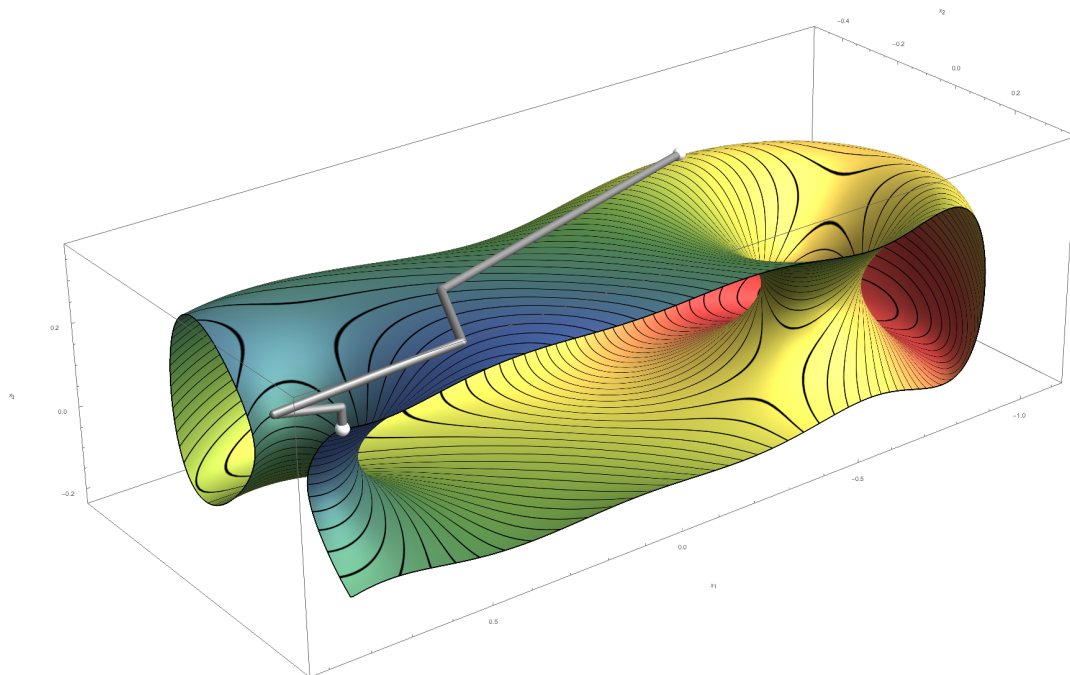


Skript zur Vorlesung Numerische Optimierung

TU Chemnitz, Wintersemester 2022

Henrik Schumacher

11. Januar 2023



Inhaltsverzeichnis

1	Einleitung	5
1.1	Was ist ein Optimierungsproblem?	5
1.2	Grobe Klassifikation von Optimierungsproblemen	5
1.3	Beispiele für Optimierungsprobleme	7
2	Grundlagen	12
2.1	Normierte Räume	12
2.2	Existenz von Lösungen	15
2.3	Hilberträume	18
2.4	Differenzierbarkeit	20
2.5	Einige Optimalitätskriterien	22
2.5.1	Notwendige Optimalitätsbedingungen erster Ordnung	22
2.5.2	Notwendige Optimalitätsbedingungen zweiter Ordnung	22
2.5.3	Hinreichende Optimalitätsbedingungen zweiter Ordnung	23
2.6	Konvexe Mengen	26
2.6.1	Beste Approximation	26
2.6.2	Hahn-Banachscher Trennungssatz	29
2.6.3	Darstellungssatz von Fréchet-Riesz	30
2.7	Konvexe Funktionen	31
3	Unrestringierte Optimierung	34
3.1	Der allgemeine Abstiegsalgorithmus	34
3.2	Richtung des steilsten Abstiegs	39
3.3	Die Methode des steilsten Abstiegs für konvexe quadratische Optimierungsprobleme	41
3.3.1	Konvergenzrate	44
3.3.2	Kantorovich-Ungleichung	46
3.4	Das lineare Konjugierte-Gradienten-Verfahren (CG-Verfahren)	48
3.4.1	Konvergenzrate	52
3.4.2	Der Algorithmus (ohne Vorkonditionierer)	54
3.4.3	Vorkonditionierer	57
3.4.4	Ein Beispiel	60
3.5	Newton-Verfahren	62
3.5.1	Satz von Newton-Kantorovich	62

3.5.2	Newton-Verfahren als Optimierungsverfahren; Residualnorm als Güte- kriterium	66
3.5.3	Übungsaufgaben	68
3.6	Schrittweitensteuerung	69
3.6.1	Exakte Schrittweiten	69
3.6.2	Armijo-Schrittweiten	73
3.6.3	Wolfe-Powell-Schrittweiten	76
3.7	Newton-artige Verfahren	80
3.8	Gauß-Newton-Verfahren	82
3.8.1	Anwendungen und Motivation	82
3.8.2	Konvergenzanalyse	84
3.8.3	Varianten	87
3.8.4	Übungsaufgaben	88
3.9	Inexakte Newton-Verfahren	89
3.10	Quasi-Newton-Verfahren	91
3.10.1	Die Broyden-Klasse	93
3.10.2	Davidon-Fletcher-Powell-Update-Formel (DFP)	95
3.10.3	Broyden-Fletcher-Goldfarb-Shanno-Update-Formel (BFGS)	97
3.10.4	Implementierung	98
3.10.5	Übungsaufgaben	102
4	Restringierte Optimierung	104
4.1	Optimalitätskriterien der restringierten Optimierung	104
4.1.1	Tangentialkegel	104
4.1.2	Notwendige Optimalitätsbedingung erster Ordnung	105
4.1.3	Hinreichende Optimalitätsbedingung zweiter Ordnung	106
4.1.4	Karush-Kuhn-Tucker-Bedingungen	107
4.1.5	Kegelgeometrie	109
4.1.6	Polyhedrische Kegel	115
4.1.7	Übungsaufgaben	119
4.2	Regularitätsbedingungen	121
4.2.1	LICQ bei Gleichungsnebenbedingungen	121
4.2.2	Mangasarian-Fromovitz Regularitätsbedingung (MFCQ)	123
4.2.3	LICQ bei Ungleichungsnebenbedingungen	125
4.2.4	Zusammenfassung	126
4.3	Kurcyusz-Robinson-Zowe-Regularitätsbedingung (KRZCQ)	128
4.3.1	Existenz von Lagrange-Multiplikatoren	129
4.3.2	Satz von der offenen Abbildung	131
4.3.3	Satz über implizite Funktionen von Robinson	134
4.3.4	KRZCQ und Abadie-Bedingung	139
4.3.5	Übungsaufgaben	141
4.3.6	Optimalitätsbedingungen zweiter Ordnung	142

4.4	Newton-Verfahren bei Gleichungsnebenbedingungen	147
4.4.1	Übungsaufgaben	150
4.5	Newton-Verfahren bei Ungleichungsnebenbedingungen	152
4.5.1	Aktive Mengen	152
4.5.2	Aktive-Mengen-Strategien	154
4.5.3	Semi-glattes Newton-Verfahren bei Ungleichungsnebenbedingungen .	157
4.5.4	Übungsaufgaben	162
4.6	Penalty-Verfahren	164
4.6.1	Übungsaufgaben	169
4.7	Barriere-Verfahren	171
4.7.1	Innere-Punkte-Verfahren für lineare Optimierungsprobleme	173
4.7.2	Übungsaufgaben	176
4.8	Augmented-Lagrangian-Verfahren	177
4.8.1	Die Lagrange-Funktion	177
4.8.2	Augmented-Lagrangian-Verfahren für Gleichungsnebenbedingungen	178
4.8.3	Quadrierte Schlupfvariablen	180
4.8.4	Augmented-Lagrangian-Verfahren für Ungleichungsnebenbedingungen	184
Literatur		187
Index		189

1 Einleitung

1.1 Was ist ein Optimierungsproblem?

Zu einer Funktion $F: M \rightarrow \mathbb{R} \cup \{\infty\}$ auf dem Raum M lautet das allgemeine *mathematische Optimierungsproblem*: Minimiere (oder maximiere) F , d. h. finde ein (oder alle) $x_* \in M$, so dass $F(x_*) \leq F(x)$ (oder $F(x_*) \geq F(x)$) gilt für alle $x \in M$. Man nennt so ein x_* dann einen *globalen Minimierer* (*globaler Maximierer*) von F . Die Menge M bezeichnet man dann als *zulässigen Bereich*, ihre Elemente als *zulässige Punkte* und die Funktion F als *Zielfunktion*.

Aus dem Englischen eingebürgert haben sich die synonymen aber etwas irreführenden Begriffe *mathematisches Programm* und *mathematische Programmierung* für mathematische Optimierungsprobleme bzw. für die mathematische Optimierung. Die Begriffe gehen auf die Bezeichnung *program* des US-Militär für z. B. Dienst- und Versorgungspläne zurück und haben nicht etwa mit Computerprogrammierung zu tun.

1.2 Grobe Klassifikation von Optimierungsproblemen

Optimierungsprobleme können beliebig kompliziert sein, so dass man sich bei einer systematischen Behandlung auf gewisse Klassen von Optimierungsproblemen beschränken muss. In dieser Vorlesung beschränken wir uns auf Optimierungsprobleme, bei denen M eine nichtleere Teilmenge eines normierten Raumes X mit Norm $\|\cdot\|$ ist. Je nach Beschaffenheit von X und M spricht man von

- *finiter Optimierung*: X ist endlichdimensional.
- *infiniter Optimierung*: X ist nicht endlichdimensional.
- *ganzzahliger Optimierung*: $M \subset \mathbb{Z}^n$.
- *diskreter Optimierung*: M ist als topologischer Raum total unzusammenhängend.
- *kontinuierlicher Optimierung*: Jeder Punkt $x \in M$ besitzt eine nicht-triviale, wegzusammenhängende Umgebung.¹

¹Das ist nur der *Versuch* einer Definition. Tatsächlich ist das gar nicht so einfach. Besitzt M mehr als eine Zusammenhangskomponente, so beinhalten auch kontinuierliche Optimierungsprobleme über M immer auch einen diskreten Anteil, nämlich die richtige Zusammenhangskomponente zu finden.

Oft beschreibt man die zulässige Menge mit Hilfe von Abbildungen $\Phi: X \rightarrow Y$ und $\Psi: X \rightarrow Z$ in andere normierte Vektorräume und einem *abgeschlossenen und konvexen Kegel* $K \subset Z$:

$$M = \{ x \in X \mid \Phi(x) = 0, \Psi(x) \in K \}$$

Ist Z endlichdimensional, so ist K typischerweise ein Quadrant bezüglich einer gewählten Basis von Z :

$$K = \{ z \in Z \mid \forall j \in \{ 1, \dots, \dim(Z) \} : z_j \leq 0 \}. \quad (1.1)$$

In diesem Fall hat man

$$M = \{ x \in X \mid \forall i \in \{ 1, \dots, \dim(Y) \} : \Phi_i(x) = 0, \forall j \in \{ 1, \dots, \dim(Z) \} : \Psi_j(x) \leq 0 \} \quad (1.2)$$

und man schreibt das Optimierungsproblem in der Form

$$\begin{array}{ll} \text{Minimiere} & F(x) \\ \text{u. d. N.} & \Phi_i(x) = 0, \quad \text{für } i \in \{ 1, \dots, \dim(Y) \}, \\ & \Psi_j(x) \leq 0 \quad \text{für } j \in \{ 1, \dots, \dim(Z) \}, \\ & x \in X. \end{array}$$

Ist $M = X$, so spricht man auch von *unrestringierter Optimierung*; ansonsten von *restringierter Optimierung*. Weiterhin spricht man auch von

- *glatter Optimierung*: F , Φ und Ψ besitzen mindestens eine Ableitung; oft auch zweite oder noch höhere Ableitungen.
- *nicht-glatter Optimierung*: F , Φ oder Ψ sind zumindest nicht im klassischen Sinn differenzierbar.
- *konvexer Optimierung*: Die Menge M ist konvex und F ist eine konvexe Funktion. Oft verlangt man auch dass die Funktionen Φ_i affin-linear und die Abbildungen Ψ_j konvex sind.
- *linearer Programmierung*: Die Funktion F ist affin-linear, die Abbildungen Φ und Ψ sind affin-linear und K ist ein Orthant.
- *quadratische Programmierung*: Die Funktion F ist quadratisch, die Abbildungen Φ und Ψ sind affin-linear und K ist ein Orthant.

Kontinuierliche Optimierung und diskrete Optimierung unterscheiden sich in ihren Methoden grundlegend. Wir werden uns hier auf kontinuierliche, glatte, finite Optimierung konzentrieren.

In der Regel ist es nicht mit vertretbaren Mitteln möglich, kontinuierliche Optimierungsprobleme exakt zu lösen. Für praktische Zwecke ist es aber meistens ausreichend, Lösungen von Optimierungsproblemen zu *approximieren*. Dazu verwendet man z. B. iterative Algorithmen, die von einem Startschätzer ausgehend diesen sukzessive verbessern. Im Idealfall kann

man zeigen, dass so ein Algorithmus für gewisse Problemklassen eine Lösung beliebig genau approximieren kann, wenn man nur genügend Ressourcen (Rechenzeit, Speicher, Implementationszeit) vorhalten kann. In der Praxis wird man aber natürlich Aufwand und Genauigkeit gegeneinander abwägen müssen.

Wir werden in dieser Vorlesung etliche solcher Algorithmen kennenlernen. Jeder von ihnen hat seine Vor- und Nachteile, die man abzuwägen in der Lage sein muss, um sie sinnvoll zur Anwendung zu bringen.

1.3 Beispiele für Optimierungsprobleme

Optimierungsprobleme können wir überall in unserem Alltag entdecken. Manchmal gehört aber auch schon ein wenig Finesse dazu, sie in mathematische Optimierungsprobleme zu überführen. Dieser Vorgang wird *Modellierung* genannt. Dazu gehört Identifikation und Kodierung sowohl der steuerbaren Größen (der Variablen) als auch der Nebenbedingungen. Am wünschenswertesten ist es natürlich, wenn ein gegebenes Optimierungsproblem in ein *gut behandelbares* mathematisches Optimierungsproblem übersetzt werden kann. Das sollte man nach Möglichkeit also schon bei der Modellierung beachten, benötigt aber auch ein wenig Erfahrung des Modellierenden.

Im Folgenden eine wahllose Sammlung von (mathematischen) Optimierungsproblemen.

Beispiel 1.1 (Das Problem des Handlungsreisenden (Traveling Salesman Problem))

Ein Handlungsreisender soll in einer Rundreise N Städte jeweils genau einmal aufsuchen und erst am Ende zur ersten Stadt zurückkehren. Dabei möchte er eine möglichst kurze Wegstrecke zurücklegen. Zunächst nummeriert der Handlungsreisende die Städte und speichert die Abstände zwischen ihnen in einer $(N \times N)$ -Matrix $\mathbf{D}$, so dass $\mathbf{D}_{ij}$ den Abstand zwischen der i -ten und j -ten Stadt beinhaltet. Danach kodiert er seine Route in einen Vektor $x \in \{0, 1\}^{N^2}$ wie folgt: Besucht er im i -ten Schritt die j -te Stadt, dann schreibt er $x_{N(i-1)+j} = 1$; ansonsten setzt er $x_{N(i-1)+j} = 0$.² Da er jede Stadt genau einmal besuchen soll, muss für jedes $j \in \{1, \dots, N\}$ gelten:

$$\Phi_j(x) = \sum_{i=1}^N x_{N(i-1)+j} = 1.$$

Da er zu jedem Zeitpunkt an genau einem dieser Orte sein muss, muss für jedes $i \in \{1, \dots, N\}$ gelten:

$$\Phi_{N+i}(x) = \sum_{j=1}^N x_{N(i-1)+j} = 1.$$

²Da die Stadt, die er im letzten Schritt aufsucht, immer diejenige sein muss, die er zuerst aufgesucht hat, braucht er diese nicht extra zu speichern.

Die Gesamtwegstrecke einer Route beträgt $F(x) = x^T Q x$ mit der folgenden Blockmatrix

$$Q = \begin{pmatrix} 0 & \mathbf{D} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & \ddots & \mathbf{D} \\ \mathbf{D} & 0 & \cdots & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{N^2 \times N^2}.$$

Dann hat der Handlungsreisenden das folgende mathematische Optimierungsproblem zu lösen:

$$\begin{aligned} &\text{Minimiere} && x^T Q x \\ &\text{u. d. N.} && \Phi_k(x) = 1, \quad \text{für } k \in \{1, \dots, 2N\}, \\ & && -x_l \leq 0, \quad \text{für } l \in \{1, \dots, N^2\}, \\ & && x_l - 1 \leq 0, \quad \text{für } l \in \{1, \dots, N^2\}, \\ & && x \in \mathbb{Z}^{N^2}. \end{aligned}$$

Unschwer ist zu sehen, dass es sich hierbei um ein diskretes, quadratisches Optimierungsproblem handelt.

Üblicherweise verwendet man eine andere mathematische Modellierung für das Problem des Handlungsreisenden. Eine Route kann man z. B. kodieren als einen Vektor aus Einsen und Nullen, der für jedes Paar $i \neq j$ von Städten einen Eintrag y_{ij} besitzt, der besagt, ob diese Verbindung Teil der Route ist ($y_{ij} = 1$) oder nicht ($y_{ij} = 0$). Die Länge der Reise ist dann $\sum_{i \neq j} \mathbf{D}_{ij} y_{ij}$. Die Bedingungen dafür, dass die Route eine Rundreise sein soll, kann man wieder durch affine Gleichheitsnebenbedingungen der Form $\Xi(y) = 0$ formulieren: Die Bedingung $\sum_{j \neq i} y_{ij} = 2$ besagt, dass Stadt i genau einmal erreicht und genau einmal wieder verlassen wird. Dass die Route zusammenhängend ist und keine sogenannten Kurzzykel auftreten, kann man ebenso durch lineare Gleichheitsnebenbedingungen garantieren, auf die wir hier aber nicht eingehen wollen.

Am Ende erhält man jedenfalls ein *lineares*, diskretes Optimierungsproblem von der Form

$$\begin{aligned} &\text{Minimiere} && \sum_{i \neq j} \mathbf{D}_{ij} y_{ij} \\ &\text{u. d. N.} && \Xi(y) = 0, \\ & && -y_{ij} \leq 0, \\ & && y_{ij} - 1 \leq 0, \\ & && y \in \mathbb{Z}^{N^2}. \end{aligned}$$

Beispiel 1.2 (Parameterschätzung im Rasch Modell) In der Psychometrie geht es oft darum, aus den Antworten von Probanden auf Testfragen gewisse Rückschlüsse auf personenbezogene Parameter zu ziehen. Klassisches Beispiel ist die Intelligenzforschung. Einer Anzahl N an Probanden mit “Intelligenz” $\theta = (\theta_1, \dots, \theta_N) \in \mathbb{R}^N$ werden jeweils M Fragen der Schwierigkeitsstufen $\sigma = (\sigma_1, \dots, \sigma_M) \in \mathbb{R}^M$ gestellt. Beantwortet der i -te Proband die j -te Frage richtig, so setzt man $x_{ij} = 1$; bei falscher Beantwortung setzt man $x_{ij} = 0$. Das Rasch-Modell geht

davon aus, dass die Antworten eines Probanden i nicht allein durch θ_i determiniert sind. Daher führt man eine Matrix-wertige Zufallsvariable X ein und modelliert die Wahrscheinlichkeit dafür, dass Proband i die Frage j richtig beantwortet durch eine logistische Wahrscheinlichkeitsverteilung

$$P(X_{ij} = 1) = \frac{\exp(\theta_i - \sigma_j)}{1 + \exp(\theta_i - \sigma_j)}.$$

Ist der i -te Proband also sehr intelligent ($\theta_i \gg \sigma_j$), so hat er eine Wahrscheinlichkeit nahe bei 1, die j -te Frage richtig zu beantworten.

Unter der Annahme, dass die skalaren Zufallsvariablen X_{ij} voneinander unabhängig sind, erhält man als Wahrscheinlichkeit für den Ausgang einer Messung mit Ergebnis $x \in \mathbb{R}^{N \times M}$:

$$L(\theta, \sigma) := P(X = x) = \prod_{i=1}^N \prod_{j=1}^M P(X_{ij} = x_{ij}) = \prod_{i=1}^N \prod_{j=1}^M \frac{\exp(x_{ij}(\theta_i - \sigma_j))}{1 + \exp(\theta_i - \sigma_j)}.$$

Die Parameter θ und σ sind nicht bekannt und sollen geschätzt werden. Dies kann man z. B. dadurch tun, dass man die *Likelihood* $L(\theta, \sigma)$ maximiert, d. h., man wählt diejenigen Parameter θ_* , σ_* , die den wahrgenommen Ausgang x des Experimentes “am wahrscheinlichsten” machen. Man nennt (θ_*, σ_*) daher auch den *Maximum-Likelihood-Schätzer*. Für numerische Zwecke ist es in der Regel ratsamer, die *Log-Likelihood*

$$\Lambda(\theta, \sigma) := \log(L(\theta, \sigma)) = \sum_{i=1}^N \sum_{j=1}^M \log \left(\frac{\exp(x_{ij}(\theta_i - \sigma_j))}{1 + \exp(\theta_i - \sigma_j)} \right)$$

zu maximieren (da der Logarithmus streng monoton steigend ist, ist dies äquivalent zum Maximieren der Likelihood),

Beispiel 1.3 (Kettenlinie) Eine Kette bestehe aus N Gliedern der Länge $\frac{L}{N}$ und der Masse $m > 0$. Für zwei Punkte P und Q im $\mathbb{R}^3$ mit $\|Q - P\| < L$ werde die Form gesucht, die die Kette unter Einfluss der Schwerkraft in ihrer stabilen Ruhelage einnimmt, wenn man sie mit ihren Enden an den Punkten P und Q befestigt. In der Mechanik wird gelehrt, dass die Kette dazu ihre Lageenergie minimieren muss. Beschreibt man die Lage der Kette durch die Lagen $x_0, x_1, \dots, x_N$ der Endpunkte der Kettenglieder und beschreibt $\xi \in \mathbb{R}^3$ den Gravitationsvektor, so beträgt die Lageenergie

$$E(x) = - \sum_{i=1}^n \frac{m}{2} \langle \xi, x_{i-1} + x_i \rangle.$$

Da die Enden der Kette fixiert sind und die Kettenglieder starr sind, müssen noch die Nebenbedingungen

$$x_0 = P, \quad x_N = Q, \quad \text{und} \quad \|x_i - x_{i-1}\|^2 = \left(\frac{L}{N}\right)^2 \quad \text{für } i \in \{1, \dots, N\}$$

berücksichtigt werden. Dies ist ein Beispiel für ein glattes Optimierungsproblem mit Gleichheitsnebenbedingungen. Die Zielfunktion ist linear und die Nebenbedingungen sind linear oder quadratisch. Lässt man die Anzahl N der Kettenglieder gegen ∞ gehen, so bekommt man ein unendlich-dimensionales Optimierungsproblem für die Lage eines Seiles, das keine Biegesteifigkeit besitzt.

Beispiel 1.4 (Das Isoperimetrische Problem)

Der Gründungsmythos Karthagos besagt, dass Dido auf der Flucht vor ihrem Bruder an der nordafrikanischen Küste anlandet. Dem dort herrschenden König versucht sie ein Stück Land zum Siedeln abzuhandeln. Dieser wirft verächtlich eine Kuhhaut vor ihr auf den Boden und verspricht ihr so viel Land wie sie darin einfasse könne. Daraufhin zerreißt sie die Kuhhaut in Streifen und knüpft daraus ein Seil der Länge L . Weil sie ja möglichst viel Land erhalten möchte, wird sie versuchen, das Seil entlang einer Kurve $\gamma: [0, 1] \rightarrow \mathbb{R}^2$ zu legen so dass der Flächeninhalt

$$A(x) = \int_0^1 \det(\gamma(t), \gamma'(t)) dt$$

unter den Nebenbedingungen

$$\int_0^1 |\gamma'(t)| dt = L \quad \text{und} \quad \gamma(0) = \gamma(1)$$

maximiert wird. Nutzt sie sogar die Küstelinie aus (vereinfacht bestehe die Landmasse aus der unteren Halbebene), so wird sie Nebenbedingungen der Form

$$\int_0^1 |\gamma'(t)| dt = L, \quad \gamma_2(0) = 0, \quad \gamma_2(1) = 0, \quad \gamma_1(0) - \gamma_1(2) \leq 0 \quad \text{und} \quad \gamma_2(t) \leq 0 \quad \text{für alle } t \in [0, 1]$$

verwenden. Dieses Optimierungsproblem ist schon etwas kniffliger denn es ist noch nicht von vorneherein klar, welche Kurven zulässig sein sollen. Zum Beispiel könnte man den Banachraum $C^1([0, 1], \mathbb{R}^2)$ der stetig differenzierbaren Kurven als zulässige Mengen nehmen. Vielleicht hat aber die optimale Lösung doch einen Knick, so dass sie nicht stetig differenzierbar ist?³ Jedenfalls ahnen wir schon, dass es sich hier wohl um ein infinites, restringiertes Optimierungsproblem handeln wird. Wir können ein endlich-dimensionales Problem dadurch erhalten, indem wir, ähnlich zum vorangegangenen Beispiel, für die Kurve γ nur Polygonzüge mit N Kanten zulassen. Für festes N können wir dann versuchen, das Problem mit dem Computer zu lösen und einen Minimierer γ_N zu bekommen. Für $N \rightarrow \infty$ würden wir dann erwarten, dass γ_N gegen eine echte Lösung γ_* des Isoperimetrischen Problems konvergiert. Dem ist auch so, aber der Nachweis dafür ist etwas komplizierter. Solche Nachweise sind typischerweise Gegenstand der numerischen Analysis, insbesondere der Theorie der *Finiten Elemente*.

³Man errät vielleicht schon, dass die optimalen Lösungen Kreise oder Halbkreise sein werden, die stetig differenzierbar parameterisiert werden können. Das mathematisch rigoros nachzuweisen ist aber gar nicht so einfach.

