf{d}_0, \dots, \mathbf{d}_{i-1}\}$ ist eine Menge von A-orthogonalen Vektoren und $u_i = r_i$ sei eine mögliche Abstiegsrichtung. Dann gilt:

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

und außerdem auch

$$\langle r_{i+1}, A\mathbf{d}_j \rangle \;=\; 0, \quad \text{für alle } j = 0, \dots, i-1.$$

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

$$\mathcal{D}_i \;:=\; \operatorname{span}\{\mathbf{d}_0, \dots, \mathbf{d}_{i-1}\}.$$

Wir gehen in diesem Beweis konstruktiv vor. Wir wählen als erste Abstiegsrichtung $u_0 = r_0 = \mathbf{d}_0$ und erhalten damit:

$$\operatorname{span}\{r_0\} \;=\; \operatorname{span}\{\mathbf{d}_0\} \;=\; \mathcal{D}_1.$$

Aus Satz **??** wissen wir, dass r_1 orthogonal zu $\mathbf{d}_0$ ist und somit folgt auch schon:

$$\langle \mathbf{d}_0, r_1 \rangle \;=\; \langle r_0, r_1 \rangle \;=\; 0.$$

Nun konstruieren wir die nächste Abstiegsrichtung $\mathbf{d}_1$ aus dem aktuellen Residuum $u_1 = r_1$ und dem Unterraum $\mathcal{D}_1$ und können 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

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

Aus Satz **??** wissen wir, dass $r_i \perp \mathcal{D}_{i+1}$ und damit wissen wir schon, dass die erste Aussage des Satzes gilt:

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

Wir drücken nun das Residuum r_i durch den Fehler e_i aus und erhalten:

$$r_i \;=\; -Ae_i \;=\; -A(e_{i-1} + \alpha_{i-1}\mathbf{d}_{i-1}) \;=\; r_{i-1} - \alpha_{i-1}A\mathbf{d}_{i-1}.$$

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

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

Wenn wir dies rekursiv entwickeln sehen wir ein, dass

$$D_i \;=\; \text{span}\{\mathbf{d}_0, A\mathbf{d}_0, \ldots, A^{i-1}d_0\}$$

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

$$\langle r_{i+1}, A\mathbf{d}_j\rangle, \qquad \text{für alle } j = 0, \ldots, i-1.$$

□

Abschnitt 4.3.5 sagt aus, dass das neue Residuum r_{i+1} ein guter Ausgangspunkt für eine neue Abstiegsrichtung $\mathbf{d}_{i+1}$ ist, da sie zu allen bisherigen A-orthogonalen Richtungen $\mathbf{d}_0, \ldots, \mathbf{d}_{i-1}$ konjugiert bezüglich der Matrix A ist. Wir müssen also nur noch dafür sorgen, dass der Vektor $u_{i+1} = \mathbf{d}_{i+1}$ A-orthogonal zur letzten Suchrichtung $\mathbf{d}_i$ ist. Dies ist numerisch wesentlich günstiger als einen beliebigen Richtungsvektor $u_{i+1} \in \mathbb{R}^n/\{0\}$ A-orthogonal zu machen bezüglich aller Vektoren $\mathbf{d}_0, \ldots, \mathbf{d}_i$.

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

THEOREM 4.36.

Wir befinden uns in einem neuen Punkt $x_{i+1} \in \mathbb{R}^n$ und wir wählen als mögliche Abstiegsrichtung $u_{i+1} = r_{i+1}$ das aktuelle Residuum, welches nach Lemma 4.3.5 bereits A-orthogonal zu fast allen vorherigen Abstiegsrichtungen $\{\mathbf{d}_0, \ldots, \mathbf{d}_{i-1}$ ist. Indem wir die neue Abstiegsrichtung definieren als

$$d_{i+1} \;:=\; r_{i+1} + \beta_{i+1}\mathbf{d}_i, \quad \beta_{i+1} \;:=\; \frac{\langle r_{i+1}, r_{i+1}\rangle}{\langle r_i, r_i\rangle}, \tag{4.44}$$

erhalten wir die Eigenschaft, dass diese Abstiegsrichtung nun A-orthogonal zu allen bisherigen Abstiegsrichtungen ist, d.h.,