In dieser Vorlesung aber wollen wir uns mit der effizienten numerischen Optimierung der endlich-dimensionalen Probleme beschäftigen. Wie wir in den Übungen sehen werden, werden die N -dimensionalen Probleme auf zwei Arten schwieriger für wachsendes N : Zum einen haben wir mir und mehr Freiheitsgrade, was mehr Rechen- und Speicheraufwand in jedem Optimierungsschritt benötigt. Wir werden aber auch sehen, dass einfach black-box-Löser für wachsendes N immer mehr Optimierungsschritte benötigen. Deswegen werden wir auch Optimierungsverfahren kennenlernen, die weitestgehend *gitterunabhängig sind*, die also beliebige N ungefähr die gleich Anzahl an Optimierungsschritten benötigen, und somit deutlich effizienter sind.

2 Grundlagen

2.1 Normierte Räume

Wir werden in dieser Vorlesung regelmäßig mit verschiedenen normierten Räumen arbeiten. Deswegen wiederholen wir kurz einige Begriffe aus den Anfängervorlesungen. Auch wenn man viele dieser Begriffe auch für Vektorräume über den komplexen Zahlen einführen kann, sind für uns nur reelle Vektorräume von Interesse.

Ein *normierter Raum* ist ein Vektorraum X über den reellen Zahlen, zusammen mit einer Abbildung $\|\cdot\|: X \rightarrow [0, \infty[, x \mapsto \|x\|$ mit den folgenden Eigenschaften:

1. Homogenität: Für alle $t \geq 0$ und alle $x \in X$ gilt $\|t x\| = |t| \|x\|$.
2. Definitheit: Für alle $x \in X$ gilt: Ist $\|x\| = 0$, dann gilt $x = 0$.
3. Dreiecksungleichung: Für alle x und $y \in X$ gilt: $\|x + y\| \leq \|x\| + \|y\|$.

Für $x, y \in X$ bezeichnet man $\|x - y\|$ auch als *Abstand* von x zu y . Ist $(x_k)_{k \in \mathbb{N}}$ eine Folge und $x \in X$ mit $\|x_n - x\| \xrightarrow{k \rightarrow \infty} 0$, so sagt man, *die Folge $(x_k)_{k \in \mathbb{N}}$ konvergiert gegen x* . Man nennt dann x auch den *Grenzwert der Folge $(x_k)_{k \in \mathbb{N}}$* und schreibt

$$x = \lim_{k \rightarrow \infty} x_k.$$

Eine Folge $(x_k)_{k \in \mathbb{N}}$ in X nennt man *Cauchyfolge*, wenn es zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$ gibt so dass für alle $m, n \geq N$ gilt: $\|x_m - x_n\| \leq \varepsilon$. Einen normierten Raum $(X, \|\cdot\|)$ nennt man *Banachraum*, wenn er *vollständig* ist, das heißt, wenn jede Cauchyfolge in X auch konvergent ist.

Seien $(X, \|\cdot\|)$ und $(Y, \|\cdot\|)$ normierte, reelle Vektorräume. Mit

$$L(X; Y) := \{ A \in \text{Hom}(X; Y) \mid A \text{ ist stetig} \}$$

bezeichnen wir den Raum der stetigen linearen Abbildungen. Statt “lineare Abbildung” sagt man auch oft “linearer Operator”, insbesondere, wenn X oder Y unendlich-dimensional ist. Ist X endlich-dimensional, so sind alle linearen Abbildungen zwischen X und Y stetig. Ist aber X ein unendlich-dimensionaler, normierter Raum, so gibt es immer auch lineare Abbildungen, die nicht stetig sind. Man überlegt sich leicht, dass ein $A \in \text{Hom}(X; Y)$ stetig ist genau dann wenn es ein $C \geq 0$ gibt mit

$$\|A x\| \leq C \|x\| \quad \text{für alle } x \in X.$$

Die *Operatornorm* $\|A\|$ ist das kleinstmögliche $C \geq 0$, dass diese Abschätzung erlaubt:

$$\|A\| := \inf \{ C \geq 0 \mid \text{für alle } x \in X \text{ gilt } \|A x\| \leq C \|x\| \}.$$

Alternativ kann man die Operatornorm auch wie folgt charakterisieren:

$$\|A\| = \sup_{x \in X \setminus \{0\}} \frac{\|A x\|}{\|x\|}. \quad (2.1)$$

Eine besondere Rolle wird der *Dualraum* $X' := L(X; \mathbb{R})$ spielen. Der Dualraum zusammen mit der Operatornorm ist stets ein Banachraum. Elemente des Dualraums nennt man oft auch (*stetige*) *lineare Funktionale auf X*. Die *duale Paarung* ist die bilineare Abbildung

$$\langle \cdot, \cdot \rangle: X' \times X \rightarrow \mathbb{R} \quad \langle \xi, x \rangle := \xi(x) \quad \text{für } \xi \in X' \text{ und } x \in X.$$

Die zu $A \in L(X; Y)$ *duale Abbildung* $A' \in L(Y'; X')$ ist definiert durch $A' \eta := \eta \circ A$ für $\eta \in Y'$, d.h. man hat

$$\langle A' \eta, x \rangle = (\eta \circ A)(x) = \eta(A x) = \langle \eta, A x \rangle \quad \text{für alle } \eta \in Y' \text{ und alle } x \in X.$$

Man rechnet leicht nach, dass $\|A'\| \leq \|A\|$ gelten muss. Es gilt sogar $\|A'\| = \|A\|$, aber um das nachzuweisen benötigt man einen der Sätze von Hahn-Banach (siehe [Satz 2.34](#)).

Analog definiert man den *Bidualraum* $X'' := (X')'$. Als *kanonische Abbildung in den Bidualraum* bezeichnet man die lineare Abbildung

$$\psi_X: X \rightarrow X'' \quad \psi_X(x) := (\xi \mapsto \langle \xi, x \rangle).$$

Diese Abbildung ist immer eine Isometrie, das heißt, es gilt $\|\psi_X x\| = \|x\|$ für alle $x \in X$. Für endlichdimensionale normierte Räume und für Hilberträume (s.u.) ist dies stets ein Isomorphismus; im Allgemeinen aber nicht. Normierte Räume, für die ψ_X ein Isomorphismus ist, sind daher etwas Besonderes; man nennt sie *reflexive Räume*. Reflexive Räume sind stets Banachräume.¹ Aus dem Dimensionssatz der linearen Algebra leitet man leicht ab, dass jeder endlichdimensionale normierte Raum reflexiv ist. Folglich ist auch jeder endlichdimensionale normierte Raum ein Banachraum.

Beispiel 2.1 Natürlichstes Beispiel für einen endlichdimensionalen, reellen Vektorraum ist der $\mathbb{R}^n$. Durch Wahl einer Basis kann man jeden endlichdimensionalen Vektorraum mit dem $\mathbb{R}^n$ identifizieren. Ein Element $x \in \mathbb{R}^n$ können wir durch Spaltenvektoren schreiben (oder auch als $(n \times 1)$ -Matrizen, wie es MATLAB tut):

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{mit } x_1, x_2, \dots, x_n \in \mathbb{R}.$$

¹Aus dem einfachen Grund, weil X'' als Dualraum von X' mit der Operatornorm ein Banachraum ist und weil ψ_X ein isometrischer Isomorphismus ist.

Es gibt hier unendliche viele Normen, die alle zueinander äquivalent sind; die üblichste ist aber die *Euklidische Norm*²

$$\|x\|_2 := \left(\sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}}.$$

Der $\mathbb{R}^n$, zusammen mit jeder beliebigen Norm ist ein Banachraum. Elemente $\xi \in (\mathbb{R}^n)'$ des Dualraumes identifizieren wir mit Zeilenvektoren (oder $(1 \times n)$ -Matrizen)

$$\xi = (\xi_1 \quad \xi_2 \quad \dots \quad \xi_n).$$

Die duale Paarung ist dann gegeben durch

$$\langle \xi, x \rangle = \sum_{i=1}^n \xi_i x_i.$$

Eine lineare Abbildung $A \in L(\mathbb{R}^n; \mathbb{R}^m)$ kodiert man durch eine $(m \times n)$ -Matrix

$$A = \begin{pmatrix} A_{11} & \dots & A_{1n} \\ \vdots & & \vdots \\ A_{m1} & \dots & A_{mn} \end{pmatrix}$$

und man hat für $x \in \mathbb{R}^n$ und $\eta \in (\mathbb{R}^m)'$:

$$A x = \begin{pmatrix} \sum_{j=1}^n A_{1j} x_j \\ \vdots \\ \sum_{j=1}^n A_{mj} x_j \end{pmatrix} \quad \text{und} \quad A' \eta = \eta A = \left(\sum_{i=1}^m A_{i1} \eta_i \quad \dots \quad \sum_{i=1}^m A_{in} \eta_i \right).$$

²Andere Normen sind zum Beispiel

$$\|x\|_p := \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}} \quad \text{für } 1 \leq p < \infty \quad \text{und} \quad \|x\|_\infty := \max \{ |x_1|, \dots, |x_n| \}.$$

2.2 Existenz von Lösungen

Definition 2.2 Sei X ein topologischer Raum³. Eine Funktion $F: X \rightarrow \mathbb{R}$ heißt *unterhalbstetig* genau dann, wenn die Subniveaumengen $L_c^-(F) := \{x \in X \mid F(x) \leq c\}$ für alle $c \in \mathbb{R}$ abgeschlossen sind.

Beispiel 2.3 Jede stetige Funktion ist auch unterhalbstetig. Wie man aber leicht nachprüft ist die Funktion $f: \mathbb{R} \rightarrow \mathbb{R}$,

$$f(x) = \begin{cases} 0, & x \in [0, 1], \\ 1, & x \notin [0, 1] \end{cases}$$

zwar unterhalbstetig, aber nicht stetig.

Definition 2.4 Sei X ein topologischer Raum. Eine Menge $K \subset X$ nennt man *folgenkompakt* genau dann, wenn jede Folge in K eine konvergente Teilfolge besitzt.

Folgender Satz wird auch oft als *direkte Methode der Variationsrechnung* bezeichnet und findet in den unterschiedlichsten Bereichen Anwendung:

Satz 2.5 (Weierstrass)

Sei X ein nichtleerer, folgenkompakter, topologischer Raum und $F: X \rightarrow \mathbb{R}$ eine unterhalbstetige Funktion. Dann besitzt F ein Minimum in X .

BEWEIS. Sei $\inf(F) := \inf \{F(x) \mid x \in X\}$. Dann gibt es eine Folge von Punkten $x_n \in X$ mit $c_n := F(x_n) \searrow \inf(F)$ für $n \rightarrow \infty$. Weil X folgenkompakt ist, gibt es eine konvergente Teilfolge $(x_{n_k})_{k \in \mathbb{N}}$ und einen Punkt $x_* \in X$ mit $\lim_{k \rightarrow \infty} x_{n_k} = x_*$. Weil die Folge $(c_n)_{n \in \mathbb{N}}$ monoton fallend ist, gilt für jedes $m \in \mathbb{N}$:

$$x_{n_k} \in L_{c_{n_m}}^-(F) \quad \text{für alle } k \geq m.$$

Für jedes $n \in \mathbb{N}$ ist $L_{c_n}^-(F)$ aber eine abgeschlossene Menge. Also gilt auch $x_* \in L_{c_{n_m}}^-(F)$ für alle $m \in \mathbb{N}$. Die führt zu $\inf(F) \leq F(x_*) \leq \lim_{m \rightarrow \infty} c_{n_m} = \inf(F)$; d. h. x_* ist in der Tat ein globales Minimum. $\square$

Eine ganz analoge Aussage zum Satz von Weierstrass bekommt man, wenn man lediglich verlangt, dass eine Subniveaumenge $L_c^-(F)$ von F gibt, die nichtleer und folgenkompakt ist und auf der F unterhalbstetig ist.⁴

Wir verwenden den Satz von Weierstrass oft in Verbindung mit dem Satz von Heine-Borel aus der Analysis II:

³Z. B. ein normierter Raum.

⁴Natürlich kann eine Funktion auch Minimierer besitzen, wenn sie keine einzige folgenkompakte, nichtleere Subniveaumenge besitzt: Die Funktionen $\sin: \mathbb{R} \rightarrow \mathbb{R}$ und $f: \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^2 e^{-x^2}$ mögen als Beispiele dienen.

Satz 2.6 (Heine-Borel) Sei X ein endlichdimensionaler, normierter Raum. Eine Menge $K \subset X$ ist genau dann folgenkompakt, wenn sie beschränkt und abgeschlossen ist.

Folgerung 2.7 Seien X ein endlichdimensionaler, normierter Raum, $M \subset X$ eine Teilmenge, sei $F: X \rightarrow \mathbb{R}$ eine unterhalbstetige Funktion und es existiere ein $c \in \mathbb{R}$, so dass die Menge $L_c^-(F) \cap M$ nichtleer, beschränkt und abgeschlossen ist. Dann besitzt das Optimierungsproblem

$$\begin{array}{ll} \text{Minimiere} & F(x) \\ \text{u. d. N.} & x \in M \end{array}$$

eine Lösung.

Ebenso kann man die Begriffe einer oberhalbstetigen Funktion und einer Superniveaumenge $L_c^+(F)$ einführen und erhält ganz analoge Resultate für das Maximieren einer Funktion.

Vorsicht ist geboten bei unendlich-dimensionalen Räumen! In der Funktionalanalysis lernt man auch den folgenden Kompaktheitssatz von Riesz:

Satz 2.8 (Kompaktheitssatz von Riesz)

Der abgeschlossene Einheitskugel eines normierten Raumes X ist genau dann folgenkompakt, wenn X endlich-dimensional ist.

BEWEIS. Siehe z. B. [13, Satz I.2.7]. □

Der Satz von Heine-Borel gilt also nur in endlich-dimensionalen Räumen! Das bedeutet, dass der Satz von Weierstrass auf unendlich-dimensionalen normierten Räumen nicht direkt anwendbar ist. In vielen wichtigen Anwendungsfällen kann man aber die Topologie des Raumes so abschwächen, dass die Subniveaumengen der Zielfunktion folgenkompakt werden, ohne dass die Zielfunktion ihre Unterhalbstetigkeit verliert. Genauer hat man den Satz von Banach-Alaoglu, den wir hier nur zitieren aber nicht in allen Details beweisen wollen:

Satz 2.9 (Satz von Banach-Alaoglu)

Sei X ein reflexiver, reeller Banachraum. Dann ist der abgeschlossene Einheitsball

$$\bar{B}(0; 1) = \{ x \in X \mid \|x\| \leq 1 \}$$

in X sowohl kompakt als auch folgenkompakt bezüglich der sogenannten schwachen Topologie.

BEWEIS. Da X reflexiv ist, können wir $X = Y'$ identifizieren mit dem Dualraum von $U := X'$. Deshalb fallen auf Y' die schwache und die schwach-*-Topologie zusammen. Nach [13, Korollar VIII.3.12] $\bar{B}(0; 1)$ schwach-*-kompakt und damit also auch schwach kompakt. Da die schwache Topologie Hausdorffsch ist (denn zwei verschiedene Punkte kann man stets durch eine Hyperebene trennen), ist $\bar{B}(0; 1)$ auch schwach-abgeschlossen. Der Satz von Eberlein-Šmulian besagt nun, dass $\bar{B}(0; 1)$ auch schwach folgenkompakt ist. □

Satz 2.10 (Satz von Eberlein–Šmulian)

Sei X ein reeller Banachraum und sei $A \subset X$ eine beliebige Teilmenge. Dann sind die folgenden Aussagen äquivalent:

1. Der schwache Abschluss von A ist schwach kompakt.
2. Jede Folge in A besitzt eine schwach konvergente Teilfolge.

Ist A also schwach abgeschlossen, so sind äquivalent:

1. A ist schwach kompakt.
2. A ist schwach folgenkompakt.

BEWEIS. Einen sehr elementaren Beweis hierfür findet man in [14]. $\square$

Dank Satz 2.9 kann man den Satz von Weierstrass auch nutzen für Funktionen, die unterhalbstetig bezüglich der schwachen Topologie sind und deren Subniveaumengen $L_c^-(F)$ beschränkt sind. Für konvexe Funktionen erhalten wir den folgenden wichtigen Spezialfall:

Folgerung 2.11 Sei X eine reflexiver Banachraum und sei $F: X \rightarrow \mathbb{R}$ eine konvexe, unterhalbstetige Funktion für die alle Subniveaumengen $L_c^-(F)$, $c \in \mathbb{R}$ beschränkt sind. Dann besitzt F ein Minimum.

BEWEIS. Da die Subniveaumengen von F alle konvex und abgeschlossen sind, sind sie auch abgeschlossen in der schwachen Topologie. Also ist F unterhalbstetig auch bezüglich der schwachen Topologie auf X . Sei $c \in \mathbb{R}$ mit $K := L_c^-(F) \neq \emptyset$. Dann ist K konvex, abgeschlossen und beschränkt. Sei $\bar{B}(0; r)$, $r > 0$ ein abgeschlossener Ball, der noch ganz K enthält. Nach dem Satz von Banach-Alaoglu ist $\bar{B}(0; r)$ schwach folgenkompakt. Also ist K als schwach abgeschlossene Teilmenge im schwach folgenkompakten Ball $\bar{B}(0; r)$ wieder schwach folgenkompakt. Der Satz von Weierstrass liefert nun die Existenz eines Minimierers. $\square$

Im allgemeinen ist die Frage nach der Existenz von Minimierern eine klassische Fragestellungen der Variationsrechnung. Der Satz von Weierstrass wird oft auch als die *direkte Methode der Variationsrechnung* bezeichnet. Wir wollen hier nicht weiter darauf eingehen. Ziel in dieser Vorlesung ist es, numerische Methoden zu entwickeln, die Minimierer zu finden, wenn sie denn existieren.

2.3 Hilberträume

Der $\mathbb{R}^n$ besitzt noch eine weitere, sehr wichtige, aber etwas versteckte Eigenschaft, nämlich ein Skalarprodukt. Um uns dieser Struktur bewusst zu werden, abstrahieren wir auch diese Eigenschaft:

Definition 2.12 Einen reellen Banachraum $(X, \|\cdot\|)$ nennt man *Hilbertraum* genau dann, wenn es eine symmetrische Bilinearform $g: X \times X \rightarrow \mathbb{R}$ gibt, so dass $g(x, x) > 0$ und $\|x\| = g(x, x)^{\frac{1}{2}}$ gilt für alle $x \in X$. Diese Bilinearform nennt man *inneres Produkt* oder auch *Skalarprodukt*.

Neben der Dreiecksungleichung ist in Hilberträumen noch eine weitere Ungleichung, die *Cauchy-Schwarz-Ungleichung*, von besonderer Bedeutung:

$$g(x, y) \leq \|x\| \|y\| \quad \text{für alle } x, y \in X. \quad (2.2)$$

Der *Satz des Pythagoras* gilt auch in Hilberträumen und wird wie folgt formuliert: Sind $x, y \in X$ *senkrecht* zueinander, d. h. $g(x, y) = 0$, so gilt $\|x + y\|^2 = \|x\|^2 + \|y\|^2$. Es gilt die *Parallelogrammgleichung*. Diese besagt für $x, y \in X$, dass

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2.$$

Tatsächlich lässt sich sogar zeigen, dass ein Banachraum $(X, \|\cdot\|)$ genau dann ein Hilbertraum ist, wenn in ihm die Parallelogrammgleichung gilt.

Außerdem definieren wir den *Riesz-Isomorphismus*

$$J_g: X \rightarrow X', \quad \langle J_g(x), y \rangle = g(x, y) \quad \text{für } x, y \in X.$$

Dass J_g linear ist, sieht man leicht ein. Dass $J_g(x)$ ein stetiges lineares Funktional ist und dass J_g eine stetige Abbildung ist, folgt aus der Cauchy-Schwarz-Ungleichung. Die Abbildung J_g ist injektiv, denn für $x \neq 0$ gilt $\langle J_g(x), x \rangle = g(x, x) > 0$. Ist X endlich-dimensional, so folgt hieraus schon mit dem Dimensionssatz aus der linearen Algebra, dass J_g in der Tat ein Isomorphismus ist. Für allgemeine Hilberträume müssen wir uns ein wenig mehr strecken. Dazu später mehr (siehe [Satz 2.35](#)).

Seien (X, g) und (Y, h) Hilberträume mit Riesz-Isomorphismen $J_g: X \rightarrow X'$ und $J_h: Y \rightarrow Y'$. Für eine stetige, lineare Abbildung $A \in L(X; Y)$ definiert man ihre *adjungierte Abbildung* $A^* \in L(Y; X)$ durch $A^* := J_g^{-1} A' J_h$. Sie ist charakterisiert durch

$$h(Ax, y) = g(x, A^*y) \quad \text{für alle } x \in X \text{ und alle } y \in Y.$$

Beispiel 2.13 (Standard-Skalarprodukt)

Wir führen [Beispiel 2.1](#) fort. Das *Standard-Skalarprodukt* auf dem $\mathbb{R}^n$ ist gegeben durch

$$g(x, y) := \sum_{i=1}^n x_i y_i \quad \text{für } x, y \in \mathbb{R}^n.$$

Man rechnet leicht nach, dass $J_g: \mathbb{R}^n \rightarrow (\mathbb{R}^n)'$ identisch ist zur Transposition von Spaltenvektoren, d. h. es gilt

$$J_g \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix}.$$

Ebenso überprüft man, dass für eine lineare Abbildung $A \in L(\mathbb{R}^n; \mathbb{R}^m)$ die Matrix-Darstellung der Adjungierten A^* aus der Matrix-Darstellung

$$A = \begin{pmatrix} A_{11} & \dots & A_{1n} \\ \vdots & & \vdots \\ A_{m1} & \dots & A_{mn} \end{pmatrix}$$

von A durch Transposition hervorgeht, d. h.

$$A^* = A^T = \begin{pmatrix} A_{11} & \dots & A_{m1} \\ \vdots & & \vdots \\ A_{1n} & \dots & A_{mn} \end{pmatrix}.$$

2.4 Differenzierbarkeit

Definition 2.14 Seien $U \subset X$ eine offene Menge im normierten Raum X , $F: U \rightarrow Y$ eine Abbildung in den normierten Raum Y und $x \in U$ ein gegebener Punkt.

1. Man sagt, F ist *Gâteaux-differenzierbar* bei x genau dann, wenn es eine stetige lineare Abbildung $A \in L(X; Y)$ gibt, so dass für alle Richtungen $u \in X$ gilt:

$$\lim_{t \rightarrow 0} \frac{F(z + t u) - F(z)}{t} = A u.$$

Die lineare Abbildung $DF(x) := A$ nennt man dann *Gâteaux-Ableitung* von F bei x .

2. Man sagt, F ist *Fréchet-differenzierbar* bei x oder kurz *differenzierbar* bei x genau dann, wenn es eine stetige lineare Abbildung $A \in L(X; Y)$ gibt, so dass es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt, so dass für alle $y \in B(x; \delta) \cap U$ gilt dass

$$\|F(y) - F(x) - A(y - x)\| \leq \varepsilon \|y - x\|.$$

Die lineare Abbildung $DF(x) := A$ nennt man dann *Fréchet-Ableitung* oder schlicht die *Ableitung* von F bei x .

Ist F bei allen $x \in U$ (Gâteaux-/Fréchet-)differenzierbar, so sagen wir schlicht, F sei (Gâteaux-/Fréchet-)differenzierbar. Die (Gâteaux-/Fréchet-)Ableitung DF ergibt dann eine Abbildung $DF: U \rightarrow L(X; Y)$ in den (mit der Operatornorm) normierten Raum $L(X; Y)$. Ist F (Fréchet-)differenzierbar und ist DF stetig, so sagen wir, F sei *stetig differenzierbar*.

Definition 2.15 Seien $U \subset X$ eine offene Menge im normierten Raum X , $F: U \rightarrow Y$ eine Abbildung in den normierten Raum Y und $x \in U$ ein gegebener Punkt. Ist F differenzierbar in einer Umgebung V von x und ist auch $DF: V \rightarrow L(X; Y)$ differenzierbar bei x , so sagen wir, F ist zweimal differenzierbar. Die *zweite Ableitung* ist dann eine stetige lineare Abbildung $DDF(x) \in L(X; L(X; Y))$. Ist F auf ganz U zweimal differenzierbar, so nennt die Abbildung $DDF: U \rightarrow L(X; L(X; Y))$ einfach die *zweite Ableitung* von F . Oft ersetzen wir $DDF(x)$ durch die folgende stetige Bilinearform

$$D^2F(x): X \times X \rightarrow Y, \quad D^2F(x)(u, v) := (DDF(x) v) u$$

und bezeichnen auch D^2F als zweite Ableitung. Analog definiert man rekursiv für jedes $k \in \mathbb{N}$ die $(k + 1)$ -fache Differenzierbarkeit und die $(k + 1)$ -te Ableitung:

$$D^{k+1}F(x): \prod_{i=1}^{k+1} X \rightarrow Y, \quad D^{k+1}F(x)(u_1, \dots, u_k, u_{k+1}) := (DD^kF(x) u_{k+1})(u_1, \dots, u_k).$$

Setzen wir den Vektor $u \in X$ k -mal in $D^kF(x)$ ein, so schreiben wir einfach

$$D^kF(x)(u)^{(k)} = D^kF(x)(u, \dots, u).$$

Satz 2.16 (Schwarz) Ist $F: U \subset X \rightarrow Y$ k -mal stetig differenzierbar, $k \geq 2$, so ist $D^k F(x)$ eine symmetrische k -Multilinearform.

BEWEIS. Man kopiert einfach den endlichdimensionalen Beweis. $\square$

Aufgabe 2.17

Seien X, Y normierte Räume und $m \in \mathbb{N}$. Eine Abbildung $B: X^m \rightarrow Y$ nennt man m -linear genau dann, wenn sie linear in jedem ihrer Argumente ist. Für eine m -lineare Abbildung $B: X^m \rightarrow Y$ definieren wir

$$\|B\| = \sup \left\{ \frac{\|B(u_1, \dots, u_m)\|_Y}{\|u_1\|_X \cdots \|u_m\|_X} \mid u_1, \dots, u_m \in X \setminus \{0\} \right\}.$$

Man nennt B beschränkt genau dann, wenn $\|B\| < \infty$ gilt. Zeigen Sie:

1. Eine m -lineare Abbildung $B: X^m \rightarrow Y$ ist genau dann beschränkt, wenn sie stetig ist.
Hinweis: Ist $\|B\| = \infty$, dann findet man eine Folge $(x_1^{(k)}, \dots, x_n^{(k)})_{k \in \mathbb{N}}$ von m -Tupeln mit $x_i^{(k)} \rightarrow 0$ aber $\|B(x_1^{(k)}, \dots, x_n^{(k)})\|_Y \rightarrow 1$ für $k \rightarrow \infty$.
2. Der Raum $L^m(X; Y)$ der beschränkten, m -linearen Abbildungen zusammen mit $\|\cdot\|$ ist ein normierter Raum.

Satz 2.18 (Taylor-Formel) Sei $k \in \mathbb{N}$, $F: U \subset X \rightarrow Y$ eine $(k+1)$ -mal stetig differenzierbare Abbildung. Sei $x \in U$ und sei $y \in U$ so dass die Verbindungsstrecke von x nach y noch komplett in U liegt. Dann gilt

$$F(y) - \sum_{j=0}^k \frac{1}{j!} D^j F(x)(y-x)^{(j)} = \int_0^1 \frac{1}{k!} (1-t)^k D^{k+1} F(x+t(y-x))(y-x)^{(k+1)} dt.$$

Die Abbildung $T^k F(x): y \mapsto \sum_{j=0}^k \frac{1}{j!} D^j F(x)(y-x)^{(j)}$ nennt man das Taylor-Polynom der Ordnung k von F bei x . Insbesondere liefert uns die Taylor-Formel die folgende Aussage: Für jedes $\delta > 0$ mit $B(x; \delta) \subset U$ und für alle $y \in B(x; \delta)$ gilt:

$$\|F(y) - T^k F(x)(y)\|_Y \leq \frac{1}{k!} \sup_{z \in B(x; \delta)} \|D^{k+1} F(z)\| \|y-x\|_X^{k+1}.$$

Die Norm $\|D^{k+1} F(z)\|$ ist hier im Sinne von [Aufgabe 2.17](#) zu verstehen.

Da wir $D^{k+1} F$ als stetig annehmen und da U offen ist, finden wir zu jedem $\varepsilon > 0$ ein $\delta > 0$ so dass für alle $y \in B(x; \delta)$ gilt:

$$\|F(y) - T^k F(x)(y)\|_Y \leq \frac{1}{k!} (\|D^{k+1} F(x)\| + \varepsilon) \|y-x\|_X^{k+1}.$$

2.5 Einige Optimalitätskriterien

Die Taylorformel besagt, dass mehrfach stetig-differenzierbare Funktionen gewissermaßen gut "vorhersehbar" sind. In der Optimierung können wir das auf verschiedene Arten ausnutzen. Hier nutzen wir das aus, um notwendige und hinreichende Bedingungen für lokale Minimierer herzuleiten.

Folgende Definitionen und Resultate sind für endlich-dimensionale normierte Räume vermutlich schon aus der Analysis II bekannt.

Definition 2.19 (Globale und lokale Extrema) Sei $M \subset X$ eine Menge in einem topologischen Raum und sei $F: M \rightarrow \mathbb{R}$ eine Zielfunktion. Ein Punkt $x_* \in M$ heißt

1. *lokales Minimum* von F , wenn es eine Umgebung $U \subset M$ von x_* gibt mit $F(x_*) \leq F(x)$ für alle $x \in U$.
2. *striktes lokales Minimum* von F , wenn es eine Umgebung $U \subset M$ von x_* gibt mit $F(x_*) < F(x)$ für alle $x \in U \setminus \{x_*\}$.
3. *globales Minimum* von F , wenn $F(x_*) \leq F(x)$ gilt für alle $x \in M$.

2.5.1 Notwendige Optimalitätsbedingungen erster Ordnung

Satz 2.20 (Notwendige Optimalitätsbedingung erster Ordnung)

Sei $M \subset X$ eine offene Menge in einem normierten Raum, sei $F: M \rightarrow \mathbb{R}$ eine differenzierbare Abbildung und sei $x_* \in M$ ein lokales Minimum von F . Dann gilt notwendig $DF(x_*) = 0$.

BEWEIS. Sei U eine Umgebung von x_* , auf der x_* ein Minimum der eingeschränkten Zielfunktion $F|_U$ ist. Für jede Richtung $u \in X$ gibt es ein $\delta > 0$, sodass $x_* + tu \in U$ liegt für alle $t \in [0, \delta]$. Für alle $t \in]0, \delta[$ gilt dann

$$\frac{1}{t} (F(x_* + tu) - F(x_*)) \geq 0$$

und durch Grenzübergang $t \rightarrow 0$ erhält man $DF(x_*)u \geq 0$. Die gleiche Argumentation liefert aber auch: $-DF(x_*)u = DF(x_*)(-u) \geq 0$. Also folgt $DF(x_*)u \geq 0$ für alle $u \in X$ und dies impliziert $DF(x_*) = 0$. $\square$

2.5.2 Notwendige Optimalitätsbedingungen zweiter Ordnung

Satz 2.21 (Notwendige Optimalitätsbedingungen zweiter Ordnung)

Sei $M \subset X$ eine offene Menge in einem normierten Raum, sei $F: M \rightarrow \mathbb{R}$ eine zweimal stetig differenzierbare Abbildung und sei $x_* \in M$ ein lokales Minimum von F . Dann ist die zweite Ableitung $D^2F(x_*)$ positiv-semidefinit, d. h. es gilt $D^2F(x_*)(u, u) \geq 0$ für alle Richtungen $u \in X$.

BEWEIS. Wir wählen wieder eine Umgebung U von x_* wie im vorangegangenen Satz. Für jede Richtung $u \in X$ gibt es wieder ein $\delta > 0$ sodass $x_* + t u \in U$ liegt für alle $t \in [0, \delta[$. Für alle $t \in]0, \delta[$ gibt es nach dem Satz von Taylor ein $\tau(t) \in]0, t[$ so dass

$$F(x_* + t u) = F(x_*) + t DF(x_*) u + \frac{t^2}{2} D^2 F(x_* + \tau(t) u) (u, u)$$

Aus dem vorangegangenen Satz wissen wir $DF(x_*) = 0$ und wir erhalten

$$0 \leq \frac{2}{t^2} (F(x_* + t u) - F(x_*)) = D^2 F(x_* + \tau(t) u) (u, u).$$

Da $D^2 F$ stetig ist, gilt $D^2 F(x_* + \tau(t) u) (u, u) \xrightarrow{t \rightarrow 0} D^2 F(x_*) (u, u)$ und wir erhalten wie gewünscht $D^2 F(x_*) (u, u) \geq 0$ für alle Richtungen $u \in X$. $\square$

Bemerkung 2.22 Satz 2.21 ist nicht hinreichend. Als Gegenbeispiel betrachten wir die Funktionen $F_1, F_2: \mathbb{R} \rightarrow \mathbb{R}$ gegeben durch

$$F_1(x) = \begin{cases} \exp(-x^{-2}), & x \neq 0 \\ 0, & x = 0, \end{cases} \quad \text{und} \quad F_2(x) = \begin{cases} \operatorname{sgn}(x) \exp(-x^{-2}), & x \neq 0 \\ 0, & x = 0. \end{cases}$$

Beide Funktionen sind bei 0 unendlich oft differenzierbar und alle ihre Ableitungen bei 0 verschwinden: $D^k F_1(0) = 0$ und $D^k F_2(0) = 0$. Insbesondere sind bei 0 die notwendigen Bedingungen zweiter Ordnung aus Satz 2.21 erfüllt. In der Tat hat F_1 auch ein lokales Minimum bei 0, F_2 aber nicht. Noch höhere Ableitung der beiden Funktionen bei 0 zu untersuchen hilft auch nicht weiter, da ja beide Funktionen dort die gleichen Ableitungen besitzen. Was wir hieraus lernen können ist, dass wir im Allgemeinen ein wenig mehr globale Information benötigen. So werden wir z. B. später sehen, dass sogar schon die notwendigen Bedingungen erster Ordnung hinreichend sind, wenn die Zielfunktion konvex ist.

2.5.3 Hinreichende Optimalitätsbedingungen zweiter Ordnung

Satz 2.23 (Hinreichende Optimalitätsbedingungen zweiter Ordnung)

Sei $M \subset X$ eine offene Menge in einem normierten Raum, sei $F: M \rightarrow \mathbb{R}$ eine zweimal stetig differenzierbare Abbildung und sei $x_* \in M$ ein Punkt, der die folgenden Bedingungen erfüllt:

1. $DF(x_*) = 0$.
2. $D^2 F(x_*)$ ist koerziv, d. h., es gebe ein $\lambda > 0$ so dass $D^2 F(x_*) (u, u) \geq \lambda \|u\|^2$ gilt für alle $u \in X$.

Dann ist x_* ein striktes lokales Minimum von F .

BEWEIS. Diesmal wählen wir $U = B_\varepsilon(x_*)$ als Ball mit $\varepsilon > 0$ so dass $\|D^2F(y) - D^2F(x_*)\| \leq \frac{\lambda}{2}$ ist für alle $y \in U$. Dabei ist die Norm einer stetigen Bilinearformen $b: X \times X \rightarrow \mathbb{R}$ so definiert:

$$\|b\| := \sup_{u,v \in X \setminus \{0\}} \frac{|b(u,v)|}{\|u\| \|v\|}.$$

Für beliebiges $x \in U \setminus \{x_*\}$ gibt es nach dem Satz von Taylor ein y auf der Verbindungsstrecke zwischen x_* und x mit

$$F(x) = F(x_*) + \cancel{DF(x_*)(x - x_*)} + \frac{1}{2} D^2F(y)(x - x_*, x - x_*).$$

Folglich ist $y \in U$ und wir können wie folgt abschätzen:

$$\begin{aligned} F(x) &= F(x_*) + \frac{1}{2} D^2F(x_*)(x - x_*, x - x_*) + \frac{1}{2} (D^2F(x_*) - D^2F(y))(x - x_*, x - x_*) \\ &\geq F(x_*) + \frac{\lambda}{2} \|x - x_*\|^2 - \frac{\lambda}{4} \|x - x_*\|^2 = F(x_*) + \frac{\lambda}{4} \|x - x_*\|^2 > F(x_*). \end{aligned}$$

Also ist x_* ein striktes lokales Minimum von F . □

Bemerkung 2.24 Für endlichdimensionale normierte Räume ist die Koerzivität von $D^2F(x_*)$ äquivalent dazu, dass $D^2F(x_*)$ *positiv-definit* ist, d.h. $D^2F(x_*)(u, u) > 0$ für alle $u \in X \setminus \{0\}$. Das ist eine einfache Folgerung aus den Sätzen von Weierstrass und Heine-Borel.

Bemerkung 2.25 Vorsicht ist geboten bei unendlichdimensionalen normierten Raum: Dort ist Koerzivität *nicht* gleichbedeutend mit Positiv-Definitheit. Mehr noch: Auf unendlichdimensionalen Hilberträumen lässt sich aus stetigen Bilinear- und Trilinearformen eine glatte Funktion F konstruieren sodass $DF(x_*) = 0$ und $D^2F(x_*)$ positiv-definit ist, aber bei 0 *kein* lokales Minimum hat (siehe [Beispiel 2.26](#)).

Beispiel 2.26 Sei (X, g) ein unendlichdimensionaler Hilbertraum. Dann gibt es eine Folge $(y_n)_{n \in \mathbb{N}}$ in X , so dass für alle $n \in \mathbb{N}$ die Vektoren $y_1, \dots, y_n$ linear unabhängig sind. Durch das Gram-Schmidt-Orthonormalisierungsverfahren erhält man ein abzählbares *Orthonormalsystem* $(e_n)_{n \in \mathbb{N}}$ in X , d. h. es gilt $g(e_i, e_j) = \delta_{ij}$ für alle $i, j \in \mathbb{N}$. Wir betrachten die Funktion

$$F: X \rightarrow \mathbb{R} \quad F(x) = B(x, x) + g(x, x) g(x, z).$$

Dabei seien für ein $m > 1$ die Bilinearform B und der Vektor z gegeben durch

$$B := \sum_{n \in \mathbb{N}} n^{-m} (J_g(e_n) \otimes J_g(e_n)) \quad \text{und} \quad z := \sum_{n \in \mathbb{N}} n^{-1} e_n.$$

Das heißt, wir haben für $u, v \in X$: $B(u, v) = \sum_{n \in \mathbb{N}} n^{-m} g(e_n, u) \cdot g(e_n, v)$. Wegen $m > 1$ hat man mit Dreiecks- und Cauchy-Schwarz-Ungleichung

$$|B(u, v)| \leq \sum_{n \in \mathbb{N}} n^{-m} |g(e_n, u)| \cdot |g(e_n, v)| \leq \|u\| \cdot \|v\| \sum_{n \in \mathbb{N}} n^{-m},$$

weshalb B stetig ist. Dass B positiv-semidefinit ist, sieht man ebenso leicht ein:

$$B(u, u) = \sum_{n \in \mathbb{N}} n^{-m} (g(e_n, u))^2 > 0 \quad \text{für } u \in X \setminus \{0\}.$$

Wegen

$$\|z\|^2 = \sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{N}} (i)^{-1} (j)^{-1} g(e_i, e_j) = \sum_{n \in \mathbb{N}} n^{-2} < \infty$$

konvergiert die Reihe, durch die wir z definieren, tatsächlich in X . Die Funktion F ist also wohldefiniert und stetig. Die Produktregel besagt, dass F auch unendlich oft stetig differenzierbar ist und dass wir als Ableitungen erhalten:

$$\begin{aligned} DF(x) u &= B(x, u) + B(u, x) + g(u, x) g(x, z) + g(x, u) g(x, z) + g(x, x) g(u, z) \\ &= 2 B(x, u) + 2 g(u, x) g(x, z) + g(x, x) g(u, z). \\ D^2 F(x) (u, v) &= 2 B(v, u) + 2 g(u, v) g(x, z) + 2 g(u, x) g(v, z) + 2 g(x, v) g(u, z). \end{aligned}$$

Insbesondere haben wir $F(0) = 0$, $DF(0) = 0$ und $D^2 F(0) = 2 B$. Wegen $B(e_n, e_n) = n^{-m} \xrightarrow{n \rightarrow \infty} 0$ ist B aber nicht koerziv. Betrachten wir nun die Folge $(x_n)_{n \in \mathbb{N}}$ gegeben durch $x_n := -n^{-k} e_n$ mit einem beliebigen $k \in]0, m - 1[$. Dies ist eine Nullfolge und wir haben für $n > 1$:

$$F(x_n) = n^{-m} \cdot n^{-2k} - n^{-3k-1} = n^{-2k} (n^{-m} - n^{-k-1}) < 0 = F(0).$$

Also ist 0 *kein* lokales Minimum, obwohl 0 ein kritischer Punkt ist und $D^2 F(0)$ positiv-definit ist.

Bemerkung 2.27 Für allgemeine Banachräume $(X, \|\cdot\|)$ liefert [Satz 2.23](#) so gut wie keinen Mehrwert, denn Koerzivität zusammen mit endlicher Norm von $g := D^2 F(x_*)$ impliziert, dass (X, g) ein Hilbertraum sein muss und dass die durch g erzeugte Norm äquivalent zu $\|\cdot\|$ ist. Auf Banachräumen, die keine Hilbertnorm besitzen⁵, kann man also Bedingung 2 aus [Satz 2.23](#) gar nicht antreffen.

⁵Solche gibt es: Ein Beispiel ist der Raum der stetigen Funktionen $(C^0([a, b]), \|\cdot\|_\infty)$ mit der Supremumsnorm.

2.6 Konvexe Mengen

Definition 2.28 Sei X ein reeller Vektorraum. Eine Menge $C \subset X$ heißt *konvex* genau dann, wenn für alle Punkte $x_1, x_2 \in C$ auch alle Punkte ihrer Verbindungsstrecke in C liegen, also wenn

$$\{x_1(1-t) + tx_2 \mid t \in [0, 1]\} \subset C.$$

Aufgabe 2.29

Sei X ein reeller Vektorraum.

1. Eine Menge $C \subset X$ ist genau dann konvex, wenn für beliebige endlichen Teilmengen $\{x_1, \dots, x_n\} \subset C$ und beliebige $t_1, \dots, t_n \in [0, 1]$ mit $\sum_{i=1}^n t_i = 1$ auch die sogenannte *Konvexkombination* $\sum_{i=1}^n t_i x_i$ wieder in C liegt.
2. Sind $A, B \subset X$ konvex, so ist auch ihre *Minkowski-Summe* $A+B := \{a+b \mid a \in A, b \in B\}$ eine konvexe Menge.
3. Sei I eine beliebige Indexmenge und seien $C \subset X$ konvexe Mengen, $i \in I$. Dann ist auch ihr Schnitt $\bigcap_{i \in I} C_i$ eine konvexe Menge.
4. Insbesondere gibt es zu jeder Menge $A \subset X$ eine kleinste konvexe Menge $\text{conv}(A)$, die A enthält, die sogenannte *konvexe Hülle*

$$\text{conv}(A) = \bigcap_{\{C \subset X \mid C \text{ ist konvex und } A \subset C\}} C.$$

5. Ist X ein normierter Vektorraum, so kann man auch die *abgeschlossene konvexe Hülle* definieren:

$$\overline{\text{conv}}(A) = \bigcap_{\{C \subset X \mid C \text{ ist abg., konvex und } A \subset C\}} C.$$

2.6.1 Beste Approximation

Aufgabe 2.30

Sei (X, g) ein reeller Hilbertraum mit innerem Produkt $g: X \times X \rightarrow \mathbb{R}$ und Norm $\|x\| := \sqrt{g(x, x)}$. Seien $x, y \in X$.

1. Zeigen Sie durch Minimieren der Funktion $F: \mathbb{R} \rightarrow \mathbb{R}$, $F(t) = g(x + ty, x + ty)$ die Cauchy-Schwarz-Ungleichung $g(x, y) \leq \|x\| \|y\|$.
2. Weisen Sie die *Parallelogrammgleichung* $\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2$ nach und leiten Sie aus ihr die *Polarisationsformel* $g(x, y) = \frac{1}{4}(\|x+y\|^2 - \|x-y\|^2)$ her.

Proposition 2.31 (Satz über die beste Approximation)

Sei $C \subset X$ eine abgeschlossene, konvexe, nichtleere Menge im Hilbertraum (X, g) . Dann gibt es zu jedem $x \in X$ eine eindeutige Lösung $P(x)$ des Optimierungsproblems

$$\begin{array}{ll} \text{Minimiere} & \|y - x\| \\ \text{u. d. N.} & y \in C. \end{array}$$

Die Lösung $P(x)$ wird beste Approximation an x in C genannt und erfüllt

$$g(x - P(x), y - P(x)) \leq 0 \quad \text{für alle } y \in C. \quad (2.3)$$

Die Abbildung $P: X \rightarrow X$, $x \mapsto P(x)$ nennt man auch den metrischen Projektor auf C . Sie erfüllt $\|P(y) - P(x)\| \leq \|y - x\|$ für alle $x, y \in X$ und ist daher stetig.

BEWEIS. Mit $r := \inf_{y \in C} \|y - x\|$ bezeichnen wir den Abstand von x zu C . Nach der Definition des Infimums finden wir zu jedem $n \in \mathbb{N}$ ein $x_n \in C$ mit

$$\|x_n - x\|^2 \leq r^2 + \frac{1}{n}.$$

Wir wollen zeigen, dass $(x_n)_{n \in \mathbb{N}}$ eine Cauchyfolge und damit konvergent ist. Weil C konvex ist, liegt der Punkt $\frac{1}{2}(x_m + x_n)$ auch in C , und wir haben $\|\frac{1}{2}(x_m + x_n) - x\|^2 \geq r^2$. Nun schreiben wir:

$$\begin{aligned} \|x_m - x_n\|^2 &= \|(x_m - x) + (x_n - x)\|^2 + \|(x_m - x) - (x_n - x)\|^2 - 4 \|\frac{1}{2}(x_m + x_n) - x\|^2 \\ &\geq \|(x_m - x) + (x_n - x)\|^2 + \|(x_m - x) - (x_n - x)\|^2 - 4 r^2. \end{aligned}$$

Mit der Parallelogrammgleichung $\|a + b\|^2 + \|a - b\|^2 = 2\|a\|^2 + 2\|b\|^2$ für $a, b \in X$ erhalten wir

$$\begin{aligned} \|x_m - x_n\|^2 &\leq \|(x_m - x) + (x_n - x)\|^2 + \|(x_m - x) - (x_n - x)\|^2 - 4 r^2 \\ &= 2\|x_m - x\|^2 + 2\|x_n - x\|^2 - 4 r^2 \\ &\leq 4 r^2 + \frac{2}{m} + \frac{2}{n} - 4 r^2 = \frac{2}{m} + \frac{2}{n}. \end{aligned}$$

Hieran lesen wir ab, dass $(x_n)_{n \in \mathbb{N}}$ eine Cauchyfolge ist. Weil X vollständig ist, konvergiert diese Folge deshalb gegen einen Punkt x_* . Weil C abgeschlossen ist, ist auch $x_* \in C$. Außerdem hat man

$$\|x_* - x\|^2 = \lim_{n \rightarrow \infty} \|x_n - x\|^2 = \lim_{n \rightarrow \infty} (r^2 + \frac{1}{n}) = r^2 = \inf_{y \in C} \|y - x\|^2.$$

Damit haben wir also eine Lösung des Optimierungsproblems gefunden.

Als Nächstes möchten wir (2.3) nachweisen. Sei $y \in C$ beliebig und sei x_* eine Lösung des Optimierungsproblems. Wir definieren die Funktion $f: [0, 1] \rightarrow \mathbb{R}$ durch

$$f(t) = \|(x_* + t(y - x_*)) - x\|^2$$

Man beachte, dass $(x_* + t(y - x_*))$ in C liegt, weil ja C konvex ist. Also gilt $\frac{f(t)-f(0)}{t} \geq 0$ für alle $t \in [0, 1]$. Außerdem haben wir

$$f(t) = \|(x_* - x) + t(y - x_*)\|^2 = \|x_* - x\|^2 - 2t g(x - x_*, y - x_*) + t^2 \|y - x_*\|^2$$

und

$$0 \leq \frac{f(t) - f(0)}{t} = -2 g(x - x_*, y - x_*) + 2t \|y - x_*\|^2 \xrightarrow{t \rightarrow 0} -2 g(x - x_*, y - x_*).$$

Das impliziert aber gerade (2.3).

Wir zeigen nun die Eindeutigkeit der Lösung. Seien dazu $y, x_* \in C$ zwei Lösungen. Dann müssen nach (2.3) die folgenden Ungleichungen gelten:

$$g(x - x_*, y - x_*) \leq 0 \quad \text{und} \quad g(y - x, y - x_*) = g(x - y, x_* - y) \leq 0$$

und Addition dieser Ungleichungen liefert

$$\|y - x_*\|^2 = g(y - x_*, y - x_*) = g(x - x_*, y - x_*) + g(y - x, y - x_*) \leq 0.$$

Schließlich beschäftigen wir uns noch mit der Stetigkeit von P . Sind $x, y \in X$, so sind $P(x)$ und $P(y) \in C$ und wir erhalten die Ungleichungen

$$0 \leq g(x - P(x), P(y) - P(x)) \quad \text{und} \quad 0 \leq g(y - P(y), P(x) - P(y)).$$

Addition liefert

$$\begin{aligned} 0 &\leq g(x - P(x), P(y) - P(x)) + g(y - P(y), P(x) - P(y)) \\ &= g(x - y, P(y) - P(x)) - g(P(y) - P(y), P(y) - P(x)) \end{aligned}$$

und daher zusammen mit der Cauchy-Schwarz-Ungleichung

$$\begin{aligned} \|P(y) - P(x)\|^2 &= g(P(y) - P(x), P(y) - P(x)) \\ &\leq g(x - y, P(y) - P(x)) \leq \|x - y\| \|P(y) - P(x)\|. \end{aligned} \quad \square$$

Folgerung 2.32 (Orthogonale Projektion)

Sei (X, g) ein Hilbertraum und $Z \subset X$ ein abgeschlossener, linearer Unterraum. Dann ist die Abbildung $P: X \rightarrow X$, die jedem $x \in X$ ihre beste Approximation $P(x) := x_*$ aus Z zuordnet, eine stetige, lineare Abbildung und erfüllt

$$P \circ P = P, \quad P^* = P \quad \text{und} \quad g(x - P(x), z) = 0 \quad \text{für alle } z \in Z.$$

BEWEIS. Seien $x, y \in X$. Die Gleichung $P(P(x)) = P(x)$ folgt direkt aus $\|P(x) - P(x)\| = 0$. Weil Z ein linearer Unterraum ist, muss für alle $z \in Z$ gelten

$$g(x - P(x), z) = g(x - P(x), (z - P(x)) - P(x)) \leq 0 \quad \text{und} \quad (2.4)$$

$$g(x - P(x), -z) = g(x - P(x), (-z - P(x)) - P(x)) \leq 0 \quad (2.5)$$

also $g(x - P(x), z) = 0$. Linearkombination der Gleichungen $g(x - P(x), z) = 0$ und $g(y - P(y), z) = 0$ mit $s \in \mathbb{R}$ und $t \in \mathbb{R}$ liefert $g(sx + ty - (sP(x) + tP(y)), z) = 0$ für alle $z \in Z$. Dies wiederum impliziert $P(sx + ty) = sP(x) + tP(y)$ wegen der Charakterisierung in (2.3) und der Eindeutigkeit von $P(sx + ty)$. Die Gleichung $P^* = P$ erhält man aus

$$\begin{aligned} 0 &= g(x - P(x), P(y)) - g(y - P(y), P(x)) \\ &= g(x, P(y)) - \overline{g(P(x), P(y))} - g(y, P(x)) + \overline{g(P(y), P(x))} = g(x, P(y)) - g(P(x), y). \quad \square \end{aligned}$$

Folgerung 2.33 *In einem Hilbertraum X besitzt jeder abgeschlossene Unterraum $Y \subset X$ ein abgeschlossenes Komplement, d. h., es gibt einen abgeschlossenen Unterraum $Z \subset X$ so dass $X = Y \oplus Z$ gilt.*

2.6.2 Hahn-Banachscher Trennungssatz

Satz 2.34 (Hahn-Banachscher Trennungssatz)

Sei $C \subset X$ eine abgeschlossene, konvexe Menge im lokalkonvexen Raum X und $x_0 \in X$. Dann liegt der Punkt x_0 entweder in C oder er wird durch eine abgeschlossene Hyperebene von ihm getrennt, d. h. es gibt ein $\xi \in X'$ und ein $s \in \mathbb{R}$ mit $\langle \xi, x_0 \rangle > s$ und $\langle \xi, x \rangle \leq s$ für alle $x \in C$.