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

Beweis. Wir müssen die mögliche Abstiegsrichtung $u_{i+1} \in \mathbb{R}^n/\{0\}$ so modifizieren, dass der resultierende Vektor $\mathbf{d}_{i+1}$ konjugiert zur $\mathbf{d}_i$ bezüglich der Matrix A ist. Mit dem Gram-Schmidtschen Orthogonalisierungsverfahren erhalten wir die Form

$$\mathbf{d}_{i+1} \;=\; r_{i+1} + \beta_{i+1}\mathbf{d}_i, \quad \beta_{i+1} \;=\; -\frac{\langle r_{i+1}, A\mathbf{d}_i\rangle}{\langle \mathbf{d}_i, A\mathbf{d}_i\rangle}.$$

Die neue Abstiegsrichtung $\mathbf{d}_{i+1} \in \mathbb{R}^n/\{0\}$ ist nach Konstruktion A-orthogonal zu allen vorherigen Abstiegsrichtungen $\{\mathbf{d}_0, \ldots, \mathbf{d}_i\}$, jedoch wollen wir den Koeffizienten β_{i+1}

noch näher charakterisieren im Folgenden. Aus dem Beweis von Lemma 4.3.5 wissen wir, dass wir das aktuelle Residuum ausdrücken können als

$$r_{i+1} \;=\; r_i - \alpha_i A\mathbf{d}_i.$$

Wir multiplizieren diese Gleichung von links mit r_i^T und erhalten

$$\langle r_{i+1}, r_{i+1}\rangle \;=\; \langle r_{i+1}, r_i\rangle - \alpha_i \langle r_{i+1}, A\mathbf{d}_i\rangle.$$

Wir wissen aus Lemma 4.3.5 jedoch auch, dass $\langle r_{i+1}, r_i\rangle = 0$ gilt und damit erhalten wir den Ausdruck

$$-\frac{1}{\alpha_i}\langle r_{i+1}, r_{i+1}\rangle \;=\; \langle r_{i+1}, A\mathbf{d}_i\rangle.$$

Wenn wir nun die optimale Schrittweite $\alpha_i = \frac{\langle r_i, \mathbf{d}_i\rangle}{\langle \mathbf{d}_i, A\mathbf{d}_i\rangle}$ aus (4.39) einsetzen erhalten wir für den Koeffizienten β_{i+1}:

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

Schlussendlich können wir den Nenner in diesem Ausdruck noch umschreiben, da die letzte Abstiegsrichtung auch mit dem Gram-Schmidtschen Orthogonalisierungsverfahren ausgedrückt werden kann

$$\langle r_i, \mathbf{d}_i\rangle \;=\; \langle r_i, r_i + \beta_i \mathbf{d}_{i-1}\rangle \;=\; \langle r_i, r_i\rangle + \beta_i \underbrace{\langle r_i, \mathbf{d}_{i-1}\rangle}_{=\,0} \;=\; \langle r_i, r_i\rangle.$$

Das Skalarprodukt in obiger Gleichung verschwindet auf Grund von Lemma **??** und somit erhalten wir schlussendlich für den Koeffizienten β_{i+1} den Ausdruck:

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

□

Mit der Herleitung von (4.44) können wir nun einen Algorithmus für das Abstiegsverfahren mit konjugierten Gradienten zum Lösen eines Gleichungssystems $Ax = b$ angeben.

ALGORITHMUS 4.37: Konjugierte Gradienten Abstieg.

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

Initialisierung
$\mathbf{d}_0 = r_0 = b - Ax_0$

Führe genau n Schritte durch

for $k = 0, \ldots, n-1$ **do**
  # Berechne Schrittweite
  $\alpha_k = \frac{\langle r_k, r_k \rangle}{\langle \mathbf{d}_k, A\mathbf{d}_k \rangle}$
  # Führe Abstiegsschritt durch
  $x_{k+1} = x_k + \alpha_k \mathbf{d}_k$
  if $k < n-1$ **then**
    # Berechne effizient neues Residuum
    $r_{k+1} = r_k - \alpha_k A\mathbf{d}_k$
    # 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
    $\mathbf{d}_{k+1} = r_{k+1} + \beta_{k+1}\mathbf{d}_k$
  end if
end for