Dieser Satz wird üblicherweise in der Funktionalanalysis mit Hilfe von transfiniter Induktion bewiesen (für endlichdimensionale normierte Räume genügt auch die gewöhnliche vollständige Induktion). Wir begnügen uns hier mit dem Spezialfall für Hilberträume, den wir elementar aus dem Satz über die beste Approximation herleiten können; der allgemeine Fall kann zum Beispiel in [13, Theorem III.2.5] nachgelesen werden.

BEWEIS (Für den Fall, dass (X, g) ein Hilbertraum ist.). Ist $x_0 \in C$, so kann man es natürlich nicht von C trennen. Nehmen wir daher an, dass x_0 nicht in C liegt. Nach Proposition 2.31 gibt es ein $x_* = P(x_0) \in C$ mit geringstem Abstand zu x_0 . Wir definieren die lineare Abbildung $\xi: X \rightarrow \mathbb{R}$ durch $\xi = J_g(x_0 - x_*)$, also durch $\xi(x) := g(x_0 - x_*, x)$ für alle $x \in X$. Wegen der Cauchy-Schwarz Ungleichung haben wir

$$|\xi(x)| = |g(x_0 - x_*, x)| \leq \|x_0 - x_*\| \|x\|.$$

Damit ist gezeigt, dass ξ stetig ist, dass also $\xi \in X'$ gilt. Außerdem setzen wir $s := \langle \xi, x_* \rangle$. Dann haben wir einerseits

$$\langle \xi, x_0 \rangle = \langle \xi, x_* \rangle + \langle \xi, x_0 - x_* \rangle = s + g(x_0 - x_*, x_0 - x_*) = s + \|x_0 - x_*\|^2 > s,$$

weil ja $x_0 \neq x_*$ gelten muss. Andererseits haben wir wegen (2.3) für beliebiges $x \in C$ die Ungleichung

$$\langle \xi, x \rangle = \langle \xi, x_* \rangle + g(x_0 - x_*, x - x_*) = s + g(x_0 - x_*, x - x_*) \leq s. \quad \square$$

2.6.3 Darstellungssatz von Fréchet-Riesz

Satz 2.35 (Darstellungssatz von Fréchet-Riesz)

Sei (X, g) ein Hilbertraum. Dann ist die Abbildung

$$J_g: X \rightarrow X', \quad \langle J_g(x), y \rangle = g(x, y) \quad \text{für } x, y \in X$$

ein linearer, isometrischer Isomorphismus, d. h. es gilt $\|J_g x\| = \|x\|$ für alle $x \in X$.

BEWEIS. Ist $x \in X$. Einerseits hat man mit der Cauchy-Schwarz-Ungleichung

$$\|J_g x\| = \sup_{y \in X \setminus \{0\}} \frac{\langle J_g x, y \rangle}{\|y\|} = \sup_{y \in X \setminus \{0\}} \frac{g(x, y)}{\|y\|} \leq \sup_{y \in X \setminus \{0\}} \frac{\|x\| \|y\|}{\|y\|} = \|x\|.$$

Ist $x = 0$, so hat man $\|J_g x\| = 0 = \|x\|$. Andernfalls kann man so rechnen:

$$\|J_g x\| = \sup_{y \in X \setminus \{0\}} \frac{\langle J_g x, y \rangle}{\|y\|} \geq \frac{\langle J_g x, x \rangle}{\|x\|} = \frac{g(x, x)}{\|x\|} = \|x\|.$$

Also hat man $\|J_g x\| = \|x\|$ und J_g ist eine Isometrie und insbesondere injektiv.

Wir müssen noch zeigen, dass J_g surjektiv ist. Dazu sei $\xi \in X'$. Ist $\xi = 0$, so haben wir auf jeden Fall $J_g 0 = \xi$. Nehmen wir also an, dass $\xi \neq 0$. Dann gibt es ein $u \in X$ so dass $\langle \xi, u \rangle > 0$ gilt. Sei P der orthogonale Projektor auf $\ker(\xi)$. (Dieser existiert weil ξ stetig ist und $\ker(\xi)$ damit abgeschlossen ist.) Wegen $\langle \xi, u - P(u) \rangle = \langle \xi, u \rangle > 0$ ist $\|u - P(u)\| \neq 0$. Deswegen können wir $x := s(u - P(u))$ mit $s = \langle \xi, u \rangle \|u - P(u)\|^{-2}$ definieren. Dann gilt

$$g(x, x) = s^2 \|u - P(u)\|^2 = s \langle \xi, u \rangle = \langle \xi, s(u - P(u)) \rangle = \langle \xi, x \rangle.$$

Für alle $z \in \ker(\xi)$ gilt nun aber

$$g(x, z) = g(\cancel{x - P(x)}, \cancel{z}) + g(P(x), z) = s g(P(u - P(u)), z) = s g(P(u) - P(P(u)), z) = 0.$$

Jedes beliebige $y \in X$ können wir zerlegen in $y = \left(y - x \frac{\langle \xi, y \rangle}{\langle \xi, x \rangle}\right) + x \frac{\langle \xi, y \rangle}{\langle \xi, x \rangle}$. Dabei ist der erste Summand in $\ker(\xi)$. Also haben wir

$$g(x, y) = g(\cancel{x - x \frac{\langle \xi, y \rangle}{\langle \xi, x \rangle}}, \cancel{y}) + g\left(x, x \frac{\langle \xi, y \rangle}{\langle \xi, x \rangle}\right) = \frac{g(x, x)}{\langle \xi, x \rangle} \langle \xi, y \rangle = \langle \xi, y \rangle.$$

Folglich gilt $J_g(x) = \xi$. □

2.7 Konvexe Funktionen

Definition 2.36 Sei $M \subset X$ eine konvexe Menge in einem reellen Vektorraum X . Eine Funktion $F: M \rightarrow \mathbb{R}$ nennt man

1. *konvex* genau dann, wenn Sekanten stets oberhalb von M verlaufen, d. h., wenn für alle $x, y \in M$ gilt:

$$F(x)(1-t) + tF(y) \geq F(x(1-t) + ty) \quad \text{für alle } t \in [0, 1].$$

2. *strikt konvex* genau dann, wenn für alle $x, y \in M$ mit $x \neq y$ gilt:

$$F(x)(1-t) + tF(y) > F(x(1-t) + ty) \quad \text{für alle } t \in]0, 1[.$$

Die Funktion F nennt man (*strikt*) *konkav* genau dann, wenn $-F$ (*strikt*) konvex ist.

Lemma 2.37 Sei $M \subset X$ eine konvexe Menge in einem reellen Vektorraum X und $F: M \rightarrow \mathbb{R}$, eine konvexe Funktion. Dann sind alle Subniveaumengen $L_c^-(F)$, $c \in \mathbb{R}$ konvex. Insbesondere ist die Menge der Minimierer

$$\arg \min(F) := \{ x \in M \mid \text{für alle } y \in M: F(x) \leq F(y) \}$$

konvex.

BEWEIS. Sei $c \in \mathbb{R}$ und $x, y \in L_c^-(F)$. Für alle $t \in [0, 1]$ hat man dann

$$F(x(1-t) + ty) \leq F(x)(1-t) + tF(y) \leq c(1-t) + tc \leq c,$$

also auch $x(1-t) + ty \in L_c^-(F)$. Die Konvexität von $\arg \min(F)$ folgt dann aus der für $L_c^-(F)$ für $c := \inf(F)$. $\square$

Aufgabe 2.38

Sei $M \subset X$ eine konvexe Menge in einem reellen Vektorraum X , I eine Indexmenge und $F_i: M \rightarrow \mathbb{R}$, $i \in I$ konvexe Funktionen. Dann gilt:

1. Für alle $t_1, t_2 \in [0, \infty[$ ist die Funktion $t_1 F_1 + t_2 F_2$ auch konvex.
2. Die Funktion $F: U \rightarrow \mathbb{R}$, gegeben durch $U = \{ x \in M \mid \sup_{i \in I} F_i(x) < \infty \}$ und $F(x) = \sup_{i \in I} F_i(x)$ ist konvex.

Satz 2.39 Sei $M \subset X$ eine konvexe, offene Menge in einem normierten Raum X und sei $F: M \rightarrow \mathbb{R}$ eine differenzierbare Funktion. Dann sind die folgenden Bedingungen äquivalent:

1. F ist konvex.
2. Für alle $x, y \in M$ gilt $F(y) \geq F(x) + \langle DF(x), y - x \rangle$.

3. Für alle $x, y \in M$ gilt $\langle DF(y) - DF(x), y - x \rangle \geq 0$.

Ist F sogar zweimal stetig differenzierbar, so sind diese Bedingungen auch äquivalent zu

4. Für alle $x \in M$ ist $D^2F(x)$ positiv-semidefinit.

BEWEIS. Sei F differenzierbar und seien stets $x, y \in M$.

“1. $\Rightarrow$ 2.”: Angenommen, F ist konvex. Subtrahiert man $F(x)$ von der Konvexitätsbedingung

$$F(x(1-t) + ty) \leq F(x)(1-t) + tF(y)$$

und teilt durch $t \in]0, 1[$, so erhält man

$$\frac{F(x(1-t) + ty) - F(x)}{t} \leq \frac{F(x)(1-t) + tF(y) - F(x)}{t} = F(y) - F(x).$$

Im Grenzübergang $t \rightarrow 0$ ergibt das genau Aussage 2.

“2. $\Rightarrow$ 3.”: Aus Aussage 2 folgen

$$\langle -DF(x), y - x \rangle \geq F(x) - F(y) \quad \text{und} \quad \langle DF(y), y - x \rangle \geq F(y) - F(x).$$

Addieren liefert Aussage 3.

“3. $\Rightarrow$ 1.”: Wir definieren die Hilfsfunktion $\varphi: [0, 1] \rightarrow \mathbb{R}$ durch $\varphi(t) := F(z(t))$ mit $z(t) := x(1-t) + ty$. Sie hat die Ableitung $\varphi'(t) = \langle DF(x(1-t) + ty), y - x \rangle$ und es gilt für $s, t \in [0, 1]$:

$$z(s) - z(t) = (x(1-s) + sy) - (x(1-t) + ty) = (s-t)(x-y).$$

Aus

$$\begin{aligned} (\varphi'(s) - \varphi'(t))(s-t) &= \langle DF(z(s)) - DF(z(t)), (y-x)(s-t) \rangle \\ &= \langle DF(z(s)) - DF(z(t)), z(s) - z(t) \rangle \geq 0. \end{aligned}$$

folgt, dass die Ableitung $\varphi': [0, 1] \rightarrow \mathbb{R}$ monoton steigend sein muss. Nach dem Mittelwertsatz gibt es $\tau \in]0, t[$ und $\sigma \in]t, 1[$ mit

$$\frac{\varphi(t) - \varphi(0)}{t-0} = \varphi'(\tau) \leq \varphi'(\sigma) = \frac{\varphi(1) - \varphi(t)}{1-t}.$$

Durchmultiplizieren mit $t(1-t) \geq 0$ und Umordnen ergibt

$$F(x(1-t) + ty) = \varphi(t) \leq \varphi(0)(1-t) + t\varphi(1) = F(x)(1-t) + tF(y).$$

Sei F von nun an zweimal stetig differenzierbar.

“3. $\Rightarrow$ 4.”: Sei $u \in X$ und $x \in M$. Dann hat man

$$D^2F(x)(u, u) = \lim_{t \searrow 0} \frac{1}{t} \langle DF(x+tu) - DF(x), u \rangle \geq 0.$$

“4. $\Rightarrow$ 2.”: Sei $D^2F(z)$ positiv-semidefinit für alle $z \in M$. Für x und $y \in M$ liegt dann auch $x + t(y-x) \in M$ für jedes $t \in [0, 1]$; folglich ist $D^2F(x + t(y-x))$ positiv semidefinit. Nach dem Satz von Taylor gilt also

$$F(y) - F(x) - DF(x)(y-x) = \int_0^1 (1-t) D^2F(x + t(y-x))(y-x, y-x) dt \geq 0. \quad \square$$

Bemerkung 2.40 (Ausblick: Monotone Operatoren)

Bedingung 3 aus Satz 2.39 führt zu folgendem Begriff: Sei $M \subset X$ eine konvexe Menge in einem reellen Vektorraum X . Eine Abbildung $A: M \rightarrow X'$ nennt man *monoton* oder auch *monotonen Operator* genau dann, wenn für alle $x, y \in M$ die Ungleichung $\langle A(y) - A(x), y - x \rangle \geq 0$ erfüllt ist. Satz 2.39 besagt also, dass eine differenzierbare Funktion F genau dann konvex ist, wenn ihre Ableitung $DF: M \rightarrow X'$ monoton ist. Monotone Operatoren und ihre Verallgemeinerungen (pseudo-monotone Operatoren, mengenwertige monotone Operatoren) spielen eine wichtige Rolle in der Funktionalanalysis nichtlinearer Differentialgleichungen, da das Browder-Minty-Theorem recht allgemeine Bedingungen dafür liefert, dass Gleichungen der Form $A(x) = b$ für gegebenes $b \in X'$ eine Lösung besitzen.

Satz 2.41 Sei $M \subset X$ eine konvexe Menge in einem normierten Raum X und sei $F: M \rightarrow \mathbb{R}$ eine konvexe Funktion. Dann gilt

1. Jedes lokale Minimum von F ist auch ein globales Minimum.
2. Ist $x_* \in M^\circ$ ein innerer Punkt, ist F auf einer Umgebung von x_* differenzierbar und gilt $DF(x_*) = 0$, so ist x_* ein globales Minimum von F .
3. Die Menge $\arg \min(F)$ der globalen Minimierer von F ist konvex.
4. Ist F strikt konvex, so gibt es höchstens einen globalen Minimierer.

BEWEIS. Zu 1.: Sei $x \in M$ ein lokales Minimum von F und sei U eine Umgebung von x , auf der x Minimierer von $F|_U$ ist. Angenommen, es gibt ein $y \in M$ mit $F(y) < F(x)$. Für alle $t \in]0, 1[$ gilt dann

$$F(x(1-t) + ty) \leq F(x)(1-t) + tF(y) < F(x)(1-t) + tF(x) = F(x).$$

Ist aber t hinreichend klein, so hat man $x(1-t) + ty \in U$ und daher

$$F(x) \leq F(x(1-t) + ty) < F(x),$$

was zu einem Widerspruch führt.

Zu 2.: Aus Aussage 2 von Satz 2.39 folgt unmittelbar $0 \leq F(y) - F(x_*)$ für alle $y \in M$.

Zu 3.: Man hat $\arg \min(F) = \bigcap_{c > \inf(F)} L_c^-(F)$ und nach Lemma 2.37 und Aussage 3 von Aufgabe 2.29 ist das eine konvexe Menge.

Zu 4.: Seien $x, y \in M$ zwei globale Minimierer. Angenommen, es gilt $x \neq y$. Dann gilt $F(x) = F(y) = \inf(F)$ und die strikte Konvexitätsbedingungen mit $t = \frac{1}{2}$ führt zu folgendem Widerspruch:

$$\inf(F) \leq F\left(\frac{x}{2} + \frac{y}{2}\right) < \frac{F(x)}{2} + \frac{F(y)}{2} = \inf(F). \quad \square$$

Bemerkung 2.42 Satz 2.41 besagt insbesondere, dass bei offenem Zulässigkeitsbereich M und bei differenzierbarer, konvexer Zielfunktion F die notwendigen Optimalitätsbedingungen erster Ordnung aus Satz 2.20 auch schon hinreichend sind. Wir sehen also, dass recht starke globale Bedingungen für Zulässigkeitsbereich und Zielfunktion die eigentliche recht schwache Bedingung aus Satz 2.20 zu einem mächtigen Werkzeug machen können.

3 Unrestringierte Optimierung

3.1 Der allgemeine Abstiegsalgorithmus

Proposition 3.1 Sei $F: X \rightarrow \mathbb{R}$ eine differenzierbare Funktion auf dem normierten Raum X . Eine Richtung $u \in X$ heißt Abstiegsrichtung genau dann, wenn $\langle DF(x), u \rangle < 0$ gilt. Ist F stetig differenzierbar und ist u eine Abstiegsrichtung, so gilt $F(x + t u) < F(x)$ für alle hinreichend kleinen $t > 0$.

BEWEIS. Ist F stetig differenzierbar, so gilt auch $\langle DF(y), u \rangle < 0$ für alle $y \in B_\varepsilon(x)$ in einer geeigneten Umgebung von x . Der Mittelwertsatz besagt dann, dass es für jedes $t \in]0, \varepsilon/\|u\|$ ein $s \in]0, 1[$ geben muss, so dass

$$F(x + t u) - F(x) = \langle DF(x + s t u), u \rangle < 0. \quad \square$$

Analog kann man auch für kurze Zeit entlang anderer (stetig) differenzierbarer Kurven $t \mapsto \Gamma_x(t u)$ durch x in Richtung u laufen, d. h. es gilt

$$\Gamma_x(0) = x \quad \text{und} \quad \frac{d}{dt} \Gamma_x(t u)|_{t=0} = u.$$

An der Beweismethode der obigen Proposition ändert sich dadurch im Wesentlichen nichts.

Wiederholt man dieses Vorgehen iterativ mit einem Startpunkt x_0 , so erhält man fast zwangsläufig einen Algorithmus wie [Algorithmus 1](#) und eine Folge $x_{k+1} = \Gamma_{x_k}(t_k u_k)$ mit $F(x_{k+1}) < F(x_k)$. Dabei gilt es aber noch Wahlen zu treffen für

- *Abbruchkriterium* $\text{crit}(x_0, \dots, x_k, F, DF, D^2F, \dots)$ mit $\text{crit}(x_0, \dots, x_k, F, DF, \dots) = \text{true}$, falls $DF(x_k) = 0$.
- *Suchrichtungsstrategie* $u_k = p(x_0, \dots, x_k, F, DF, D^2F, \dots)$ mit $\langle DF(x_k), u_k \rangle < 0$, falls $DF(x_k) \neq 0$.
- *Schrittweitenregel* $t_k = \tau(x_0, \dots, x_k, u_k, \Gamma, F, DF, D^2F, \dots) > 0$.
- *Aufdatierungsmethode* $\Gamma \in C^1(X \times X; X)$ mit $\Gamma_x(0) = x$ und $\frac{d}{dt} \Gamma_x(t u)|_{t=0} = u$.

Freilich werden wir zunächst der Einfachheit halber $\Gamma_x(v) = x + v$ wählen. Später, in der restringierten Optimierung werden wir uns aber wieder auf diese Wahlfreiheit besinnen. Außerdem werden wir nur Schrittweitenregeln und Abbruchkriterien von der Form $\tau(x_k, u_k, F, DF, D^2F)$

Algorithm 1 Allgemeiner Abstiegsalgorithmus der unrestringierten Optimierung

```

1: function DESCENT( $x_0, F, DF, D^2F, \dots$ )
2:    $k \leftarrow 0$ 
3:   while  $\neg \text{crit}(x_0, \dots, x_k, F, DF, \dots)$  do                                 $\triangleright$  Abbruchkriterium
4:      $u_k \leftarrow p(x_0, \dots, x_k, F, DF, D^2F, \dots)$                          $\triangleright$  Wahl einer Suchrichtung
5:      $t_k \leftarrow \tau(x_0, \dots, x_k, u_k, F, DF, D^2F, \dots)$                    $\triangleright$  Schrittweitenregel
6:      $x_{k+1} \leftarrow \Gamma_{x_k}(t_k u_k)$                                            $\triangleright$  Aufdatierung
7:      $k \leftarrow k + 1$ 
8:   end while
9:   return  $x_k$ 
10: end function

```

bzw. $\text{crit}(x_k, DF)$ betrachten, das heißt bei uns hängen diese nur vom aktuellen Ort und der gegebenen Suchrichtung ab. Bei den Suchrichtungsstrategien werden wir aber auch Verfahren der Form $\tau(x_0, \dots, x_k, u_k, F, DF)$ kennenlernen, die also Nutzen aus der kompletten Suchhistorie $x_0, \dots, x_k$ ziehen.

Als Abbruchkriterium nimmt man in der Regel etwas von der Form

$$\text{crit}(x, DF) = \begin{cases} \text{true}, & \|DF(x)\| < \text{TOL}, \\ \text{false}, & \|DF(x)\| \geq \text{TOL} \end{cases} \quad (3.1)$$

mit einer vom Nutzer zu wählenden Fehlertoleranz $\text{TOL} > 0$. In der Praxis wird TOL etliche Größenordnungen über der Maschinengenauigkeit liegen, damit [Algorithmus 1](#) überhaupt terminieren kann. In wieweit die *Residualnorm* $\|DF(x_k)\|$ ein sinnvolles Abbruchkriterium liefert (letztlich ist man ja eher am Abstand der letzten Iterierten x_k zu einer Lösung interessiert), werden wir später erörtern (siehe [Satz 3.29](#) und [Bemerkung 3.30](#)). Oft wird man den Abstiegsalgorithmus beenden, wenn die Iterierten sich kaum noch ändern ($\|x_{k+1} - x_k\|$ ist kleiner als eine vorzugegebene Toleranz) oder wenn keine Konvergenz mehr erwartet wird (die anderen Abbruchkriterien werden auch nach “vielen” Iterationen nicht erfüllt; die Entscheidung, was “viel” bedeuten soll, wird dann in der Regel dem Nutzer überlassen).

Konkrete Strategien für Suchrichtungen und Schrittweiten werden wir in den folgenden Abschnitten behandeln. Zunächst untersuchen wir einige “Gütekriterien” für Abstiegsverfahren.

Definition 3.2 Ein Tripel (p, τ, Γ) bestehend aus einer Auswahlstrategie für Suchrichtungen p , einer Schrittweitenstrategie τ und einer Aufdatierungsmethode Γ nennen wir *Suchstrategie*. Für Suchstrategien (p, τ, Γ) und eine Menge von Startwerten $K \subset M$ führen wir folgende Qualitätsmerkmale ein (hierbei sind x_k und u_k stets die von Algorithmus 1 erzeugten Folgen zu einem beliebigem Startwert x_0):

1. *Winkelbedingung* mit Konstante $\alpha > 0$:

$$-\frac{\langle DF(x_k), u_k \rangle}{\|DF(x_k)\| \|u_k\|} \geq \alpha \quad \text{für alle } x_0 \in K \text{ und } k \in \mathbb{N}. \quad (3.2)$$

2. *Geschwindigkeitsbedingung* mit Konstante $\beta > 0$:

$$\|u_k\| \geq \beta \|DF(x_k)\| \quad \text{für alle } x_0 \in K \text{ und } k \in \mathbb{N}. \quad (3.3)$$

3. *Effizienzbedingung*¹ mit Konstante $\vartheta > 0$:

$$F(x_{k+1}) \leq F(x_k) - \vartheta \min\left(-\langle DF(x_k), u_k \rangle, \frac{\langle DF(x_k), u_k \rangle^2}{\|u_k\|^2}\right) \quad \text{für alle } x_0 \in K \text{ und } k \in \mathbb{N}. \quad (3.4)$$

Für $u_k \neq 0$ bedeutet die Winkelbedingung gerade, dass der Einheitsvektor $\frac{u_k}{\|u_k\|}$ dem Kern von $DF(x_k)$ im Laufe des Abstiegsalgorithmus nicht zu nahe kommt, die Suchrichtung mit dem Kern also einen nicht zu kleinen “Winkel” einschließt. Andernfalls sucht man mit der Zeit eher entlang einer Niveaumenge $\{x \in X \mid F(x) = c\}$ als transversal dazu; entlang der Niveaumenge ist aber so gut wie nichts “zu holen”. Die Geschwindigkeitsbedingung stellt lediglich sicher, dass die Abstiegsrichtungen u_k nicht zu kurz werden (dies ist natürlich immer durch geeignete Umskalierung zu erreichen). Die Effizienzbedingung schließlich soll für einen hinreichenden Abstieg im Funktionswert sorgen.

Bei einfacheren Suchstrategien hängt die Suchrichtung nur von der aktuellen Iterierten x_k und der Ableitungen $DF(x_k)$, $D^2F(x_k)$ der Funktion an dieser Stelle ab. Die Winkel- und Geschwindigkeitsbedingungen sind dann einzig und allein Bedingungen an die Suchrichtungswahl p . In der Regel kann man dann die Effizienz durch eine geeignete Schrittweitenregel erreichen. Mehr dazu in Abschnitt 3.6. Alle drei Bedingungen zusammen ergeben jedenfalls:

Lemma 3.3 Sei $F: X \rightarrow \mathbb{R}$ differenzierbar. Die Suchstrategie (p, τ, Γ) erfülle die Winkel-, Geschwindigkeits- und Effizienzbedingung mit den Konstanten $\alpha, \beta, \vartheta > 0$ auf der Menge $K \subset X$. Dann gilt für alle $x_0 \in K$:

$$F(x_{k+1}) \leq F(x_k) - \vartheta \alpha \min(\alpha, \beta) \|DF(x_k)\|^2.$$

¹Üblicherweise wird eine Suchstrategie *effizient* genannt, wenn $F(x_{k+1}) \leq F(x_k) - \vartheta \frac{\langle DF(x_k), u_k \rangle^2}{\|u_k\|^2}$ gilt. Manche Autoren nennen eine Suchstrategie *semi-effizient*, wenn sie die hier definierte Effizienzbedingung erfüllt.

BEWEIS. Wir setzen nacheinander die Winkel- und die Geschwindigkeitsbedingung in die Effizienzbedingung ein, so erhalten wir

$$\begin{aligned}
 F(x_{k+1}) &\leq F(x_k) - \vartheta \min\left(-\langle DF(x_k), u_k \rangle, \frac{\langle DF(x_k), u_k \rangle^2}{\|u_k\|^2}\right) \\
 &\leq F(x_k) - \vartheta \min\left(\alpha \|DF(x_k)\| \|u_k\|, \alpha^2 \|DF(x_k)\|^2\right) \\
 &\leq F(x_k) - \vartheta \min\left(\alpha \beta \|DF(x_k)\|^2, \alpha^2 \|DF(x_k)\|^2\right) \\
 &= F(x_k) - \vartheta \alpha \min(\beta, \alpha) \|DF(x_k)\|^2. \quad \square
 \end{aligned}$$

Satz 3.4 (Globale Konvergenz)

Sei $F: X \rightarrow \mathbb{R}$ stetig differenzierbar. Die Suchstrategie (p, τ, Γ) erfülle je eine Winkel-, Geschwindigkeits- und Effizienzbedingung auf der Menge K . Mit dem Abbruchkriterium

$$\text{crit}(x, DF) = \begin{cases} \text{true}, & DF(x) = 0, \\ \text{false}, & DF(x) \neq 0, \end{cases}$$

gibt es für [Algorithmus 1](#) mit Startwert $x_0 \in K$ nur die folgenden drei Möglichkeiten:

1. Abbruch nach endlich vielen Iterationen mit $DF(x_k) = 0$.
2. Die Folge $(F(x_k))_{k \in \mathbb{N}}$ konvergiert gegen $-\infty$.
3. Es gilt $\|DF(x_k)\| \rightarrow 0$ für $k \rightarrow \infty$ und jeder Häufungspunkt von $(x_k)_{k \in \mathbb{N}}$ ist ein kritischer Punkt von F .

BEWEIS. Angenommen, der Algorithmus bricht nicht ab. Wegen [Lemma 3.3](#) ist die Folge der Funktionswerte monoton fallend:

$$F(x_{k+1}) \leq F(x_k) - \vartheta \alpha \min(\alpha, \beta) \|DF(x_k)\|^2 < F(x_k).$$

Nun kann die Folge der Funktionswerte *entweder* uneigentlich gegen $-\infty$ konvergieren *oder* aber, sie ist von unten beschränkt. In letzterem Fall ist sie nach dem Monotoniekriterium aber auch konvergent und man hat

$$\|DF(x_k)\|^2 \leq \frac{F(x_k) - F(x_{k+1})}{\vartheta \alpha \min(\alpha, \beta)} \xrightarrow{k \rightarrow \infty} 0. \quad \square$$

Folgerung 3.5 Ist F zusätzlich zu den Bedingungen aus [Satz 3.4](#) nach unten beschränkt und wird die Abbruchbedingung (3.1) verwendet, so terminiert [Algorithmus 1](#) nach endlich vielen Iteration mit $\|DF(x_k)\| \leq \text{TOL}$.

Der Terminus “Globale Konvergenz” verspricht ein wenig mehr, als er halten kann. Schließlich ist Konvergenz von $(DF(x_k))_{k \in \mathbb{N}}$ in der Regel nur notwendig, nicht hinreichend für Konvergenz der Iterierten $(x_k)_{k \in \mathbb{N}}$ gegen einen kritischen Punkt. Aber zumindest für gewisse strikt konvexe Funktionen, deren Konvexität wir quantifizieren können, führt der Satz über Globale Konvergenzsatz auf eine “echte” globale Konvergenz:

Folgerung 3.6 Sei F zusätzlich zu den Bedingungen aus [Satz 3.4](#) nach unten beschränkt und erfülle die folgende quantifizierte Version von Aussage 3. von [Satz 2.39](#): Es gibt eine streng monoton steigende Funktion $\psi: [0, \infty[\rightarrow [0, \infty[$ mit $\psi(0) = \lim_{s \searrow 0} \psi(s) = 0$, so dass

$$\langle DF(y) - DF(x), y - x \rangle \geq \psi(\|y - x\|) \|y - x\| \quad \text{gilt für alle } x, y \in X.$$

Ist dann x_* ein kritischer Punkt von F (und damit also ein globales Minimum), so gilt

$$\psi(\|x_k - x_*\|) \|x_k - x_*\| \leq \langle DF(x_k) - \underline{DF}(x_*), x_k - x_* \rangle \leq \|x_k - x_*\| \|DF(x_k)\|,$$

also

$$\psi(\|x_k - x_*\|) \leq \|DF(x_k)\|.$$

Da ψ streng monoton steigend ist, besitzt es eine stetige Inverse und wir erhalten

$$\|x_k - x_*\| \leq \psi^{-1}(\|DF(x_k)\|) \xrightarrow{k \rightarrow \infty} 0.$$

3.2 Richtung des steilsten Abstiegs

Die plausibelsten Abstiegsrichtungen sind wohl die des *steilsten* Abstiegs.

Definition 3.7 Sei $(X, \|\cdot\|)$ ein normierter Raum, $M \subset X$ eine offene Menge und $F: M \rightarrow \mathbb{R}$ eine differenzierbare Funktion. Einen Vektor $u \in X$ nennt man *Richtung des steilsten Anstiegs* (bzw. des steilsten Abstiegs) am Punkt $x \in M$ genau dann, wenn er eine Lösung des folgenden Optimierungsproblems ist:

$$\begin{array}{ll} \text{Maximiere} & \langle DF(x), v \rangle \\ \text{u. d. N.} & \|v\| = \|DF(x)\| \end{array} \quad \text{bzw.} \quad \begin{array}{ll} \text{Minimiere} & \langle DF(x), v \rangle \\ \text{u. d. N.} & \|v\| = \|DF(x)\|. \end{array} \quad (3.5)$$

Ist die Richtung u des steilsten Anstiegs eindeutig, so bezeichnet man sie auch als *Gradienten* von F an der Stelle x und schreibt $\text{grad}(F)(x) := u$. Die Richtung des steilsten Abstiegs ist dann $-\text{grad}(F)(x)$.

Bemerkung 3.8 *Per definitionem* erfüllen Richtungen des steilsten Abstiegs immer die Geschwindigkeitsbedingung mit $\beta = 1$.

Lemma 3.9 Sei $(X, \|\cdot\|)$ ein normierter Raum, $M \subset X$ eine offene Menge und $F: M \rightarrow \mathbb{R}$ eine differenzierbare Funktion. Dann gilt:

1. Ist X endlichdimensional, so gibt es mindestens eine Richtung des steilsten Abstiegs.
2. Ist $\|\cdot\|$ strikt konvex², so gibt es höchstens eine Richtung des steilsten Abstiegs.

BEWEIS. Sei X endlichdimensional und $x \in M$. Wir wählen eine Minimalfolge u_k mit $\|u_k\| = \|DF(x)\|$ und $\lim_{k \rightarrow \infty} \langle DF(x), u_k \rangle = \inf_{\|v\|=\|DF(x)\|} \langle DF(x), v \rangle$. Nach dem Satz von Weierstrass gibt es eine konvergente Teilfolge und weil $DF(x)$ stetig ist, ist ihr Grenzwert eine Richtung des steilsten Abstiegs.

Sei nun $(X, \|\cdot\|)$ ein normierter Raum mit strikt konvexer Norm. Im Falle $DF(x) = 0$ ist klar, dass $u = 0$ die einzige Richtung des steilsten Abstiegs ist. Im Falle $DF(x) \neq 0$ seien u und v zwei Richtungen des steilsten Abstiegs. Dann ist der Vektor $\frac{1}{2}(u + v)$ kürzer als u und v :

$$\left\| \frac{1}{2}u + \frac{1}{2}v \right\| < \frac{1}{2}\|u\| + \frac{1}{2}\|v\| = \|DF(x)\|.$$

Der Vektor $w := \frac{s}{2}(u + v)$ mit $s := 2 \frac{\|DF(x)\|}{\|u+v\|} > 1$ hat dann gerade die richtige Länge und erfüllt wegen $\langle DF(x), u \rangle = \langle DF(x), v \rangle < 0$ die Ungleichung

$$\langle DF(x), w \rangle = \frac{s}{2} \langle DF(x), u + v \rangle = s \langle DF(x), u \rangle < \langle DF(x), u \rangle.$$

Dies ist aber ein *Widerspruch* dazu, dass u und v schon Richtungen des steilsten Abstiegs sind. $\square$

Bemerkung 3.10 Mit dem Satz von Alaoglu aus der Funktionalanalysis kann man zeigen, dass es in reflexiven Räumen³ auch immer eine Richtung des steilsten Abstiegs gibt. Für normierte

²Eine Norm heißt *strikt konvex*, wenn für $x \neq y$ mit $\|x\| = \|y\|$ stets gilt: $\|x(1-t) + ty\| < \|x\|(1-t) + t\|y\|$ für alle $t \in]0, 1[$. Als Funktion kann eine Norm nie strikt konvex sein, denn entlang vom Nullpunkt ausgehender Strahlen sind Normen ja linear.

³Wie schon in [Abschnitt 2.1](#) erwähnt ist ein Banachraum X reflexiv genau dann, wenn die kanonische Abbildung in den Bidualraum $\psi_X: X \rightarrow X''$ ein Isomorphismus ist.

Räume, die keine Banachräume sind, sollte man die Existenz einer Richtung des steilsten Abstiegs aber nicht erwarten.

Auch wenn sie existieren sind Richtungen des steilsten Abstiegs in allgemeinen Banachräumen nicht so einfach zu bestimmen. Zwar kann man das Problem (3.5) auf das *konvexe* Optimierungsproblem

$$\begin{aligned} &\text{Minimiere} \quad \langle DF(x), v \rangle \\ &\text{u. d. N.} \quad \|v\| \leq \|DF(x)\| \end{aligned}$$

zurückführen, aber dieses braucht nicht einfacher zu sein als das eigentlich zu lösende Optimierungsproblem (das Minimieren von F).

Mehr Glück hat man in Hilberträumen: Dort wird die Richtung des steilsten An-/Abstiegs immerhin durch eine lineare Gleichung gegeben.

Satz 3.11 *Sei (X, g) ein Hilbertraum, $M \subset X$ eine offene Menge und $F: M \rightarrow \mathbb{R}$ eine differenzierbare Funktion. Dann gibt es immer genau eine Richtung des steilsten Anstiegs und sie ist gegeben durch $\text{grad}(F)(x) = J_g^{-1} DF(x)$. Hierbei ist $J_g: X \rightarrow X'$ der Riesz-Isomorphismus d. h., es gilt*

$$g(\text{grad}(F)(x), v) = \langle DF(x), v \rangle \quad \text{für alle } v \in X.$$

Außerdem erfüllt die Richtung des steilsten Abstiegs in Hilberträumen immer die Winkelbedingung mit $\alpha = 1$.

BEWEIS. Nach Satz 2.35 ist J_g eine Isometrie, d. h., wir haben für $u := J_g^{-1} DF(x)$ schon $\|u\| = \|DF(x)\|$. Andererseits haben wir

$$\langle DF(x), v \rangle \leq \|DF(x)\| \|v\| = \|DF(x)\|^2 \quad \text{für alle } v \in X \text{ mit } \|v\| = \|DF(x)\|.$$

Wegen $\langle DF(x), u \rangle = g(u, u) = \|u\|^2 = \|DF(x)\|^2$ muss dann u schon eine Richtung des steilsten Anstiegs sein. Da die Norm in Hilberträumen strikt konvex ist, gibt es nur eine solche Richtung und man hat $\text{grad}(F)(x) = u$. Wegen $\|u\| = \|DF(x)\|$ erfüllt die Suchrichtung $\text{grad}(F)(x)$ die Winkelbedingung mit $\alpha = 1$. Man hat sogar

$$g(\text{grad}(F)(x), v) = \langle DF(x), v \rangle = 0, \quad \text{für alle } v \in \ker(DF(x)),$$

d. h., $\text{grad}(F)(x)$ ist senkrecht zum Kern von $\ker(DF(x))$, schließt mit diesem also einen Winkel von 90° ein. $\square$

Für numerische Zwecke ist es aber nicht unbedingt notwendig (und möglich), *exakte* Richtungen des steilsten Abstiegs zu finden; approximative Suchrichtungen sind vollkommen ausreichend, solange die Winkelbedingung erfüllt bleibt. Insbesondere für Optimierungsprobleme in hochdimensionalen Euklidischen Räumen wird man unter Umständen iterative Methoden für das Lösen der linearen Gleichungssysteme

$$J_g u_k = -DF(x_k)$$

verwenden. Bei diesen kann man in der Regel Rechenzeit direkt gegen Genauigkeit einhandeln und so bessere Laufzeiten als mit exakten Suchrichtungen erzielen, auch wenn mehr Iterationen von Algorithmus 1 benötigt werden.

3.3 Die Methode des steilsten Abstiegs für konvexe quadratische Optimierungsprobleme

Nimmt man im allgemeinen Abstiegsalgorithmus den Gradienten als Abstiegsrichtung, so erhält man die *Methode des steilsten Abstiegs* (engl. *method of steepest descent*); oft bezeichnet man dieses Verfahren etwas unsauber auch einfach als das *Gradientenverfahren*. Dieses Verfahren ist in der Regel sehr einfach zu implementieren. Man könnte aufgrund des Namens (“steilster Abstieg”) auf die Idee kommen, das Verfahren konvergiere besonders schnell gegen ein lokales Minimum. Dem ist überraschenderweise aber nicht so. Im Gegenteil: Die Methode des steilsten Abstiegs hat den Ruf, in der Praxis auch mit guten Schrittweitenregeln recht langsam zu sein. Dem wollen wir hier ein wenig auf den Grund gehen, indem wir ein vereinfachtes Modellproblem betrachten.

Während dieses Abschnitts seien (X, g) ein reeller Hilbertraum und $a: X \times X \rightarrow \mathbb{R}$ eine stetige Bilinearform. Hierzu definieren wir die eine stetige, lineare Abbildung

$$A: X \rightarrow X, \quad \text{durch} \quad g(Ax, y) := a(x, y).$$

Umgekehrt erhalten wir zu jeder stetigen, linearen Abbildung $A: X \rightarrow X$ eine stetige Bilinearform $a: X \times X \rightarrow \mathbb{R}$, so dass obige Gleichung erfüllt ist.

Nehmen wir an, dass a symmetrisch ist. Dann ist A bezüglich g *selbstadjungiert*, d.h.,

$$A^{*g} = A \quad \text{oder auch} \quad g(Au, v) = g(u, Av) \quad \text{für alle } u, v \in X.$$

Wir nennen A *koerziv* genau dann, wenn die dazugehörige Bilinearform a koerziv ist, d. h. wenn ein $\lambda > 0$ gibt, so dass

$$a(x, x) = g(x, Ax) \geq \lambda g(x, x) \quad \text{gilt für alle } x \in X.$$

Da A auch stetig ist, gibt es zudem ein Λ mit $\lambda \leq \Lambda < \infty$ so dass gilt:

$$\lambda g(x, x) \leq a(x, x) = g(x, Ax) \leq \Lambda g(x, x) \quad \text{für alle } x \in X.$$

Es ist nun nicht schwer einzusehen, dass a ebenfalls ein Skalarprodukt auf X ist, und dass dessen Norm $\|\cdot\|_a$ äquivalent ist zu $\|\cdot\|_g$ und somit die gleiche Topologie erzeugt. Wir beobachten außerdem, dass

$$Ax = J_g^{-1} J_a x \quad \text{für alle } x \in X$$

gilt, wobei J_g und J_a die jeweiligen Riesz-Abbildungen zu g und a sind.

Definition 3.12 Zu einer stetigen, symmetrischen Bilinearform $a: X \times X \rightarrow \mathbb{R}$ im Hilbertraum (X, g) und der dazugehörigen selbstadjungierten Abbildung $A = J_g^{-1} J_a$ definieren wir

$$\begin{aligned} \lambda(A, g) &:= \lambda(a, g) := \sup \{ \lambda \in \mathbb{R} \mid \text{für alle } x \in X: a(x, x) \geq \lambda g(x, x) \} \quad \text{und} \\ \Lambda(A, g) &:= \Lambda(a, g) := \inf \{ \Lambda \in \mathbb{R} \mid \text{für alle } x \in X: a(x, x) \leq \Lambda g(x, x) \}. \end{aligned}$$

Offensichtlich gilt stets $\lambda(a, g) \leq \Lambda(a, g)$. Die Zahl

$$\kappa(a, g) := \frac{\Lambda(a, g)}{\lambda(a, g)} \geq 1$$

nennt man die *Konditionszahl* von a bezüglich g .

Wir werden bald sehen, dass die Konditionszahl gewissermaßen die Schwierigkeit bemisst, für vorgegebenes $\beta \in X'$ die lineare Gleichung

$$J_a x = \beta \quad \text{oder äquivalent} \quad A x = b \quad \text{mit } b := J_g^{-1} \beta$$

durch g -gradientenbasierte Verfahren zu lösen.

Aufgabe 3.13 (Satz von Lax-Milgram)

Sei (X, g) ein Hilbertraum und sei $A \in L(X; X)$ selbstadjungiert mit $0 < \lambda(A) \leq \Lambda(A) < \infty$. Dann ist A stetig invertierbar und es gilt für die Operatornormen, dass

$$\|A\| = \Lambda(A, g) \quad \text{and} \quad \|A^{-1}\| = \frac{1}{\lambda(A, g)}.$$

Dazu gehe man wie folgt vor:

1. Man zeige, dass $\|A x\| \geq \lambda(A, g) \|x\|$ für alle $x \in X$ gilt.
2. Man folgere hieraus, dass A injektiv ist.
3. Man zeige, dass $\text{im}(A)^\perp = \{0\}$ gilt, wobei $\text{im}(A) := \{A x \mid x \in X\}$ und folgere daraus $\overline{\text{im}(A)} = (\text{im}(A)^\perp)^\perp = X$, so dass A ein dichtes Bild in X hat.
4. Schließlich benutze man 1., um zu zeigen, dass jede Cauchyfolge in $\text{im}(A)$ konvergiert gegen ein Element von $\text{im}(A)$ und schließe hieraus, dass $\text{im}(A)$ bereits abgeschlossen sein und A somit surjektiv sein muss.

Sei von nun an a eine stetige, symmetrische und koerzive Bilinearform auf (X, g) und sei $A = J_g^{-1} J_a$ der dazugehörige stetige und selbstadjungierte Operator. Weiters fixieren wir $\beta \in X'$ und den Vektor $b := J_g^{-1} \beta \in X$ und studieren das folgende quadratische, konvexe, unrestringierte Optimierungsproblem:

$$\text{Minimiere} \quad F(x) := \frac{1}{2} a(x, x) - \langle \beta, x \rangle = \frac{1}{2} g(A x, x) - g(b, x) \quad \text{über alle } x \in X. \quad (3.6)$$

Man rechnet leicht nach, dass F zweimal stetig differenzierbar ist und dass die ersten zwei Ableitungen gegeben sind durch

$$DF(x) u = \frac{1}{2} g(A u, x) + \frac{1}{2} g(A x, u) - g(b, u) = g(A x - b, u) \quad \text{für alle } u \in X \quad (3.7)$$

und

$$D^2F(x)(u, v) = (DDF(x)u)v = g(Av, u) = a(u, v) \quad \text{für alle } u, v \in X.$$

Die Koerzivität von A garantiert, dass auch $D^2F(x) = a$ koerziv ist. Insbesondere heißt das, dass F gleichmäßig konvex ist. Also ist x ein globaler Minimierer genau dann, wenn $DF(x) = 0$ gilt. Außerdem sehen wir an (3.7), dass $DF(x) = J_g(Ax - b)$ gilt. Und da der Riesz-Isomorphismus stetig invertierbar ist, ist x_* ein globaler Minimierer genau dann, wenn es die lineare Gleichung

$$Ax_* = b$$

erfüllt. Das heißt, wir können das Optimierungsproblem lösen, indem wir diese lineare Gleichung lösen und umgekehrt! Und in der Tat verwendet man ähnliche Verfahren wie das des steilsten Abstiegs im wissenschaftlichen Rechnen, um große lineare Gleichungssysteme mit symmetrischer, positiv-definiter Systemmatrix zu lösen. Ein Grund mehr also, dieses Verfahren genauer zu analysieren!

Wir möchten die Methode des steilsten Abstiegs mit *exakten* Schrittweiten benutzen. Das heißt, wir verwenden das allgemeine Abstiegsverfahren mit irgendeinem Startpunkt $x_0 \in X$ der Suchrichtung

$$u_k := -\text{grad}(F)(x_k) = b - Ax_k.$$

Mit einer geeigneten Schrittweite t_k erhalten wir die sogenannte *Richardson-Iteration*

$$x_{k+1} = x_k + t_k(b - Ax_k). \quad (3.8)$$

Die Schrittweite t_k wählen wir so, dass sie die univariate Funktion

$$\varphi(t) := F(x_k + t u_k)$$

minimiert. Dies ist eine quadratische, konvexe Funktion in t und wir finden das Minimum durch Lösen von

$$\begin{aligned} 0 &\stackrel{!}{=} \varphi'(t) = DF(x_k + t u_k) u_k = g(u_k, A(x_k + t u_k) - b) \\ &= g(u_k, Ax_k - b) + t g(u_k, A u_k) = -g(u_k, u_k) + t g(u_k, A u_k). \end{aligned}$$

Daher haben wir

$$t_k := \frac{g(u_k, u_k)}{g(u_k, A u_k)}, \quad (3.9)$$

so dass wir folgende Iterationsformel erhalten:

$$x_{k+1} = x_k + \frac{g(u_k, u_k)}{g(u_k, A u_k)} (b - Ax_k). \quad (3.10)$$

3.3.1 Konvergenzrate

Satz 3.14 Die Iterierten x_k des Verfahrens des steilsten Abstiegs mit exakter Schrittweite (siehe (3.10)) konvergieren gegen das Minimum x_* mit folgender Konvergenzrate:

$$\|x_k - x_*\| \leq \sqrt{\kappa} \left(\frac{\kappa - 1}{\kappa + 1} \right)^k \|x_0 - x_*\|,$$

wobei $\kappa := \kappa(A, g)$ die Konditionszahl von A bezüglich g ist.

BEWEIS. Die zu A gehörende symmetrische Bilinearform a ist selbst auch wieder ein Skalarprodukt. Die dazugehörige Norm bezeichnen wir mit

$$\|x\|_a := \sqrt{a(x, x)} = \sqrt{g(Ax, x)}.$$

Mit $x_{k+1} = x_k + t_k u_k$ und $Ax_* = b$ bekommen wir

$$\begin{aligned} \|x_{k+1} - x_*\|_a^2 &= g(A(x_{k+1} - x_*), x_{k+1} - x_*) \\ &= g(A(x_k - x_*), x_k - x_*) + g(A(x_k - x_*), u_k) t_k + g(A u_k, x_k - x_*) t_k + g(A u_k, u_k) t_k^2 \\ &= \|x_k - x_*\|_a^2 + 2 g(A(x_k - x_*), u_k) t_k + g(A u_k, u_k) t_k^2 \\ &= \|x_k - x_*\|_a^2 - 2 g(b - A x_k, u_k) t_k + g(A u_k, u_k) t_k^2. \end{aligned}$$

Setzen wir nun

$$u_k = b - A x_k \quad \text{und} \quad t_k = \frac{g(u_k, u_k)}{g(u_k, A u_k)}$$

ein, so bekommen wir

$$\begin{aligned} \|x_{k+1} - x_*\|_a^2 &= \|x_k - x_*\|_a^2 - 2 g(u_k, u_k) \frac{g(u_k, u_k)}{g(A u_k, u_k)} + g(A u_k, u_k) \frac{g(u_k, u_k)^2}{g(A u_k, u_k)^2} \\ &= \|x_k - x_*\|_a^2 - \frac{g(u_k, u_k)^2}{g(A u_k, u_k)} \\ &= \|x_k - x_*\|_a^2 \left(1 - \frac{g(u_k, u_k)^2}{g(A u_k, u_k) \|x_k - x_*\|_a^2} \right). \end{aligned}$$

Den Term $\|x_k - x_*\|_a^2$ im Nenner ersetzen wir durch

$$\|x_k - x_*\|_a^2 = g(A^{-1} A(x_k - x_*), A(x_k - x_*)) = g(A^{-1}(A x_k - b), (A x_k - b)) = g(A^{-1} u_k, u_k)$$

und erhalten

$$\|x_{k+1} - x_*\|_a^2 = \|x_k - x_*\|_a^2 \left(1 - \frac{g(u_k, u_k)^2}{g(A u_k, u_k) g(A^{-1} u_k, u_k)} \right).$$

Der Knackpunkt hier ist, dass der Ausdruck

$$f(u) := \frac{g(u, u)^2}{g(A u, u) g(A^{-1} u, u)}$$

invariant unter Skalierung ist, d. h., es gilt $f(t u) = f(u)$ für alle $t \in \mathbb{R} \setminus \{0\}$. Offensichtlich gilt stets $f(u) > 0$. Wegen der Skalierungsinvarianz hat man deshalb eine Chance, eine gleichmäßige untere Schranke für $f(u)$ zu finden. In der Tat bekommen wir wegen

$$g(A u, u) g(A^{-1} u, u) \leq \Lambda(A, g) \Lambda(A^{-1}, g) \|u\|^4 = \Lambda(A, g) / \lambda(A, g) \|u\|^4 = \kappa \|u\|^4.$$

ohne große Mühen bereits

$$f(u) \geq \frac{1}{\kappa}.$$

Oben eingesetzt führt uns dies zu

$$\|x_{k+1} - x_*\|_a \leq q \|x_k - x_*\|_a \quad \text{mit} \quad q = \sqrt{1 - \frac{1}{\kappa}} < 1.$$

Per Induktion über k bekommen wir

$$\|x_k - x_*\|_a \leq q^{k+1} \|x_0 - x_*\|_a$$

und mit $\|x_k - x_*\|_g \leq \Lambda^{-1} \|x_k - x_*\|_a$ und $\|x_0 - x_*\|_a \leq \Lambda \|x_0 - x_*\|_g$ führt dies zu

$$\|x_k - x_*\|_g \leq \kappa q^k \|x_0 - x_*\|_g.$$

Weil $0 \leq q < 1$ gilt, folgt damit $x_k \xrightarrow{k \rightarrow \infty} x_*$. Weil die Distanz zum Grenzwert in jedem Schritt um den Faktor $q < 1$ verbessert wird, nennt man diese Art der Konvergenz auch *q-linear*.⁴ Das q kann man noch zu

$$q = \frac{\kappa - 1}{\kappa + 1}$$

verbessern, denn mit der Kantorovich-Ungleichung [Lemma 3.15](#) folgt

$$1 - \frac{g(u, u)^2}{g(A u, u) g(A^{-1} u, u)} \leq 1 - \frac{4 \lambda(A, g) \Lambda(A, g)}{(\lambda(A, g) + \Lambda(A, g))^2} = \frac{(\kappa - 1)^2}{(\kappa + 1)^2}. \quad \square$$

Wir sehen also, dass die Methode des steilsten Abstiegs *exponentiell* im Sinne von

$$\|x_k - x_*\| = \sqrt{\kappa} q^k \|x_0 - x_*\|$$

mit einem $0 < q < 1$ gegen das Minimum konvergiert. Allerdings haben wir

$$q = \frac{\kappa - 1}{\kappa + 1}.$$

Wenn die Konditionszahl κ von A sehr groß ist, dann liegt q also sehr nahe an 1. Es kann in der Praxis daher sehr lange dauern, bis man das Minimum mit erwünschter Genauigkeit gefunden hat.