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

4.4 Wahl der Schrittweite

Wir haben nun verschiedene Abstiegsverfahren der Form (4.3) kennen gelernt, die uns ein p_k liefern. Offen bleibt noch wie wir am besten (adaptiv während der Optimierung) die Schrittweite α_k wählen. Im Folgenden werden wir immer annehmen, dass p_k eine Abstiegsrichtung ist, d.h. es gilt

$$\nabla F(x_k) \cdot p_k < 0. \tag{4.45}$$

Ist α_k sehr klein, dann bleiben wir sicher in einem Bereich wo die Taylor-Approximation

$$F(x_k + \alpha_k p_k) \approx F(x_k) + \alpha_k \nabla F(x_k) \cdot p_k < F(x_k)$$

gilt, aber das Iterationsverfahren könnte aber sehr langsam werden. Ist andererseits α_k zu groß, dann ist die Abstiegsbedingung nicht mehr garantiert, es könnte z.B. passieren dass $x_k + \alpha_k p_k$ über ein Minimum springt. Deshalb benötigt man im wesentlichen zwei Bedingungen an α_k, die zu kleine und zu große Schritte verhindern.

Zunächst wollen wir eine theoretische Möglichkeit der optimalen Wahl von α_k untersuchen, nämlich jenes α_k das zum größtmöglichen Abstieg führt:

$$\overline{\alpha}_k = \arg\min_{\alpha} F(x_k + \alpha p_k).$$

Um $\overline{\alpha}_k$ ausrechnen zu können, müssen wir eindimensionale Probleme exakt lösen können, was im allgemeinen schwierig ist. Deshalb versuchen wir Bedingungen zu finden, die

wir leicht überprüfen können und für die wir Konvergenz garantieren können. Dafür benutzen wir den Vergleich mit der Linearisierung des Problems und definieren den daraus *erwarteten Abstieg*

$$E_k(\alpha) = F(x_k) + \alpha F'(x_k)p_k - F(x_k) = \alpha F'(x_k)p_k \tag{4.46}$$

und den tatsächlichen Abstieg

$$D_k(\alpha) = F(x_k + \alpha p_k) - F(x_k). \tag{4.47}$$

Unsere beiden Bedingungen können wir über die Abweichung von $D_k(\alpha)$ und $E_k(\alpha)$ formulieren. Dies machen die *Armijo-Goldstein Bedingungen*, die fordern, dass

$$c_1 E_k(\alpha) > D_k(\alpha) > c_2 E_k(\alpha) \tag{4.48}$$

gilt mit gegebenen Konstanten $0 < c_1 < c_2 < 1$ (wir beachten die Negativität von E_k). Die erste Bedingung garantiert, dass zumindest ein gewisser Teil des Abstiegs erreicht wird, die zweite verhindert, dass wir zu stark den Grenzwert Fall $\alpha \to 0$ annähern, in dem die Gleichheit mit Konstante gleich eins gilt. Eine typische Wahl der Parameter ist $c_1 = 0.1$ und $c_2 = 0.9$. Praktisch kann man den Algorithmus folgendermaßen realisieren: wir beginnen mit einem Wert von α, der im letzten Iterationsschritt erfolgreich war und testen die Armijo–Goldstein Bedingung. Ist die erste Ungleichung nicht erfüllt, verkleinern wir α (z.B. durch Halbierung), ist die zweite Ungleichung nicht erfüllt, dann vergrößern wir α. Um nicht in einen periodischen Zyklus zu geraten, sollte man zur Vergrößerung einen anderen Faktor wählen, etwa 1.5. Die Wahl der Schrittweite nach den Armijo-Goldstein Regeln ist also relativ einfach durchführbar und führt auch zu beweisbarer Konvergenz:

THEOREM 4.38.

Gegeben sei eine Wahl an Abstiegsrichtungen p_k, die für jedes x_k, das kein stationärer Punkt ist, eine uniforme Abstiegsrichung liefert, d.h. es gibt $\beta > 0$ mit

$$F'(x_k)p_k < -\gamma \|F'(x_k)\|^\beta$$

mit $\gamma > 0$. Dazu sei $F : \mathbb{R}^n \to \mathbb{R}$ eine nach unten beschränkte stetig differenzierbare Funktion, sodass die Niveaumenge $\{x \in \mathbb{R}^n \mid F(x) \leq F(x_0)\}$ beschränkt ist. Ist darüber hinaus p_k beschränkt, dann hat die Folge x_k eine konvergente Teilfolge und jeder Häufungspunkt ist ein stationärer Punkt von F.

Beweis. Ist $p_k = 0$ für endliches k, dann haben wir bereits einen stationären Punkt erreicht. Also können wir annehmen, dass $p_k \neq 0$ für alle k gilt und wir damit einen Abstieg haben. Dann impliziert die erste Armijo–Goldstein Bedingung

$$F(x_{k+1}) - F(x_k) < c_1 \alpha_k F'(x_k)p_k < 0$$

und damit induktiv

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

Also liegt die gesamte Folge (x_k) in der beschränkten Menge $\{x \in \mathbb{R}^n \mid F(x) \leq F(x_0)\}$ und hat somit eine konvergente Teilfolge x_{k_ℓ} mit Grenzwert $\overline{x}$. Tatsächlich erhalten wir sogar die stärkere Bedingung

$$F(x_k) + c_1 \sum_{j=0}^{k-1} (-\alpha_j F'(x_j) p_j) \leq F(x_0).$$

Mit der uniformen Schranke folgt dann

$$\sum_{j=0}^{k-1} \alpha_j \|F'(x_j)\|^{\beta+1} \leq \frac{1}{\gamma} \sum_{j=0}^{k-1} (-\alpha_j F'(x_j) p_j) \leq \frac{1}{\gamma c_1} F(x_0).$$

Damit gilt $\sqrt{\alpha_j} F'(x_j) \to 0$. Nun müssen wir noch zeigen, dass α_j nicht gegen Null konvergiert. Nehmen wir das Gegenteil für eine Teilfolge an, dann gilt auch $\alpha_{k_\ell} p_{k_\ell} \to 0$ und damit gilt für jedes ϵ für k hinreichend groß

$$F(x_{k_\ell} + \alpha_{k_\ell} p_{k_\ell}) - F(x_{k_\ell}) \leq (1 - \epsilon) \alpha_{k_\ell} F'(x_{k_\ell}) p_k < c_2 \alpha_{k_\ell} F'(x_{k_\ell}) p_{k_\ell},$$

was nicht möglich ist, da die Armijo-Bedingungen erfüllt sind. □

Für das Gradientenverfahren bzw. Varianten davon können wir mehr zeigen, weil p_k direkt in Verbindung mit $-F'(x_k)$ ist.

KOROLLAR 4.39.

Gegeben sei eine Wahl an Abstiegsrichtungen der Form

$$p_k = -A_k F'(x_k),$$

die wobei für jedes $k \in \mathbb{N}$ die Matrix $A_k \in \mathbb{R}^{n \times n}$ symmetrisch positiv definit ist, A_k uniform beschränkt und mit kleinstem Eigenwert durch $\lambda > 0$ für jedes k. Dazu sei $F : \mathbb{R}^n \to \mathbb{R}$ eine nach unten beschränkte stetig differenzierbare Funktion, sodass die Niveaumenge $\{x \in \mathbb{R}^n \mid F(x) \leq F(x_0)\}$ beschränkt ist. Dann hat die Folge x_k eine konvergente Teilfolge und jeder Häufungspunkt ist ein stationärer Punkt von F.

Beweis. Die Bedingungen für den Konvergenzsatz sind erfüllt mit $\gamma = \lambda$, $\beta = 1$. □

4.5 Nicht-differenzierbare Optimierung

Im Folgenden widmen wir uns noch einigen nicht-differenzierbaren Problemen, die man häufig in der Datenanalyse und Bildverarbeitung findet. Ein Beispiel ist das sogenannte Lasso-Problem der Minimierung von

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

für $A \in \mathbb{R}^{m\times n}$, typischerweise mit $m < n$. Mit der Lösung dieses Problems für $\alpha \to 0$ approximiert man die Lösung von $Ax = b$ mit minimaler ℓ^1-Norm, die auch oft die Lösung mit den meisten Nulleinträgen ist. Das Problem dabei ist, dass die ℓ^1-Norm

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

nicht differenzierbar ist. Eine interessante Klasse von Problemen, die wir betrachten wollen, ist von der Form

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

mit einer stetig differenzierbaren Funktion $G : \mathbb{R}^n \to \mathbb{R}$ und einer konvexen nicht notwendigerweise differenzierbaren Funktion $H : \mathbb{R}^n \to \mathbb{R}$. In diesem Fall können wir kein Gradienten-basiertes Optimierungsverfahren verwenden, sondern benötigen einen anderen Ansatz. Zunächst benötigen wir aber noch einige Definition um mit konvexen Funktionen umgehen zu können.

DEFINITION 4.40.