⁴Mit dem Newtonverfahren werden wir bald ein Beispiel für ein *q-quadratisch* konvergentes Verfahren sehen, also ein Verfahren mit $\|x_{k+1} - x_*\|_a \leq q \|x_k - x_*\|_a^2$.

3.3.2 Kantorovich-Ungleichung

Lemma 3.15 (Kantorovich-Ungleichung) Sei (X, g) ein reeller Hilbertraum und sei $A \in L(X; X)$ selbstadjungiert und koerziv. Dann gilt für alle $x \in X \setminus \{0\}$, dass

$$1 \leq \frac{g(Ax, x) g(A^{-1}x, x)}{g(x, x)^2} \leq \frac{(\lambda(A, g) + \Lambda(A, g))^2}{4 \lambda(A, g) \Lambda(A, g)}.$$

BEWEIS. Sei $x \in X \setminus \{0\}$. Wir betrachten den dreidimensionalen Untervektorraum $Y \subset X$, der durch x , Ax und $A^{-1}x$ aufgespannt wird. Dazu definieren wir das Skalarprodukt h auf Y und den Operator $B: Y \rightarrow Y$ durch

$$h(y_1, y_2) := g(y_1, y_2) \quad \text{und} \quad h(By_1, y_2) := g(Ay_1, y_2) \quad \text{für alle } y_1, y_2 \in Y.$$

Wir beobachten, B bezüglich h selbstadjungiert ist und dass aufgrund der Definition von λ und Λ die folgenden Ungleichungen gelten müssen:

$$0 < \lambda(A, g) \leq \lambda(B, h) \leq \Lambda(B, h) \leq \Lambda(A, g) < \infty. \quad (3.11)$$

Für $0 < m \leq M < \infty$ gelten

$$\frac{\partial}{\partial m} \frac{(m+M)^2}{4mM} = \frac{(m-M)(m+M)}{4m^2M} \leq 0 \quad \text{und} \quad \frac{\partial}{\partial M} \frac{(m+M)^2}{4mM} = \frac{(M-m)(m+M)}{4mM^2} \geq 0. \quad (3.12)$$

Mit (3.11) folgt hieraus

$$\frac{(\lambda(B, h) + \Lambda(B, h))^2}{4 \lambda(B, h) \Lambda(B, h)} \leq \frac{(\lambda(A, g) + \Lambda(A, g))^2}{4 \lambda(A, g) \Lambda(A, g)}$$

Aufgrund der Konstruktion von Y gilt $Bx = Ax$ und $B^{-1}x = A^{-1}x$ und deshalb

$$\frac{g(Ax, x) g(A^{-1}x, x)}{g(x, x)^2} = \frac{h(Bx, x) h(B^{-1}x, x)}{h(x, x)^2}.$$

Folglich reicht es, wenn wir die folgende *endlichdimensionale* Version der Kantorovich-Ungleichung Lemma 3.16 beweisen. $\square$

Lemma 3.16 Sei (Y, h) ein höchstens dreidimensionaler Hilbertraum und es sei $B \in L(Y; Y)$ ein bezüglich h selbstadjungierter Operator. Dann gilt die Kantorovich-Ungleichung

$$1 \leq \frac{h(By, y) h(B^{-1}y, y)}{h(y, y)^2} \leq \frac{(\lambda(B, h) + \Lambda(B, h))^2}{4 \lambda(B, h) \Lambda(B, h)} \quad \text{für alle } y \in Y \setminus \{0\}.$$

BEWEIS. Der Einfachheit halber nehmen wir $\dim(Y) = 3$ an. Der Fall $\dim(Y) = 1$ ist trivial und der Fall $\dim(Y) = 2$ ist analog wie $\dim(Y) = 3$. Mit Hilfe des Satzes über die Hauptachsentransformation finden wir eine Orthonormalbasis aus Eigenvektoren e_1, e_2, e_3 und Eigenwerte $0 < \lambda(B, h) = \lambda_1 \leq \lambda_2 \leq \lambda_3 = \Lambda(B, h)$, so dass gilt

$$B e_i = \lambda_i e_i \quad \text{für alle } i \in \{1, \dots, 3\}.$$

Es genügt, die Ungleichung für $y \in Y$ mit $h(y, y) = 1$ zu zeigen. Ein solches y können wir schreiben als $y = \sum_{i=1}^3 y_i e_i$ mit $y_i = h(e_i, y)$; es gilt dann $\sum_{i=1}^3 y_i^2 = 1$. Es gilt dann

$$\frac{h(By, y) h(B^{-1}y, y)}{h(y, y)^2} = \left(\sum_{i=1}^3 \lambda_i y_i^2 \right) \left(\sum_{i=1}^3 \lambda_i^{-1} y_i^2 \right) \quad \text{und} \quad \frac{(\lambda(B, h) + \Lambda(B, h))^2}{4 \lambda(B, h) \Lambda(B, h)} = \frac{(\lambda_1 + \lambda_3)^2}{4 \lambda_1 \lambda_3}. \quad (3.13)$$

Um die Extremwerte der linken Seite von (3.13) zu bekommen, diskutieren wir nun die Einschränkung der Funktion

$$f(u_1, u_2, u_3) := \left(\sum_{i=1}^3 \lambda_i u_i \right) \left(\sum_{i=1}^3 \lambda_i^{-1} u_i \right)$$

auf dem Standardsimplex

$$S := \{ (u_1, u_2, u_3) \in \mathbb{R}^3 \mid u_1 + u_2 + u_3 = 1, u_1 \geq 0, u_2 \geq 0, u_3 \geq 0 \}.$$

Behauptung: Minimalwert und Maximalwert von f auf S werden auch auf dem relativen Rand von S realisiert, d. h., dass

$$\min \{ f(u) \mid u \in S \} = \min \{ f(u) \mid u \in R \} \quad \text{und} \quad \max \{ f(u) \mid u \in S \} = \max \{ f(u) \mid u \in R \}$$

mit dem relativen Rand

$$R := \{ u \in S \mid u_1 = 0 \text{ oder } u_2 = 0 \text{ oder } u_3 = 0 \}.$$

Wenn mindestens zwei der drei Eigenwerte $\lambda_1, \lambda_2, \lambda_3$ übereinstimmen, dann ist das sicherlich richtig: Gilt zum Beispiel $\lambda_1 = \lambda_2$ und ist $u = (u_1, u_2, u_3)$ ein Extremum mit $u_i > 0$ für alle i , dann liegt $v = (u_1 + u_2, 0, u_3)$ in R und es gilt $f(u) = f(v)$. Nehmen wir nun also an, dass $0 < \lambda_1 < \lambda_2 < \lambda_3$ gilt und dass (u_1, u_2, u_3) ein Extremum von f in $S \setminus R$ ist. Dann wäre $(u_1, u_2, u_3) = (u_1, u_2, 1 - u_1 - u_2)$ auch ein lokales Extremum von f auf der Hyperebene

$$H := \{ (u_1, u_2, u_3) \in \mathbb{R}^3 \mid u_1 + u_2 + u_3 = 1 \}.$$

Folglich müsste (u_1, u_2) ein lokales Extremum von $\varphi(u_1, u_2) := f(u_1, u_2, 1 - u_1 - u_2)$ sein. Dann müsste nach den notwendigen Bedingungen zweiter Ordnung die zweite Ableitung

$$D^2 \varphi(u_1, u_2) = \begin{pmatrix} 2 \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_3} \right) (\lambda_1 - \lambda_3) & \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_3} \right) (\lambda_1 - \lambda_3) + \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_3} \right) (\lambda_2 - \lambda_3) \\ \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_3} \right) (\lambda_1 - \lambda_3) + \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_3} \right) (\lambda_2 - \lambda_3) & 2 \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_3} \right) (\lambda_2 - \lambda_3) \end{pmatrix}$$

positiv-semidefinit oder negativ-semidefinit sein. Der linke obere Eintrag ist aber auf jeden Fall negativ, weil $\lambda_1 < \lambda_3$. Außerdem ist die Determinante

$$\det(D^2\varphi(u_1, u_2)) = -\frac{(\lambda_1 - \lambda_2)^2 (\lambda_1 - \lambda_3)^2 (\lambda_2 - \lambda_3)^2}{\lambda_1^2 \lambda_2^2 \lambda_3^2}$$

negativ. Nach dem Sylvester-Kriterium folgt aber, dass $D^2\varphi(u_1, u_2)$ indefinit sein muss. Das ist ein Widerspruch! Die Behauptung ist also wahr.

Folglich können wir die Extremewerte von f auf S finden, indem wir auf den Randkanten von R laufen. Dazu brauchen wir nur Extrema der drei folgenden univariaten Funktionen auf dem Einheitsintervall $[0, 1]$ zu finden:

$$\begin{aligned} f_{23}(t) &:= f(0, t, 1 - t) \\ f_{13}(t) &:= f(1 - t, 0, t) \\ f_{12}(t) &:= f(t, 1 - t, 0) \end{aligned}$$

Kurze Rechnung ergibt

$$f'_{ij}(t) = (2t - 1) \frac{(\lambda_i - \lambda_j)^2}{\lambda_i \lambda_j}.$$

Im Falle $\lambda_i \neq \lambda_j$ bedeutet dies, dass als Extremwerte nur

$$f_{ij}(0) = 1, \quad f_{ij}(1) = 1, \quad \text{oder} \quad f_{ij}\left(\frac{1}{2}\right) = \frac{\lambda_i + \lambda_j}{2} \frac{\lambda_i^{-1} + \lambda_j^{-1}}{2} = \frac{(\lambda_i + \lambda_j)^2}{4 \lambda_i \lambda_j}$$

in Frage kommen. Gilt $\lambda_i = \lambda_j$, so ist f_{ij} konstant gleich 1. Wegen (3.12) erhalten wir schließlich

$$\begin{aligned} \min \left\{ \frac{h(B y, y) h(B^{-1} y, y)}{h(y, y)^2} \mid u \in Y \setminus \{0\} \right\} &= \min \{ f(u) \mid u \in R \} = 1 \quad \text{und} \\ \max \left\{ \frac{h(B y, y) h(B^{-1} y, y)}{h(y, y)^2} \mid u \in Y \setminus \{0\} \right\} &= \max \{ f(u) \mid u \in R \} = \frac{(\lambda_1 + \lambda_3)^2}{4 \lambda_1 \lambda_3}, \end{aligned}$$

was den Beweis abschließt. □

3.4 Das lineare Konjugierte-Gradienten-Verfahren (CG-Verfahren)

In der Praxis verwendet man für das Lösen von (3.6) für gewöhnlich lieber das *Konjugierte-Gradienten-Verfahren*, oder kurz: das *CG-Verfahren*. Der Grund dafür ist, dass es deutlich effizienter ist. Das wollen wir nun untersuchen.

Die Idee des CG-Verfahrens ist es, Suchrichtungen $u_0, u_1, u_2, \dots$ zu benutzen, die zueinander *konjugiert*, d.h., die orthogonal sind bezüglich a :

$$a(u_i, u_j) = 0 \quad \text{für } i \neq j.$$

Dabei verwendet man den g -Gradient als erste Abstiegsrichtung:

$$u_0 := -r_0 := b - A x_0.$$

Die $k + 1$ -te Suchrichtung u_{k+1} ermittelt man dann als Linearkombination des *Residuums*

$$r_k = \text{grad}_g(F)(x_k) = A x_k - b.$$

und der bisherigen konjugierten Suchrichtungen $u_0, \dots, u_k$:

$$u_k = -r_k + \sum_{i=0}^{k-1} \gamma_{k,i} u_i.$$

Dann bestimmt man die $\gamma_{k,i}$ so, dass u_k zu $u_0, \dots, u_{k-1}$ konjugiert ist. Bildet man nämlich das a -Skalarprodukt mit u_j für $j \in \{0, \dots, k-1\}$, so bekommt man die Gleichung

$$0 \stackrel{!}{=} a(u_k, u_j) = -a(r_k, u_j) + \sum_{i=0}^k \gamma_{k,i} a(u_i, u_j) = -a(r_k, u_j) + \gamma_{k,j} a(u_j, u_j).$$

Wir erhalten also

$$\gamma_{k,j} = \frac{a(r_k, u_j)}{a(u_j, u_j)}, \quad j \in \{0, \dots, k-1\}.$$

Danach laufen wir von x_k aus in Richtung u_k um die Schrittweite $t_k > 0$, bis wir das Minimum der Funktion

$$\varphi_k(t) = F(x_k + t u_k) = \frac{1}{2} a(x_k + t u_k, x_k + t u_k) + g(b, x_k + t u_k)$$

erreicht haben. Dazu muss also

$$0 \stackrel{!}{=} \dot{\varphi}_k(t_k) = a(x_k, u_k) + t_k a(u_k, u_k) + g(b, u_k)$$

gelten. Umstellen ergibt

$$t_k = -\frac{g(r_k, u_k)}{a(u_k, u_k)} \tag{3.14}$$

und also

$$u_k = -r_k + \sum_{i=0}^{k-1} \frac{a(r_k, u_i)}{a(u_i, u_i)} u_i \quad \text{und} \quad x_{k+1} = x_k - \frac{g(r_k, u_k)}{a(u_k, u_k)} u_k. \tag{3.15}$$

Das Wunderbare am CG-Verfahren ist nun aber, dass stets gilt:

$$\gamma_{k,0} = \dots = \gamma_{k,k-2} = 0.$$

Das heißt, wir haben

$$u_k = -r_k + \gamma_k u_{k-1} \quad \text{mit} \quad \gamma_k := \gamma_{k,k-1} = \frac{a(r_k, u_{k-1})}{a(u_{k-1}, u_{k-1})}.$$

Also darf man $u_0, \dots, u_{k-2}$ getrost vergessen und braucht die anderen $\gamma_{k,j}$ (und die dafür notwendigen Skalarproduktausdrücke) gar nicht erst zu berechnen! Dies und noch eine weitere sehr wichtige Eigenschaft weisen wir nun in folgendem Lemma nach.

Lemma 3.17 *Wir definieren den Krylov-Unterraum k -ter Stufe mit Startwert r_0 durch*

$$\mathcal{K}(r_0; k) := \text{span} \{ r_0, A r_1, \dots, A^k r_k \}.$$

Es gilt für jedes $k \in \mathbb{N} \cup \{0\}$, für das x_k noch nicht globaler Minimierer von F ist:

1. $\text{span} \{ r_0, r_1, \dots, r_k \} = \mathcal{K}(r_0; k);$
2. $\text{span} \{ u_0, u_1, \dots, u_k \} = \mathcal{K}(r_0; k);$
3. $x_k \in x_0 + \mathcal{K}(r_0; k);$
4. $r_k \perp_g \mathcal{K}(r_0; k-1)$, insbesondere $g(r_k, r_j) = g(r_k, u_j) = 0$ für alle $j \in \{0, \dots, k-1\};$
5. x_{k+1} ist Minimierer von F im affinen Unterraum $x_0 + \mathcal{K}(r_0; k).$

Das heißt, was eigentlich zueinander konjugiert (a-orthogonal) ist, sind die Suchrichtungen; die Gradienten r_i sind lediglich g-orthogonal zueinander.

BEWEIS. Den Beweis für Aussagen 1. bis 4. führen wir durch Induktion nach k . Für $k = 0$ ist $\mathcal{K}(r_0; k) = \text{span} \{ r_0 \}$ und 1. ist somit erfüllt. Außerdem haben wir $u_0 = -r_0$ und dies zeigt die zweite Aussage für $k = 0$. Die Aussagen 3. und 4. sind für $k = 0$ auch offensichtlich.

Nun zum Induktionsschritt von k zu $k+$. Wir haben nach Induktionsannahme, dass $x_k = x_0 + v$ mit $v \in \mathcal{K}(r_0; k)$. Also gilt

$$r_{k+1} = A x_k - b = A x_0 - b + A v = r_0 + A v \in r_0 + A \mathcal{K}(r_0; k) \subset \mathcal{K}(r_0; k+1).$$

Weiters gilt wegen $u_0, \dots, u_k \in \mathcal{K}(r_0; k)$ und $r_{k+1} \in \mathcal{K}(r_0; k+1)$, dass

$$u_{k+1} = -r_{k+1} + \sum_{j=0}^{k-1} \gamma_{k,j} u_j \in \mathcal{K}(r_0; k+1).$$

Folglich muss

$$x_{k+2} = x_{k+1} + t_{k+1} u_{k+1} \in x_0 + \mathcal{K}(r_0; k) + \mathcal{K}(r_0; k+1) = x_0 + \mathcal{K}(r_0; k+1)$$

gelten. Wegen $x_k = x_{k-1} + t_k u_{k-1}$ haben wir

$$r_{k+1} = A x_k - b = A x_{k-1} + t_k A u_{k-1} - b = r_k + t_k A u_{k-1}.$$

Aus der Induktionsannahme und der Konjugiertheit der u_i folgt daher

$$g(r_{k+1}, u_i) = g(r_k, u_i) + t_k g(A u_{k-1}, u_i) = g(r_k, u_i) + t_k a(u_{k-1}, u_i) = 0 \quad \text{für alle } i \in \{0, \dots, k-1\}. \quad (3.16)$$

Dies zeigt schon einmal, dass $r_{k+1} \perp_g \mathcal{K}(r_0; k-1)$ ist. Da wir t_k so gewählt haben, dass x_{k+1} Minimierer von F auf der Geraden $x_k + \mathbb{R} u_k$ ist, muss auch $r_{k+1} = \text{grad}_g(F)(x_{k+1}) \perp_g u_k$ gelten:

$$0 = \varphi'_k(t_k) = DF(x_{k+1}) u_k = g(\text{grad}_g(F)(x_{k+1}), u_k) = g(r_{k+1}, u_k).$$

Also ist r_{k+1} auch g -orthogonal zu $\mathcal{K}(r_0; k) = \mathcal{K}(r_0; k-1) + \mathbb{R} u_k$. Dies zeigt Aussage 4.

Schließlich folgt Aussage 5. aus Aussage 4. und der Tatsache, dass F strikt konvex ist: Denn $\text{grad}_g(F)(x_{k+1}) \perp_g \mathcal{K}(r_0; k)$ bedeutet ja gerade, dass

$$DF(x_{k+1}) v = 0 \quad \text{gilt für alle } v \in \mathcal{K}(r_0; k).$$

Das heißt, x_{k+1} ist ein kritischer Punkt im affinen Unterraum $x_0 + \mathcal{K}(r_0; k)$. Und da F wie gesagt strikt konvex ist, muss x_{k+1} schon der Minimierer sein muss. $\square$

Lemma 3.18 *Die Iterierte x_{k+1} des CG-Verfahrens ist die beste Approximation an x_* innerhalb des affinen Raumes $x_0 + \mathcal{K}(r_0; k)$ bezüglich der Norm $\|\cdot\|_A$.*

BEWEIS. Mit $A x_* = b$ können wir wie folgt rechnen:

$$\begin{aligned} \frac{1}{2} \|x - x_*\|_A^2 &= \frac{1}{2} g(A(x - x_*), x - x_*) \\ &= \frac{1}{2} g(Ax, x) - g(Ax_*, x) + \frac{1}{2} g(Ax_*, x_*) \\ &= \frac{1}{2} g(Ax, x) - g(b, x) + g(b, x_*) - \frac{1}{2} g(Ax_*, x_*) \\ &= F(x) - F(x_*). \end{aligned}$$

Nach Lemma 3.17 minimiert x_{k+1} die Zielfunktion F innerhalb des affinen Raumes $x_0 + \mathcal{K}(r_0; k)$. Gleiches gilt dann natürlich auch für $F(\cdot) - F(x_*) = \frac{1}{2} \|\cdot - x_*\|_A^2$. $\square$

Dieses Lemma sagt gewissermaßen, dass das CG-Verfahren optimal ist unter allen Verfahren, die lediglich Matrix-Vektor-Multiplikation nutzen. Außerdem bekommen wir, dass das CG-Verfahren für niedrig-dimensionale Räume die exakte Lösung liefern kann:

Folgerung 3.19 *Wenn X ein n -dimensionaler Hilbertraum ist und wenn man exakte Arithmetik annimmt (also keine Rundungsfehler hat), dann ist spätestens nach n Iterationen der globale Minimierer x_* exakt gefunden, denn entweder bricht das CG-Verfahren mit der exakten Lösung vorher ab, oder es gilt $\mathcal{K}(r_0; n-1) = X$. In letzterem Fall muss dann wegen der letzten Aussage von Lemma 3.17 aber $x_n = x_*$ gelten.*

3.4.1 Konvergenzrate

Wenn X hoch- oder sogar unendlich-dimensional ist, dann hilft [Folgerung 3.19](#) nicht besonders. Außerdem kann man in der Praxis nicht von exakter Arithmetik ausgehen. Wir wollen nun beweisen, dass das CG-Verfahren auch brauchbare Ergebnisse für große k mit $k \ll \dim(X)$ liefern kann. Zwar berücksichtigen wir hier auch nicht den Einfluss von Fehlern der Fließkommaarithmetik; unser Konvergenzresultat legt aber trotzdem nahe, dass der Algorithmus auch solche Fehler ausbessert.

Der erste Schritt ist, den Fehler auf Ergebnisse über Polynome in einer Variablen zurückzuspielen.