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

$$\partial H(x) = \{p \in \mathbb{R}^n \mid H(x) + p \cdot (y - x) \leq H(y) \quad \forall\, y \in \mathbb{R}^n\}.$$

Ein Element des Subdifferentials heißt Subgradient.

Wir sehen sofort, dass ein Minimum von H durch die Bedingung $0 \in \partial H(\overline{x})$ charakterisiert ist. Diese Bedingung ist äquivalent zu $H(\overline{x}) \leq H(y)$ für alle $y \in \mathbb{R}^n$. Für die Summe einer differenzierbaren und einer konvexen Funktion können wir eine entsprechende notwendige Bedingung herleiten:

LEMMA 4.41.

Sei $F = G + H$ mit G stetig differenzierbar und H konvex. $\overline{x} \in \mathbb{R}^n$ sei ein Minimierer von F, dann gilt $0 \in G'(\overline{x}) + \partial H(\overline{x})$, d.h. es existiert ein $p \in \partial H(\overline{x})$ mit $p + G'(\overline{x}) = 0$.

Beweis. Sei $\overline{x}$ ein Minimierer und $p = -F'(\overline{x})$. Dann gilt für x in einer Umgebung von $\overline{x}$

$$G(\overline{x}) + H(\overline{x}) = F(\overline{x}) \leq F(x) = G(x) + H(x) = G(\overline{x}) + G'(\overline{x})(x - \overline{x}) + H(x) + r(x)\|x - \overline{x}\|$$

mit $r(x) \to 0$ für $x \to \overline{x}$. Also folgt

$$H(\overline{x}) + p \cdot (\overline{x} - x) \leq H(x).$$

□

Wir beginnen mit einem einfachen Beispiel:

BEISPIEL 4.42.

Wir betrachten $H(x) = |x|$. Dann gilt für $x \neq 0$

$$\partial H(x) = \{\text{sign}(x)\},$$

während für $x = 0$

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

folgt.

Damit können wir auch das Subdifferential der ℓ^1-norm ausrechnen:

BEISPIEL 4.43.

Sei $H(x) = \|x\|_{\ell^1}$. Dann gilt

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

Wir können also von der Optimalitätsbedingung $F'(\overline{x}) + p = 0$ ausgehen und damit könnten wir ein Analogon zum Gradientenverfahren als

$$x^{k+1} = x^k - \tau^k(G'(x^k) + p^k), \qquad p^k \in \partial H(x^k)$$

betrachten. Allerdings ist nicht klar welchen Subgradienten wir auswählen sollen wenn mehrere existieren oder ob überhaupt einer existiert. Deshalb werden wir im Folgenden eine Variante betrachten, bei der automatisch Iterierte x^k mit nichtlinearem Subdifferential ausgewählt werden und ebenso ein $p^k \in \partial H(x^k)$.

4.5.1 Proximales Splitting

Die Idee des proximalen Splitting, auch *Forward-Backward Splitting* genannt ist es den differenzierbaren Teil genauso wie beim Gradientenverfahren auszuwerten (Vorwärts bei x^k), während der Subgradient bei der nächsten Iterierten ausgewertet wird (rückwärts bei x^{k+1}), d.h.

$$x^{k+1} = x^k - \tau^k(G'(x^k) + p^{k+1}), \qquad p^{k+1} \in \partial H(x^{k+1}).$$

Wir können die Iteration umschreiben zu

$$\frac{1}{\tau^k}x^{k+1} - \frac{1}{\tau^k}x^k + G'(x^k) + p^{k+1} = 0.$$

Diese Gleichung ist die Optimalitätsbedingung für die Minimierung der strikt konvexen Funktion

$$F^k(x) = \frac{1}{2\tau^k}\|x - x^k + \tau^k G'(x^k)\|^2 + H(x).$$

Wir können das Verfahren also effizient durchführen, wenn wir Funktionen der Form

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

effizient minimieren können. Dies ist bei der ℓ^1-Norm tatsächlich der Fall, wie wir im Folgenden sehen werden. Zuvor definieren wir noch den sogennanten Proximaloperator $\text{prox}_{\tau H}(y)$, als den eindeutigen Minimierer von Φ.

BEISPIEL 4.44.

Sei $H(x) = |x|$, dann ist $z = \text{prox}_{\tau H}(y)$ der Minimierer von

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

mit der Optimalitätsbedingung

$$\frac{1}{\tau}(y-z) \in \partial|z|.$$

Ist $z > 0$, so muss gelten $z = y - \tau$, dies ist nur möglich für $y \geq \tau$. Analog muss für $z < 0$ gelten $z = y + \tau$, was nur für $y \leq -\tau$ funktioniert. Ist $-\tau < y < \tau$, dann folgt $p = \frac{y}{\tau} \in [-1, 1] = \partial|0|$.

Den Proximaloperator für den Betrag $|x|$ ist der sogenannte Shrinkage-Operator

$$\text{shrink}_\tau(y) = \text{prox}_{\tau|\cdot|}(y) = \begin{cases} y - \tau & \text{falls } y > \tau \\ y + \tau & \text{falls } y < -\tau \\ 0 & \text{sonst.} \end{cases}$$

Damit können wir auch den Proximaloperator für $H(x) = \|x\|_{\ell^1}$ berechnen als

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

Mit Hilfe des Proximaloperators können wir allgemein das Vorwärts-Rückwärts-Splitting Verfahren schreiben als

$$x^{k+1} = \text{prox}_{\tau^k H}(x^k - \tau^k G'(x^k)).$$

Proximaloperatoren sind in gewisser Weise kontraktiv, d.h. es gilt:

LEMMA 4.45.

Sei J eine konvexe Funktion, dann ist prox_J Lipschitz stetig mit Modul 1.

Beweis. Ist $x_i = \text{prox}_J(y_i)$, dann gilt $x_i + p_i = y_i$, für ein $p_i \in \partial J(x_i)$. Subtrahieren wir diese Identitäten für $i = 1, 2$ so folgt

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

Ein Skalarprodukt mit $x_1 - x_2$ liefert

$$\|x_1 - x_2\|^2 + \langle p_1 - p_2, x_1 - x_2\rangle = \langle y_1 - y_2, x_1 - x_2\rangle \leq \|y_1 - y_2\| \, \|x_1 - x_2\|.$$

Nun haben wir aus der Definition eines Subgradienten

$$\langle p_1, x_1 - x_2 \rangle \geq F(x_1) - F(x_2)$$
$$\langle p_2, x_2 - x_1 \rangle \geq F(x_2) - F(x_2)$$

und addieren wir diese beiden so erhalten wir $\langle p_1 - p_2, x_1 - x_2 \rangle \geq 0$. Also gilt

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

was genau die gewünschte Lipschitz-Stetigkeit des Proximaloperators bedeutet. □

Analog wie im obigen Beweis können wir auch vorgehen um die Konvergenz des Forward-Backward Splitting Verfahrens zu verstehen. Wir schreiben

$$\frac{1}{\tau^k}(x^{k+1} - x^k) + p^{k+1} = -G'(x^k)$$

und nehmen ein Skalarprodukt mit $x^{k+1} - x^k$, dann folgt

$$\frac{1}{\tau^k}\|x^{k+1} - x^k\|^2 + \langle p^{k+1}, x^{k+1} - x^k \rangle = -\nabla G(x^k)(x^{k+1} - x^k).$$

Nehmen wir an, dass G zweimal stetig differenzierbar ist, dann folgt

$$-\nabla G(x^k)(x^{k+1} - x^k) = G(x^k) - G(x^{k+1}) + r^k,$$

mit Restglied $r^k \leq \frac{C^k}{2}\|x^{k+1} - x^k\|^2$, wobei C_k eine Schranke für die Norm der Hesse-Matrix von G in einer Umgebung von x^k mit Radius $\|x^{k+1} - x^k\|$ ist. Darüber hinaus gilt

$$\langle p^{k+1}, x^{k+1} - x^k \rangle \geq H(x^{k+1}) - H(x^k)$$

und damit

$$(\frac{1}{\tau^k} - C^k)\|x^{k+1} - x^k\|^2 + F(x^{k+1}) \leq F(x^k).$$

Theoretisch können wir τ^k so klein wählen, dass $\frac{1}{\tau^k} - C^k > \epsilon$ für kleines $\epsilon > 0$ gilt. Dann ist das Proximale Splitting ein Verfahren, dass F verkleinert und unter analogen Bedingungen wie in Satz 4.38 erhalten wir, dass x^k eine konvergente Teilfolge hat und

$$\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|^2 < \infty$$

gilt. Da die Iterierten auf einer beschränkten Menge bleiben, ist die zweite Ableitung von F dort beschränkt und τ^k kann nach unten beschränkt werden. Aus der Konvergenz

$$\nabla G(x^k) + p^{k+1} = \frac{1}{\tau^k}(x^{k+1} - x^k) = 0$$

folgern wir dann, dass jeder Häufungspunkt die Optimalitätsbedingung erfüllt.