Satz 3.20 Sei $\sigma(A)$ das Spektrum von A und für eine stetige Funktion $f: \mathbb{R} \rightarrow \mathbb{R}$ und eine kompakte Menge $M \subset \mathbb{R}$ sei

$$\|f\|_{L^\infty(M)} := \sup \{ |f(t)| \mid t \in M \}.$$

Dann gilt für die Iterierte x_{k+1} des CG-Verfahrens:

$$\|x_{k+1} - x_*\|_a \leq K_{k+1} \|x_0 - x_*\|_a \quad \text{und} \quad \|x_{k+1} - x_*\|_g \leq \sqrt{\kappa(a, g)} K_{k+1} \|x_0 - x_*\|_g$$

mit

$$K_k := \inf \{ \|Q\|_{L^\infty(\sigma(A))} \mid Q \text{ ist Polynom vom Grad } k \text{ mit } Q(0) = 1 \}.$$

BEWEIS. Da $x_{k+1} - x_0 \in \mathcal{K}(r_0, k)$ gilt, gibt es $\alpha_i \in \mathbb{R}$ so dass

$$x_k - x_0 = \sum_{i=0}^k \alpha_i A^i r_0.$$

Mit dem Polynom k -ten Grades

$$P_k(t) := \sum_{i=0}^k \alpha_i t^i$$

gilt also

$$x_k - x_0 = P_k(A) r_0.$$

Jedes andere $y \in x_0 + \mathcal{K}(r_0, k)$ kann man ebenfalls schreiben als $y = x_0 + P(A) r_0$ mit einem geeigneten Polynom P vom Grade höchstens k . Nach [Lemma 3.18](#) gilt folglich

$$\|x_0 + P_k(A) r_0 - x_*\|_a^2 = \|x_{k+1} - x_*\|_a^2 \leq \|y - x_*\|_a^2 = \|x_0 + P(A) r_0 - x_*\|_a^2$$

für alle Polynome vom Grad k . Hieraus folgt

$$\|x_{k+1} - x_*\|_a^2 = \inf \left\{ \|x_0 + P(A) r_0 - x_*\|_a^2 \mid P \text{ ist Polynom vom Grad } k \right\}.$$

Außerdem haben wir

$$\begin{aligned}
 (x_0 + P(A) - x_*) - x_* &= (x_0 - x_*) + P(A) r_0 \\
 &= (x_0 - x_*) + P(A) (A x_0 - b) \\
 &= (x_0 - x_*) + P(A) A (x_0 - x_*) \\
 &= (\text{Id} + P(A) A) (x_0 - x_*) = Q(A) (x_0 - x_*),
 \end{aligned}$$

wobei

$$Q(t) := 1 + P(t) t.$$

Das Polynom Q ist ein Polynom vom Grad $k + 1$ und erfüllt $Q(0) = 1$. Und jedes solche Polynom kann man darstellen als $Q(t) = 1 + P(t) t$ mit einem Polynom P vom Grad k . Hieraus folgt

$$\|x_{k+1} - x_*\|_a^2 = \inf \left\{ \|Q(A)(x_0 - x_*)\|_a^2 \mid Q \text{ ist Polynom vom Grad } k + 1 \text{ mit } Q(0) = 1 \right\}.$$

Um $\|Q(A)(x_0 - x_*)\|_a^2$ abzuschätzen, verwenden wir den Spektralsatz für beschränkte, selbstadjungierte Operatoren. Sei E_A das Projektorwertige Spektralmaß von A bezüglich g . Für beliebige Vektoren $v \in X$ haben wir dann

$$\begin{aligned}
 \|Q(A) v\|_a^2 &= g(A Q(A) v, Q(A) v) = g(Q(A) A Q(A) v, v) \\
 &= \int_{\sigma(A)} \lambda Q(\lambda)^2 d\langle E_A(\lambda) v, v \rangle \leq \|Q\|_{L^\infty(\sigma(A))}^2 \int_{\sigma(A)} \lambda d\langle E_A(\lambda) v, v \rangle \\
 &= \|Q\|_{L^\infty(\sigma(A))}^2 g(A v, v) = \|Q\|_{L^\infty(\sigma(A))}^2 \|v\|_a^2.
 \end{aligned}$$

Dies führt uns zu

$$\|x_{k+1} - x_*\|_a^2 = \inf \left\{ \|Q\|_{L^\infty(\sigma(A))}^2 \|x_0 - x_*\|_a^2 \mid Q \text{ ist Polynom vom Grad } k + 1 \text{ mit } Q(0) = 1 \right\},$$

woraus die Aussagen des Satzes unmittelbar hervorgehen. $\square$

Folgerung 3.21 *Sei das Spektrum $\sigma(A)$ von A diskret und bestehe aus nur $n \in \mathbb{N}$ Eigenwerten $0 < \lambda_1 \leq \dots \leq \lambda_n$. Dann konvergiert das CG-Verfahren in exakter Arithmetik in höchstens n Iterationen, selbst, wenn X unendlich-dimensional ist.*

BEWEIS. Wir betrachten dazu das Polynom

$$Q_*(t) := \left(1 - \frac{t}{\lambda_1}\right) \left(1 - \frac{t}{\lambda_2}\right) \cdots \left(1 - \frac{t}{\lambda_n}\right).$$

Dann gilt $Q_*(0) = 1$ und $Q_*(\lambda_i) = 0$ für alle $\lambda_i \in \sigma(A)$. Folglich haben wir

$$K_{k+1} \leq \|Q_*\|_{L^\infty(\sigma(A))} = 0,$$

also

$$\|x_{k+1} - x_*\|_a \leq K_{k+1} \|x_0 - x_*\|_a = 0. \quad \square$$

Wie wir sehen, reicht es vollkommen, ein Polynom Q mit guter L^∞ -Norm zu finden. Je nachdem, was man über das Spektrum weiß, kann man ganz unterschiedliche Konstruktionen verwenden. Kennt man lediglich untere und obere Schranke von $\sigma(A)$ so gilt immerhin der folgende Satz:

Satz 3.22 Für die Iterierten x_k des CG-Verfahrens gilt:

$$\|x_k - x_*\|_a \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x_0 - x_*\|_a \quad \text{und} \quad \|x_k - x_*\|_g \leq \sqrt{\kappa} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x_0 - x_*\|_g,$$

wobei $\kappa := \kappa(a, g)$ die Konditionszahl von a bezüglich g ist.

BEWEIS. Nach Satz 3.20 brauchen wir lediglich ein Polynom Q vom Grad k mit $Q(0) = 1$ und mit möglichst kleiner Supremumsnorm auf $[\lambda, \Lambda] := [\lambda(a, g), \Lambda(a, g)]$ zu konstruieren. Tatsächlich kann man das beste solche Polynom mit Hilfe von Chebyscheff-Polynomen angeben. Wir gehen nicht im Detail darauf ein und verweisen stattdessen auf [10, Section 3.2.2]. Als Ergebnis bekommt man jedenfalls

$$K_k \leq \inf \{ \|Q\|_{L^\infty([\lambda, \Lambda])} \mid Q \text{ ist Polynom vom Grad } k+1 \text{ mit } Q(0) = 1 \} = \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k. \quad \square$$

Wie beim einfachen Gradientenverfahren konvergieren die Iterierten des CG-Verfahrens exponentiell oder q -linear gegen den Minimierer. Wir erinnern uns: Die Konditionszahl $\kappa(a, g)$ ist immer größer oder gleich eins. In der Praxis können natürlich sehr große Konditionszahlen vorkommen. Und für große κ konvergiert $q_{\text{CG}} := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ deutlich langsamer gegen 1 als $q_{\text{Gradient}} := \frac{\kappa-1}{\kappa+1}$ (siehe Abbildung 3.1). Für große Konditionszahlen κ sollte das CG-Verfahren also deutlich schneller konvergieren und somit effizienter sein.⁵

3.4.2 Der Algorithmus (ohne Vorkonditionierer)

Für eine möglichst sparsamen Algorithmus führt man noch ein paar Umformungen durch. Zum ersten verwendet man

$$g(r_k, u_k) = g(r_k, -r_k + \gamma_k u_{k-1}) = -g(r_k, r_k) + \gamma_k g(r_k, u_{k-1}).$$

Wegen $u_{k-1} \in \mathcal{K}(r_0; k-1)$, und weil r_k wie wir wissen g -orthogonal zu $\mathcal{K}(r_0; k-1)$ ist, bekommen wir

$$t_k = -\frac{g(r_k, u_k)}{a(u_k, u_k)} = -\frac{g(r_k, r_k)}{a(u_k, u_k)}.$$

⁵Der Rechenaufwand für einen Schritt des CG-Verfahrens ist natürlich ein wenig größer als für einen Gradientenschritt; beide sind aber trotzdem ungefähr gleich groß.

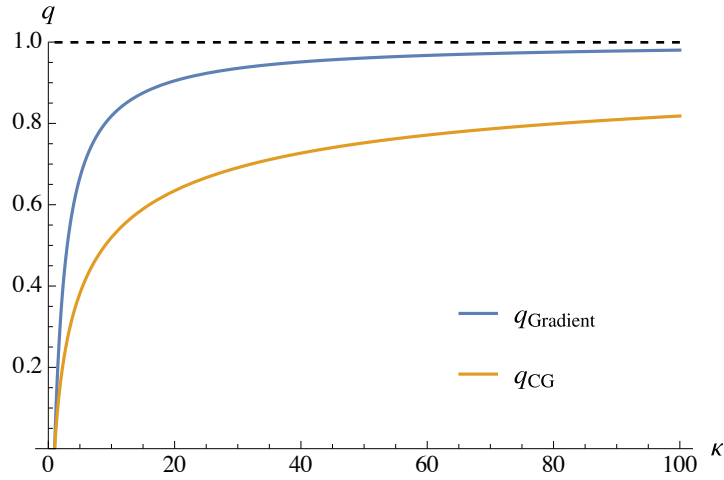


Abbildung 3.1: Vergleich von $q_{\text{CG}} := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ gegen $q_{\text{Gradient}} := \frac{\kappa-1}{\kappa+1}$: Für große Werte von κ sollte das CG-Verfahren deutlich effizienter sein als das einfache Gradientenverfahren.

Mit

$$v_k := A u_k$$

ergibt das

$$t_k = -\frac{g(r_k, r_k)}{g(u_k, v_k)}.$$

Aus (3.16) entnehmen wir, dass

$$r_{k+1} = r_k + t_k v_{k-1}$$

gilt. Umgestellt ergibt dies

$$A u_k = v_{k-1} = \frac{1}{t_k} (r_{k+1} - r_k),$$

also

$$a(r_{k+1}, u_k) = g(r_{k+1}, A u_k) = \frac{1}{t_k} g(r_{k+1}, r_{k+1})$$

Außerdem hat man

$$a(u_k, u_k) = g(A u_k, u_k) = g(A u_k, -r_k + \gamma_{k-1} u_{k-1}) = g(A u_k, -r_k) = \frac{1}{t_k} g(r_{k+1} - r_k, -r_k) = \frac{1}{t_k} g(r_k, r_k).$$

Algorithm 2 Das CG-Verfahren zum Lösen der linearen Gleichung $Ax = b$.

```

1: function CG( $x_0, A, b$ )
2:    $k \leftarrow 0$ 
3:    $r_0 \leftarrow Ax_0 - b$ 
4:    $\varrho_0 \leftarrow g(r_0, r_0)$ 
5:    $u_0 \leftarrow -r_0$ 
6:   while  $r_k \neq 0$  do
7:      $v_k \leftarrow Au_k$ 
8:      $t_k \leftarrow -\frac{\varrho_k}{g(v_k, u_k)}$ 
9:      $x_{k+1} \leftarrow x_k + t_k u_k$ 
10:     $r_{k+1} \leftarrow r_k + t_k v_{k-1}$ 
11:     $\varrho_{k+1} \leftarrow g(r_{k+1}, r_{k+1})$ 
12:     $\gamma_{k+1} \leftarrow \frac{\varrho_{k+1}}{\varrho_k}$ 
13:     $u_{k+1} \leftarrow -r_{k+1} + \gamma_{k+1} u_k$ 
14:     $k \leftarrow k + 1$ 
15:   end while
16:   return  $x_k$ 
17: end function

```

Zusammen ergibt dies

$$\gamma_{k+1} = \frac{a(r_{k+1}, u_k)}{a(u_k, u_k)} = \frac{g(r_{k+1}, r_{k+1})}{g(r_k, r_k)}.$$

Insgesamt erhält man [Algorithmus 2](#), eine Variante des CG-Verfahrens, die recht häufig verwendet wird. Wie wir sehen, benötigt jede Iteration genau eine Matrix-Vektor-Multiplikation mit A , drei Vektor-Skalar-Multiplikationen, drei Vektor-Additionen (die beiden letzten kann man zu einer Fuse-Multiplication-Add für Vektoren zusammenführen) und zwei g -Skalarprodukte. Auf dem Papier sehen die Skalarprodukte nicht teurer aus als die anderen Vektor-Operationen; trotzdem sind sie der Flaschenhals bei parallelisierten Implementierungen für Gleichungssysteme aus Finite-Elemente-Verfahren, weil sie eine Synchronisierung aller parallelen Prozesse verlangen.

3.4.3 Vorkonditionierer

Wir haben also gesehen, dass das q der q -linearen Konvergenzrate beim CG-Verfahren für $\kappa(a, g) \rightarrow \infty$ langsamer wächst als beim Gradientenverfahren. Trotzdem nimmt die Konvergenzrate des CG-Verfahrens rapide ab, wenn extrem große Konditionszahlen auftreten. Leider kommt das in der Praxis sehr häufig vor, insbesondere, wenn unendlich-dimensionale Optimierungsprobleme diskretisiert werden. Nicht selten wächst dann die Konditionszahl mit der Feinheit der Diskretisierung. Das liegt meistens daran, dass für g einfach das Euklidische Skalarprodukt auf dem $\mathbb{R}^n$ verwendet wird, der die Freiheitsgrade des diskretisierten Systems beinhaltet, ganz unabhängig davon, wie genau a eigentlich aussieht. D. h., a und g haben so gut wie nichts miteinander zu tun. Warum benutzt man dann g überhaupt? Nehmen wir also an, wir hätten Zugriff auf ein besser geeignetes Skalarprodukt h auf X , eines, für das $\kappa(a, h)$ deutlich kleiner ist. Dann können wir das CG-Verfahren mit h , statt g verwenden. Insbesondere müssen wir dafür den h -Gradienten bestimmen $\text{grad}_h(F)$. Das machen wir zuerst: Zu h gibt es einen stetigen g -selbstadjungierte linearen Operator $P \in L(X)$, so dass

$$h(u, v) = g(Pu, v) \quad \text{gilt für alle } u, v \in X.$$

Für den h -Gradienten von F muss dann einerseits gelten:

$$h(\text{grad}_h(F)(x), v) = g(P \text{grad}_h(F)(x), v) \quad \text{für alle } v \in X.$$

Andererseits gilt natürlich auch

$$h(\text{grad}_h(F)(x), v) = \langle DF(x), v \rangle = g(\text{grad}_g(F), v) = g(Ax - b, v).$$

Wir bekommen also

$$P \text{grad}_h(F)(x) = \text{grad}_g(F)(x) = Ax - b.$$

Das heißt, um $\text{grad}_h(F)(x)$ ermitteln, müssen wir ein lineares Gleichungssystem mit P lösen können. Kurz: Wir müssen

$$\text{grad}_g h(F)(x) = P^{-1}(Ax - b)$$

möglichst effizient auswerten können. Das ist die Krux. Aus Sicht der Konditionszahl $\kappa(a, h)$ wäre natürlich $h = a$ am besten. Allerdings müssten wir dann A invertieren können; aber dafür machen uns wir ja die ganze Mühe überhaupt! Es gilt also einen g -selbstadjungierten Operator P zu finden, der einfach zu invertieren ist und für den $\kappa(a, h)$ möglichst klein ist. Ein solches P nennt man einen *Vorkonditionierer* (englisch: *preconditioner*) für A . In der Praxis ist das Finden eines guten Vorkonditionierers häufig die kniffligste Aufgabe. Die konkrete Konstruktion hängt oft sehr stark von den Eigenschaften von A ab. Ist A zum Beispiel diagonal dominant, so reicht es manchmal schon, für P einfach die Diagonale von A zu verwenden. Ist A durch eine dünnbesetzte Matrix repräsentiert, so liefern unvollständige LU- oder Cholesky-Faktorisierungen oft brauchbare Vorkonditionierer.

Wir benügen uns hier erst einmal damit, den Algorithmus für das vorkonditionierte CG-Verfahren (englisch: *preconditioned CG method*) darzustellen (siehe [Algorithmus 3](#)) und noch einmal die Konvergenzsatz anzuführen (siehe [Satz 3.23](#))

Algorithm 3 Das vorkonditionierte CG-Verfahren zum Lösen der linearen Gleichung $Ax = b$.

```

1: function CG( $x_0, A, b$ )
2:    $k \leftarrow 0$ 
3:    $r_0 \leftarrow Ax_0 - b$ 
4:    $z_0 \leftarrow -P^{-1}r_0$ 
5:    $\varrho_0 \leftarrow g(z_0, r_0)$ 
6:    $u_0 \leftarrow -r_0$ 
7:   while  $r_k \neq 0$  do
8:      $v_k \leftarrow Av_k$ 
9:      $t_k \leftarrow -\frac{\varrho_k}{g(v_k, u_k)}$ 
10:     $x_{k+1} \leftarrow x_k + t_k u_k$ 
11:     $r_{k+1} \leftarrow r_k + t_k z_{k-1}$ 
12:     $z_{k+1} \leftarrow P^{-1}r_{k+1}$ 
13:     $\varrho_{k+1} \leftarrow g(z_{k+1}, r_{k+1})$ 
14:     $\gamma_{k+1} \leftarrow \frac{\varrho_{k+1}}{\varrho_k}$ 
15:     $u_{k+1} \leftarrow -z_{k+1} + \gamma_{k+1} u_k$ 
16:     $k \leftarrow k + 1$ 
17:  end while
18:  return  $x_k$ 
19: end function

```

Satz 3.23 Für die Iterierte x_k des vorkonditionierten CG-Verfahrens (siehe [Algorithmus 3](#)) gilt:

$$\begin{aligned} \|x_k - x_*\|_a &\leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x_0 - x_*\|_a, \\ \|x_k - x_*\|_g &\leq \sqrt{\kappa(a, g)} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x_0 - x_*\|_g, \\ \|x_k - x_*\|_h &\leq \sqrt{\kappa(a, h)} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x_0 - x_*\|_h, \end{aligned}$$

wobei $\kappa := \kappa(a, h)$ die Konditionszahl von a bezüglich h ist.

Übrigens kann man analog eine verbesserte Konvergenzrate für das entsprechend vorkonditionierte Verfahren des steilsten Abstiegs herleiten.

Satz 3.24 Für die Iterierte x_k des vorkonditionierten Verfahrens des steilsten Abstiegs gilt:

$$\begin{aligned} \|x_k - x_*\|_a &\leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x_0 - x_*\|_a, \\ \|x_k - x_*\|_g &\leq \sqrt{\kappa(a, g)} \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x_0 - x_*\|_g, \\ \|x_k - x_*\|_h &\leq \sqrt{\kappa(a, h)} \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x_0 - x_*\|_h, \end{aligned}$$

wobei $\kappa := \kappa(a, h)$ die Konditionszahl von a bezüglich h ist.

Bemerkung 3.25 (Abbruchkriterium) Oft wird empfohlen, sowohl das CG-Verfahren [Algorithmus 2](#) als auch das vorkonditionierte CG-Verfahren [Algorithmus 3](#) abubrechen, sobald die g -Norm des Residuums klein wird. Das ist die g -Norm des g -Gradienten. In vielen Anwendungen ist aber die a -Norm viel aussagekräftiger; nicht selten bricht $g(r_k, r_k) < \text{TOL}^2$ daher deutlich zu spät ab. Aber ein früher Abbruch wäre wünschenswert, sobald die a -Norm des a -Gradienten klein ist, denn das ist gerade der Abstand zur Lösung:

$$\begin{aligned} a(\text{grad}_a(F)(x), \text{grad}_a(F)(x)) &= a(A^{-1} \text{grad}_g(F)(x), A^{-1} \text{grad}_g(F)(x)) \\ &= a(A^{-1}(Ax - b), A^{-1}(Ax - b)) \\ &= a(x - x_*, x - x_*) = \|x - x_*\|_a^2 \end{aligned}$$

Wenn aber der Vorkonditionierer leistet, was er soll, nämlich $P \approx A$ zu gewährleisten, dann kann man stattdessen die h -Norm des h -Gradienten nehmen, denn es gilt analog

$$\begin{aligned} h(\text{grad}_h(F)(x), \text{grad}_h(F)(x)) &= h(P^{-1} \text{grad}_g(F)(x), P^{-1} \text{grad}_g(F)(x)) \\ &= h(P^{-1}(Ax - b), P^{-1}(Ax - b)) \\ &= h((P^{-1}A)(x - x_*), (P^{-1}A)(x - x_*)) \\ &= g((AP^{-1}A P^{-1}A)(x - x_*), (x - x_*)) \\ &\approx g(A(x - x_*), (x - x_*)) \\ &= a(x - x_*, x - x_*) = \|x - x_*\|_a^2. \end{aligned}$$

Andererseits hat man aber

$$\begin{aligned} h(\text{grad}_h(F)(x), \text{grad}_h(F)(x)) &= h(P^{-1} \text{grad}_g(F)(x_k), P^{-1} \text{grad}_g(F)(x_k)) \\ &= g(P^{-1} \text{grad}_g(F)(x_k), \text{grad}_g(F)(x_k)) \\ &= g(z_k, r_k) = \varrho_k. \end{aligned}$$

Das heißt, man kann

$$\varrho_k < \text{TOL}^2$$

als Abbruchkriterium verwenden. Besonders praktisch daran ist, dass diese Zahl eh schon berechnet werden muss. Das heißt, dieses Abbruchkriterium ist gratis!

3.4.4 Ein Beispiel

Ein typisches Beispiel ist dieses hier:

Beispiel 3.26 Sei $I = [0, 1] \subset \mathbb{R}$ ein abgeschlossenes Intervall. Wie definieren die Hilberträume $(L^2(I), g)$ und $(H_0^1(I), a)$ wie folgt: Für C^1 -Funktionen $u, v \in C_0^1(I)$,

$$C_0^1(I) := \{ w \in C^1(I) \mid w(0) = w(1) = 0 \}$$

definieren wir die folgenden symmetrischen Bilinearformen:

$$g(u, v) := \int_0^1 u(t) v(t) dt \quad \text{und} \quad a(u, v) := \int_0^1 \dot{u}(t) \dot{v}(t) dt.$$

Die zu g gehörende Halbnorm $\|\cdot\|_g$ ist in der Tat eine Norm, die bekannte L^2 -Norm. Wie definieren daher $L^2(I)$ als Vervollständigung von $C_0^1(I)$ unter der Norm $\|\cdot\|_g$. Gleichzeitig setzen wir g fort zu einem Skalarprodukt auf $L^2(I)$ (das wir der Einfachheit wegen ebenfalls g nennen wollen).

Auch a definiert eine Halbnorm $\|\cdot\|_a$ auf $C_0^1(I)$. Wegen

$$\|u\|_a^2 = \int_0^1 |\dot{u}(t)|^2 dt$$

Wegen

$$u(t) = u(0) + \int_0^t \dot{u}(s) ds = \int_0^t \dot{u}(s) ds$$

gilt mit der Cauchy-Schwarz-Ungleichung:

$$|u(t)| \leq \int_0^t |\dot{u}(s)| ds \leq \int_0^1 1 \cdot |\dot{u}(s)| ds \leq \left(\int_0^1 1 ds \right)^{1/2} \left(\int_0^1 |\dot{u}(s)|^2 ds \right)^{1/2} = \|u\|_a.$$

Hieraus folgt auch

$$\|u\|_g \leq \left(\int_0^1 |u(t)|^2 dt \right)^{1/2} \leq \left(\int_0^1 \|u\|_a^2 dt \right)^{1/2} \leq \|u\|_a.$$

Folglich ist $\|\cdot\|_a$ tatsächlich eine Norm, und wir können $H_0^1(I)$ definieren als Abschluss von $C_0^1(I)$ bezüglich dieser Norm.

Nehmen wir nun an, wir möchten zu $f \in L^2(I)$ das folgende System lösen:

$$a(u, v) = g(f, v) \quad \text{für alle } v \in H_0^1(I).$$

Motiviert wird dies wie üblich dadurch, dass wir partiell integrieren können:

$$a(u, v) = \int_0^1 \dot{u}(t) \dot{v}(t) dt = - \int_0^1 \ddot{u}(t) v(t) dt \stackrel{!}{=} \int_0^1 f(t) v(t) dt = g(f, v).$$

Da man dies für alle $v \in H_0^1(I)$ verlangt, müsste u also das folgende Randwertproblem lösen:

$$\begin{cases} -\ddot{u}(t) = f(t) & t \in]0, 1[, \\ u(0) = 0, \\ u(1) = 0. \end{cases}$$

Soviel zum Setup und zur Motivation. Aber was hat das alles mit unseren iterativen Verfahren zu tun? Das Problem hier ist, dass a zwar koerziv bezüglich g ist (wir haben gezeigt dass $\lambda(a, g) = 1$ ist als wir nachgewiesen haben, dass a eine Norm definiert), aber es ist nicht stetig bezüglich g . Das heißt, es gilt $\Lambda(a, g) = \infty$. Das kann man ganz leicht mit den Funktionen $u_k(t) := \sin(2\pi kt)$, $k \in \mathbb{N}$ überprüfen, denn diese sind glatt und erfüllen die Randbedingungen (d. h. es gilt $u_k \in H_0^1(I)$) und für diese gilt gerade $a(u_k, u_k) = (2\pi k)^2 g(u_k, u_k)$. Das heißt, im Unendlichdimensionalen kann das CG-Verfahren ohne Vorkonditionierer überhaupt nicht funktionieren. In der Praxis wählt man üblicherweise einen endlichdimensionalen Unterraum X_n von $H_0^1(I)$, zum Beispiel (i) stetige, stückweise lineare Funktionen bezüglich einer Unterteilung $0 = t_0 < t_1, \dots, t_n < t_{n+1} = 1$ oder (ii) die endliche Teilmenge $\{u_1, \dots, u_n\}$ der Fourierbasis mit den u_k von oben. Einschränken von a und g auf X_n liefert Bilinearformen a_n und g_n . Im Falle der Fourierbasis sieht man dann sofort ein, dass

$$\lambda(a_n, g_n) = 1, \quad \Lambda(a_n, g_n) = (2\pi n)^2, \quad \text{und} \quad \kappa(a_n, g_n) = (2\pi n)^2$$

gelten muss. Das heißt, das unvorkonditionierte CG-Verfahren funktioniert zwar, benötigt aber mehr und mehr Iterationen für $n \rightarrow \infty$. Findet man aber einen guten Vorkonditionierer h_n so dass $\kappa(a_n, h_n)$ beschränkt ist für alle $n \in \mathbb{N}$, dann braucht man in etwa die gleich Anzahl an PCG-Iterationen unabhängig von n . In der Praxis kann das einen gigantischen Unterschied machen.

3.5 Newton-Verfahren

3.5.1 Satz von Newton-Kantorovich

Wegen seiner immensen Bedeutung für die Numerik im Allgemeinen und der Optimierung im Speziellen stellen wir uns hier dem Satz von Newton-Kantorovich.

Wir bezeichnen mit $\bar{B}(a; r)$ und mit $B(a; r)$ den abgeschlossenen bzw. offenen Ball mit Radius r .