4.5.2 Primal-Duale Verfahren

In vielen Fällen kann man den Proximaloperator von H selbst nicht gut berechnen, dann ist das Proximale Splitting nicht so leicht durchzuführen. Oft liegt dies daran, dass man die Form $H(x) = J(Bx)$ hat, mit einer nichtdiagonal Matrix B und einer konvexen Funktion J, deren Proximaloperator man gut berechnen kann. Ein häufiges Beispiel ist eine Variante des Lasso-Problems, in der man

$$\frac{1}{2}\|Ax - b\|_{\ell^2}^2 + \alpha\|Bx\|_{\ell^1} \to \min_{x\in\mathbb{R}^n}.$$

Der Proximaloperator von $\|Bx\|_{\ell^1}$ kann im allgemeinen nich einfach berechnet werden, währen wir für $B = I$ bereits eine explizite Form haben . Um dies zu umgehen führt man oft eine Nebenbedingung mit einer zusätzlichen Variable ein, die wir hier y nennen. Setzen wir $y = Bx$, so können wir

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

unter der Nebenbedingung $y = Bx$ minimieren.Um eine Nebenbedingung der obigen Form in der Minimierung einfach zu berücksichtigen, können wir die Idee der Lagrange Multiplikatoren verwenden. Wir definieren das Lagrange-Funktional

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

und überzeugen uns, dass

$$\inf_{x,y,Bx=y} \tilde{F}(x, y) = \inf_{x,y}\sup_{z} L(x, y, z)$$

gilt. Dies liegt daran, dass das Supremum über $L(x, y, z)$ gleich unendlich ist, wenn $Bx \neq y$, d.h. das Infimum wird dort sicher nicht angenähert und es ist reicht das Lagrange-Funktional auf der Menge wo $Bx = y$ gilt zu betrachten. Dort ist aber $L(x, y, z) = \tilde{F}(x, y)$.

Wir sehen also, dass wir im Fall der Optimierung mit Nebenbedingungen eigentlich ein Sattelpunktsproblem lösen wollen

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

Nehmen wir an, dass wir inf und sup vertauschen können (dies gilt wenn ein Sattelpunkt existiert), dann lösen wir

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

Nun können wir duale oder primal-duale Formulierungen herleiten, in dem wir die konvex Konjugierte definieren als

$$J^*(z) = \sup_y(\langle z, y\rangle - J(y)).$$

Dann haben wir äquivalent zum ursprünglichen Problem die primal-duale Formulierung

$$\inf_z \sup_x(J^*(z) + \langle -B^T z, x\rangle - G(x)),$$

die als Grundlage vieler primal-dualer Verfahren dient. In dieser Formulierung können wir um einen Sattelpunkt zu berechnen abwechselnd einen Abstiegsschritt in z und einen Aufstiegsschritt in x durchführen, z.B. einen Proximalschritt in z und einen Gradientenaufstiegschritt in x. Dies führt auf eine Variante des sogenannten Uzawa-Verfahrens

$$z^{k+1} = \text{prox}_{\tau^k J^*}(z^k + Bx^k), \quad x^{k+1} = -\sigma^k(\nabla G(x^k) + B^T z^k).$$

In diesem Fall müssen wir die Matrix B und B^T nur einmal mit einem Vektor multiplizieren in jedem Schritt, sowie den Proximaloperator von J^* ausrechnen. Letzteres ist aber einfach, wenn wir den Proximaloperator von J ausrechnen können, es gilt die sogenannte Moreau-Identität

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

Ist G ebenfalls konvex und hat einen einfach auszurechnenden Proximaloperator, so können wir auch hier statt dem Gradientenschritt einen Proximaloperator verwenden.

Wir bemerken zum Abschluss noch eine Dualitätsrelation: rechnen wir auch das Supremum in x aus, so erhalten wir das voll duale Problem

$$\inf_z(J^*(z) + G^*(-B^T z)),$$

die sogenannte *Fenchel-Dualität*, das zum ursprünglichen äquivalent ist, wenn J und G konvex sind.

Literatur