Satz 3.27 (Newton-Kantorovich) *Seien X, Y Banachräume, $U \subset X$ eine offene Menge, $\Phi \in C^1(U; Y)$ eine Abbildung und $a \in U$ ein Punkt, so dass $D\Phi(a)$ invertierbar ist. Es gebe Konstanten $\delta \geq 0$ und $L > 0$, mit $2L\delta < 1$ so dass die folgenden Bedingungen erfüllt sind:*

1. $\|D\Phi(a)^{-1}\Phi(a)\| \leq \delta$,
2. $\|D\Phi(a)^{-1}D\Phi(x) - D\Phi(a)^{-1}D\Phi(y)\| \leq L\|x - y\|$ für alle $x, y \in U$,
3. $\bar{B}(a; r) \subset U$,

wobei $r := \frac{1 - \sqrt{1 - 2L\delta}}{L}$ und $R := \frac{1 + \sqrt{1 - 2L\delta}}{L}$. Dann gibt es genau einen Punkt $x_* \in \bar{B}(a; r)$ mit $\Phi(x_*) = 0$ und die Iterierten $x_0 := a$, $x_n := x_{n-1} - D\Phi(x_{n-1})^{-1}\Phi(x_{n-1})$, $n \in \mathbb{N}$ des Newton-Verfahrens konvergieren gegen x_* mit

$$\|x_{k+1} - x_*\| \leq \frac{L}{2\sqrt{1 - 2L\delta}} \|x_k - x_*\|^2. \quad (3.17)$$

Außerdem ist x_* die einzige Lösung der Gleichung $\Phi(x) = 0$ innerhalb von

$$V := \{x \in B(a; R) \mid \forall t \in [0, 1] : a(1 - t) + tx \in U\}.$$

BEWEIS (Nach [3]). Zunächst ersetzen wir die Abbildung Φ durch $x \mapsto D\Phi(a)^{-1}\Phi(x)$. Die ändert nichts an den Newton-Iterationen, denn benennen wir die “alte” Abbildung kurz in Ψ um, sieht ein Schritt des Newton-Verfahren bei x wie folgt aus:

$$\begin{aligned} x_{\text{new}} &= x - D\Phi(x)^{-1}\Phi(x) \\ &= x - (D\Psi(a)^{-1}D\Psi(x))^{-1}(D\Psi(a)^{-1}\Psi(x)) \\ &= x - D\Psi(x)^{-1}D\Psi(a)D\Psi(a)^{-1}\Psi(x) \\ &= x - D\Psi(x)^{-1}\Psi(x). \end{aligned}$$

Wir können also ohne Einschränkung die folgenden vereinfachende Annahmen treffen: Die Abbildung $\Phi: U \rightarrow X$ bildet nach X ab, $D\Phi(a) = \text{id}_X$, $\|\Phi(a)\| \leq \delta$ und $\|D\Phi(x) - D\Phi(y)\| \leq L\|x - y\|$.

Die zentrale Idee bei diesem Beweis ist, ein eindimensionales Vergleichsproblem zu konstruieren. Dazu definieren wir die Funktion

$$\varphi: \mathbb{R} \rightarrow \mathbb{R}, \quad \varphi(t) := \frac{1}{2}Lt^2 - t + \delta.$$

Wir beobachten, dass sie selbst im Wesentlichen den Voraussetzungen des Satzes genügt:

$$\varphi(0) = \delta, \quad \varphi'(0) = -\text{id}_{\mathbb{R}}, \quad \text{und} \quad \varphi''(t) = L.$$

Insbesondere ist die Lipschitz-Konstante von φ' gleich L . Außerdem beobachten wir, dass $t = r$ und $t = R$ gerade die beiden Nullstellen von φ sind. Weiters definieren wir

$$\theta(t) := t - \frac{\varphi(t)}{\varphi'(t)}.$$

Dies ist gerade eine Newton-Iteration für die Funktion φ . Wir benötigen gleich das folgende Hilfsresultat:

Behauptung 1: θ ist streng monoton steigend auf $[0, r]$ und es gilt $\theta([0, r]) \subset [0, r]$. Für $t \in [0, r]$ gilt $\varphi'(t) = Lt - 1 < 0$; φ ist für diese t also streng monoton fallend. Deswegen gilt auch

$$\theta'(t) = 1 - \frac{\varphi'(t)^2 - \varphi(t)\varphi''(t)}{(\varphi'(t))^2} = L \frac{\varphi(t)}{(\varphi'(t))^2} \geq L \frac{\varphi(r)}{(\varphi'(r))^2} = 0$$

mit $\theta'(t) > 0$, falls $t < r$. Also ist θ streng monoton steigend. Man kann leicht nachrechnen, dass $\theta(r) = r$ gilt und hieraus folgt $\theta([0, r]) \subset [0, r]$.

Als weiteres Werkzeug definieren wir die Mengen

$$K(t) := \{x \in B(a; t) \mid \|\Phi(x)\| \leq \varphi(t)\} \quad \text{und} \quad K := \bigcup_{t \in [0, r]} K(t).$$

Die Motivation für diese Definition ist, dass wir für $x \in K$ garantieren können, dass wir wenigstens einen Newton-Schritt machen können. Dafür muss $D\Phi(x)$ stetig invertierbar sein. Wegen

$$\|\text{id}_X - D\Phi(x)\| = \|D\Phi(a) - D\Phi(x)\| \leq L\|a - x\| \leq Lr < 1$$

und [Aufgabe 3.33](#) ist das auch der Fall und wir erhalten

$$\|D\Phi(x)^{-1}\| \leq \frac{1}{1 - L\|a - x\|} = -\frac{1}{\varphi'(\|a - x\|)}. \quad (3.18)$$

Deshalb ist die folgende Newton-Iterationsabbildung wohldefiniert:

$$\Theta: K \rightarrow X, \quad \Theta(x) := x - D\Phi(x)^{-1}\Phi(x).$$

Unser nächstes Ziel ist es, dass man von K aus startend die Menge K mit Θ nicht verlassen kann, dass also $\Theta(K) \subset K$ gilt. Denn dann wissen wir per vollständiger Induktion, dass die Folge der Newton-Iterierten $x_k := \Theta^k(a)$, $k \in \mathbb{N}$ wohldefiniert sein muss.

Behauptung 2: Für alle $t \in [0, r]$ gilt $\Theta(K(t)) \subset K(\theta(t))$.

Sei $x \in K(t)$. Das heißt wir haben $\|x - a\| \leq t$ und $\|\Phi(x)\| \leq \varphi(t)$. Zusammen mit (3.18) führt dies unmittelbar zu

$$\|D\Phi(x)^{-1}\Phi(x)\| \leq \|D\Phi(x)^{-1}\| \|\Phi(x)\| \leq -\frac{\varphi(t)}{\varphi'(\|a - x\|)} \leq -\frac{\varphi(t)}{\varphi'(t)}. \quad (3.19)$$

Dies impliziert

$$\|a - \Theta(x)\| \leq \|a - x\| + \|x - \Theta(x)\| \leq t + \|D\Phi(x)^{-1}\Phi(x)\| \leq \theta(t).$$

Mit dem Satz von Taylor bekommen wir für $x, y \in V$:

$$\|\Phi(y) - (\Phi(x) + D\Phi(x)(y - x))\| \leq \frac{L}{2} \|y - x\|^2 \quad (3.20)$$

Damit und mit $\Phi(x) + D\Phi(x)(\Theta(x) - x) = 0$ bekommen wir

$$\|\Phi(\Theta(x))\| = \left\| \Phi(\Theta(x)) - (\Phi(x) + D\Phi(x)(\Theta(x) - x)) \right\| \quad (3.21)$$

$$\leq \frac{L}{2} \|\Theta(x) - x\|^2 = \frac{L}{2} \|D\Phi(x)^{-1}\Phi(x)\|^2 \leq \frac{L}{2} \left(-\frac{\varphi(t)}{\varphi'(t)} \right)^2 = \varphi(\theta(t)), \quad (3.22)$$

wobei man die letzte Gleichung wieder durch elementare Rechnung überprüfen kann. Dies impliziert aber $\Theta(x) \in \varphi(\theta(t))$, was ja zu zeigen war.

Hieraus folgt nun aber auch schon $\Theta(K) \in \varphi(K)$, denn wir haben

$$\Theta(K) = \bigcup_{t \in [0, r]} \Theta(K(t)) \subset \bigcup_{t \in [0, r]} K(\theta(t)) \subset \bigcup_{t \in \theta([0, r])} K(t) \subset \bigcup_{t \in [0, r]} K(t) = K.$$

Damit sind Newton-Iterierten also alle definiert. Da die Menge $K \subset \bar{B}(a; r)$ beschränkt ist, würde in endlichdimensionalen Räumen bereits folgen, dass die Folge $(x_k)_{k \in \mathbb{N}}$ eine konvergente Teilfolge besitzt. Im Allgemeinen müssen wir uns aber ein klein wenig mehr strecken. Überraschenderweise können wir das auf den Satz von Weierstrass über monotone Konvergenz zurückspielen. Das tun wir in den kommenden beiden Behauptungen:

Behauptung 3: Die Folge $(x_k)_{k \in \mathbb{N}}$ konvergiert gegen ein $x_* \in \bar{B}(a; r)$ mit $\Phi(x_*) = 0$.

Wir definieren die Folge $t_k := \theta^k(0)$, wobei $\theta^k = \theta \circ \dots \circ \theta$ die k -fach iterierte Abbildung bedeute. Wegen Behauptung 1 ist $(t_k)_{k \in \mathbb{N}}$ eine wohldefinierte, monoton steigende Folge in der beschränkten Menge $[0, r]$. Nach dem Monotonie-Kriterium ist sie aber auch konvergent gegen ein $t_* \in [0, r]$ und insbesondere eine Cauchy-Folge. Wegen Behauptung 2 gilt $x_k \in K(t_k)$, d. h. $\|a - x_k\| \leq t_k$ und $\|\Phi(x_k)\| \leq \varphi(t_k)$. Wegen (3.19) haben wir

$$\|x_{k+1} - x_k\| = \|D\Phi(x_k)^{-1}\Phi(x_k)\| \leq -\frac{\varphi(t_k)}{\varphi'(t_k)} = t_{k+1} - t_k.$$

Für $m > k$ rechnen wir $\|x_m - x_k\| \leq \sum_{i=k}^{m-1} \|x_{i+1} - x_i\| \leq \sum_{i=k}^{m-1} (t_{i+1} - t_i) = |t_m - t_k|$. Dies zeigt, dass auch $(x_k)_{k \in \mathbb{N}}$ eine Cauchy-Folge ist und somit gegen ein $x_* \in X$ konvergieren muss. Da die Folge in $K \subset \bar{B}(a; r)$ liegt, muss auch $x_* \in \bar{B}(a; r)$ gelten. Weil Φ stetig ist und weil $\|D\Phi(x_k)\| \leq 1 + Lr$ gilt, hat man nun auch

$$\|\Phi(x_*)\| = \lim_{k \rightarrow \infty} \|\Phi(x_{k+1})\| = \lim_{k \rightarrow \infty} \|D\Phi(x_k)(x_{k+1} - x_k)\| \leq \lim_{k \rightarrow \infty} (1 + Lr) \|x_{k+1} - x_k\| = 0.$$

Behauptung 4: Für jedes $y \in \bar{B}(x_0; r)$ mit $\Phi(y) = 0$ gilt

$$\|x_{k+1} - y\| \leq \frac{L}{2} \frac{1}{|\varphi'(t_k)|} \|x_k - y\|^2 \leq \frac{L}{2\sqrt{1-2L\delta}} \|x_k - y\|^2. \quad (3.23)$$

Mit $\Phi(y) = 0$ und (3.20) schätzen wir wie folgt ab:

$$\begin{aligned} \|x_{k+1} - y\| &= \|x_k - D\Phi(x_k)^{-1}\Phi(x_k) + D\Phi(x_k)^{-1}\Phi(y) - y\| \\ &= \|D\Phi(x_k)^{-1}(\Phi(y) - \Phi(x_k) - D\Phi(x_k)(y - x_k))\| \leq \|D\Phi(x_k)^{-1}\| \frac{L}{2} \|y - x_k\|^2. \end{aligned}$$

Aus (3.18) und der Monotonie von φ' folgt

$$\|D\Phi(x_k)^{-1}\| \leq -\frac{1}{\varphi'(\|a - x_k\|)} \leq -\frac{1}{\varphi'(t_k)} = \frac{1}{1 - Lt_k} \leq \frac{1}{1 - Lr} = \frac{1}{\sqrt{1 - 2L\delta}}$$

und wir erhalten wie gewünscht (3.23). Insbesondere zeigt dies auch die quadratische Konvergenzrate, wenn wir $y = x_*$ einsetzen.

Behauptung 5: Sei $y \in V$ eine Nullstelle von Φ . Dann gilt $y = x_*$.

Einerseits liefert uns (3.20) die Ungleichung

$$\|\Phi(y) - \Phi(a) - D\Phi(a)(y - a)\| \leq \frac{L}{2} \|y - a\|^2.$$

Andererseits gilt auch

$$\|\Phi(y) - \Phi(a) - D\Phi(a)(y - a)\| = \|\Phi(a) + (y - a)\| \geq \|y - a\| - \|\Phi(a)\| \geq \|y - a\| - \delta.$$

Zusammen ergibt dies $\varphi(\|y - a\|) = \frac{L}{2}\|y - a\|^2 - \|y - a\| + \delta \geq 0$. Wegen $\|y - a\| < R$ muss dann aber sogar $\|y - a\| \leq r$ gelten, denn φ ist auf $]r, R[$ negativ. Also ist $y \in \bar{B}(a; r)$ und wir können Behauptung 4 benutzen und mit $(r - t_k)^2$ erweitern:

$$\|y - x_{k+1}\| \leq -\frac{L}{2} \frac{1}{\varphi'(t_k)} \|y - x_k\|^2 = -\frac{L}{2} \frac{(r - t_k)^2}{\varphi'(t_k)} \frac{\|y - x_k\|^2}{(r - t_k)^2}.$$

Analog bekommen wir aber durch explizites Ausrechnen, dass

$$r - t_{k+1} = r - t_k + \frac{\varphi'(t_k)}{\varphi'(t_k)} = \frac{1}{-\varphi'(t_k)} (\varphi(r) - \varphi(t_k) - \varphi'(t_k)(r - t_k)) = -\frac{L}{2} \frac{(r - t_k)^2}{\varphi'(t_k)}.$$

Oben eingesetzt erhalten wir

$$\|y - x_{k+1}\| \leq (r - t_{k+1}) \frac{\|y - x_k\|^2}{(r - t_k)^2}.$$

Per Induktion zeigen wir damit nun $\|y - x_k\| \leq r - t_k$: Für $k = 0$ folgt dies aus $\|y - x_0\| \leq r$ und $t_0 = 0$. Den Induktionsschritt erhalten wir durch die obige Rechnung kombiniert mit der Induktionsannahme:

$$\|y - x_{k+1}\| \leq (r - t_{k+1}) \frac{\|y - x_k\|^2}{(r - t_k)^2} \leq (r - t_{k+1}).$$

Da t_k gegen r konvergiert, folgt aus $\|y - x_k\| \leq r - t_k$, dass $y = \lim_{k \rightarrow \infty} x_k = x_*$ gelten muss. Dies schließt das den Beweis ab. $\square$

Die Formulierung des Satzes ist vielleicht etwas sperrig. Eine etwas griffigere Formulierung ist die folgende:

Folgerung 3.28 Seien X, Y Banachräume, $U \subset X$ eine offene Menge, $C \geq 0$ eine Konstante und $\Phi \in C^{1,1}(U; Y)$ eine Abbildung mit $\|D\Phi(x) - D\Phi(y)\| \leq C \|x - y\|$ für alle $x, y \in U$.

Sei $a \in U$ ein Punkt, an dem $D\Phi(a)$ stetig invertierbar ist und für den gilt

$$\|\Phi(a)\| \|D\Phi(a)^{-1}\|^2 < \frac{1}{2C} \quad \text{und} \quad \bar{B}(a; 2 \|\Phi(a)\| \|D\Phi(a)^{-1}\|) \subset U.$$

Dann gibt es eine Nullstelle $x_* \in U$ von Φ mit $\|x_* - a\| \leq 2 \|\Phi(a)\| \|D\Phi(a)^{-1}\|$.

BEWEIS. Man verwende $\delta := \|\Phi(a)\| \|D\Phi(a)^{-1}\|$ und $L := C \|D\Phi(a)^{-1}\|$. Nach Voraussetzung gilt dann $2L\delta < 1$ und man hat folgende Abschätzung für den Radius r aus den Satz von Kantorovich

$$r = \frac{1 - \sqrt{1 - 2L\delta}}{L} \leq 2\delta = 2 \|\Phi(a)\| \|D\Phi(a)^{-1}\|.$$

Der Satz von Kantorovich impliziert dann gerade die Existenz der Nullstelle und die Abschätzung gegen die Residualnorm $\|\Phi(a)\|$. $\square$

3.5.2 Newton-Verfahren als Optimierungsverfahren; Residualnorm als Gütekriterium

Satz 3.4 besagt freilich nicht, dass im dritten genannten Fall die Folge $(x_k)_{k \in \mathbb{N}}$ der Iterierten konvergieren muss und schon gar nicht, dass Häufungspunkt etwa lokale Minimierer sein müssen. Ohne zusätzliche Annahmen an die Zielfunktion F lässt sich das in der Regel auch nicht garantieren. Dazu später mehr in [Satz 3.29](#).

Für eine Funktion der Klasse $C^{2,1}(M; \mathbb{R})$ mit (in einer Umgebung) hinreichend koerziven zweiten Ableitungen hat man zumindest das folgende ermutigende Resultat; um es zu beweisen benötigen wir auch den Satz von Newton-Kantorovich (siehe [Satz 3.27](#)), der uns gleichzeitig ein weiteres Optimierungsverfahren liefert.

Satz 3.29 Sei $M \subset X$ eine offene Menge im Hilbertraum (X, g) , sei $F: M \rightarrow \mathbb{R}$ eine zweimal stetig differenzierbare Funktion, sei $a \in M$ ein Punkt so dass es positive Konstanten $\varepsilon, c, C > 0$ gibt mit $0 < 2C\varepsilon < c^2$ und so dass gelten:

1. $\|DF(a)\| = \varepsilon$.
2. $D^2F(a)(u, u) \geq c\|u\|^2$ für alle $u \in X$.
3. $\|D^2F(x) - D^2F(y)\| \leq C\|x - y\|$ für alle $x, y \in M$.
4. $B(a; r) \subset M$, wobei $r := \frac{c - \sqrt{c^2 - 2C\varepsilon}}{C} \leq \frac{2\varepsilon}{c}$.

Dann gibt es in $B(a; r)$ genau ein lokales Minimum x_* von F und man hat

$$\|a - x_*\| \leq r \leq \frac{2\varepsilon}{c} = \frac{2}{c} \|DF(a)\|.$$

BEWEIS. Wir betrachten die Abbildung $\Phi := DF: M \rightarrow X'$ und die stetigen, symmetrischen Bilinearformen $b(x) := D^2F(x)$, $x \in M$. Die notwendigen Bedingungen erster Ordnung besagen gerade, dass jedes Minimum von x_* eine Nullstelle von Φ sein muss. Diese Nullstelle können wir versuchen mit dem Newton-Verfahren

$$x_0 := a, \quad x_{k+1} := x_k - D\Phi(x_k)^{-1}\Phi(x_k)$$

zu bestimmen. Unter gewissen Bedingungen besagt der Satz von Newton-Kantorovich [Satz 3.27](#), dass dies klappen muss und dass diese Nullstelle lokal eindeutig ist. Diese Bedingungen gilt es nun zu überprüfen.

Für alle $x \in B(a; r)$ und alle $u \in X$ gilt

$$D^2F(x)(u, u) = D^2F(a)(u, u) - (D^2F(a) - D^2F(x))(u, u) \quad (3.24)$$

$$\geq c\|u\|^2 - C\|a - x\|\|u\|^2 \geq \left(c - \frac{2C\varepsilon}{c}\right)\|u\|^2 \quad (3.25)$$

und analog $D^2F(x)(u, u) \leq (\|D^2F(a)\| + C)\|u\|^2$. Deswegen ist $b(x)$ ein inneres Produkt auf X , das eine zu $\|\cdot\|$ äquivalente Norm liefert. Nach [Satz 2.35](#) ist die Riesz-Abbildung $J_{b(x)}: X \rightarrow X'$ für diese x stetig invertierbar. Für $u, v \in X$ beobachten wir

$$\langle D\Phi(x)u, v \rangle = D^2F(x)(u, v) = b(x)(u, v) = \langle J_{b(x)}u, v \rangle,$$

und wir folgern, dass insbesondere $D\Phi(a) = J_{b(a)}$ stetig invertierbar ist mit $\|D\Phi(a)^{-1}\| \leq \frac{1}{c}$. Der Satz von Newton-Kantorovich (siehe [Satz 3.27](#) unten) lässt sich nun auf die Abbildung Φ anwenden mit $\delta := \frac{\varepsilon}{c}$ und $L := \frac{C}{c}$. Dieser besagt insbesondere, dass es genau einen Punkt x_* in $B(a; r)$ mit $D\Phi(x_*) = F(x_*) = 0$ gibt. Wegen (3.25) erfüllt dieser die hinreichenden Bedingungen zweiter Ordnung (siehe [Satz 2.23](#)) und ist daher ein lokales Minimum. $\square$

[Satz 3.29](#) liefert uns viele nützliche Informationen.

Bemerkung 3.30 Satz 3.29 besagt, dass unter gewissen, nicht allzu abwegigen Umständen, die Residualnorm $\|DF(x)\|$ tatsächlich einen zuverlässigen Schätzer für den tatsächlichen Abstand zum Ziel abgibt. In diesen Fällen ist also das Abbruchkriterium (8)refeq:ResidualNormStopCrit bei geeigneter Wahl von $TOL > 0$ sehr sinnvoll.

Bemerkung 3.31 Ist $b(x) := D^2F(x)$ koerziv und also ein inneres Produkt, so liefert die *Newtonsuchrichtung* $p_{\text{Newton}}(x, DF, D^2F) = -J_{b(x)}^{-1} DF(x)$ eine Abstiegsrichtung (nämlich die Richtung des steilsten Abstiegs bezüglich des ortsabhängigen inneren Produktes $b(x)$). Das heißt, das Newtonverfahren ist ein Gradientenverfahren mit *punktabhängigen* inneren Produkten $b(x)$. In der Differentialgeometrie wird eine solche Familie von inneren Produkten auch als *Riemannsche Metrik* bezeichnet.

Bemerkung 3.32 Der Satz von Newton-Kantorovich zeigt insbesondere auch quadratische Q -Konvergenz des allgemeinen Abstiegsalgorithmus mit Newtonsuchrichtung und konstanter Schrittweite $\tau = 1$, d. h., es gilt

$$\|x_{k+1} - x_*\| \leq \gamma \|x_k - x_*\|^2$$

mit einer geeigneten Konstante $\gamma > 0$. In den Übungen werden wir auch experimentell bestätigen, dass das Newtonverfahren deutlich effizienter sein kann, als die Methode des steilsten Abstiegs bezüglich eines *gegebenen* inneren Produktes.

3.5.3 Übungsaufgaben

Aufgabe 3.33 (Neumann-Reihe)

Sei $(A, \|\cdot\|, +, \cdot)$ eine *reelle Banachalgebra mit Eins*, d. h., $(A, \|\cdot\|)$ ist ein reeller Banachraum, $(A, +, \cdot)$ ist eine Algebra mit Einheit I über $\mathbb{R}$ und es gilt $\|x \cdot y\| \leq \|x\| \|y\|$ für alle $x, y \in A$ und $\|I\| = 1$. Zeigen Sie: Ist $a \in A$ ein Element mit $\|I - a\| < 1$, so ist die Reihe $b := I + \sum_{k=1}^{\infty} (I - a)^k$ absolut konvergent, es gilt $b = a^{-1}$ und man hat $\|a^{-1}\| \leq (1 - \|I - a\|)^{-1}$.

3.6 Schrittweitensteuerung

Vorab ein einfaches Beispiel, dass uns zeigen soll, dass man ohne geschickte Schrittweitenwahl im Allgemeinen nichts Gutes vom allgemeinen Abstiegsalgorithmus erwarten kann.

Beispiel 3.34 Wir betrachten die Funktion $F: \mathbb{R} \rightarrow \mathbb{R}$, $F(x) := x^4 + x^2$. Diese ist offensichtlich strikt konvex und besitzt $x_* = 0$ als globales Minimum. Angenommen, wie starten das allgemeine Abstiegsverfahren bei $x_0 = 1$ und verwenden die Methodes des steilsten Abstiegs mit konstanter Schrittweite $\tau = 1$. Wir erhalten $u = -4 - 2 = -6$ und $x_1 = \Gamma_1(-6) = -5$. Leider ist dann $F(x_1) = 650 \gg 2 = F(x_0)$. Für die nachfolgenden Iterierten hat man sogar schon $F(x_2) > 6 \times 10^{10}$, $F(x_3) > 7 \times 10^{34}$ und $F(x_4) > 8 \times 10^{106}$. Diese Suchstrategie ist also nicht nur ineffizient, sondern geradezu kontraproduktiv.

Der folgenden Situation werden wir regelmäßig begegnen: Sei $x \in M$ und $u \in X$ eine Abstiegsrichtung. Wir suchen nun entlang der Kurve $t \mapsto \Gamma_x(tu)$ nach einer möglichst geeigneten Schrittweite $t > 0$. Diese Fragestellung lässt sich sehr gut mit der folgenden Funktion diskutieren:

$$\varphi_{x,u}: [0, \infty[\rightarrow \mathbb{R}, \quad \varphi_{x,u}(t) := F(\Gamma_x(tu))$$

Für die Abbildung $\Gamma: X \times X \rightarrow X$ verlangen wir zunächst, dass sie mindestens einmal stetig differenzierbar ist und dass für alle $x \in X$ und alle $v \in X \setminus \{0\}$ gilt

1. $\Gamma_x(0) = x$,
2. $D\Gamma_x(0) = \text{id}_X$,
3. $\|\Gamma_x(tv)\| \rightarrow \infty$ für $t \rightarrow \infty$.

Wegen $\varphi'_{x,u}(t) = \langle DF(\Gamma_x(tu)), D\Gamma_x(tu)u \rangle$, $\Gamma_x(0) = x$ und $D\Gamma_x(0) = \text{id}_X$ haben wir $\varphi_{x,u}(0) = F(x)$. $\varphi'_{x,u}(0) = \langle DF(x), u \rangle$.

3.6.1 Exakte Schrittweiten

Definition 3.35 Die globalen Minimierer von $\varphi_{x,u}$ bezeichnet man als *optimale Schrittweiten*. Als *Cauchy-Schrittweite* bezeichnen wir, falls vorhanden, die kleinste optimale Schrittweite; ansonsten ist die Cauchy-Schrittweite gleich ∞ :

$$\tau_{\text{Cauchy}}(x, u) := \inf(\arg \min(\varphi_{x,u})).$$

Unter der *Curry-Schrittweite*⁶ versteht man, falls vorhanden, den kleinsten kritischen Punkt von φ ; ansonsten ist auch die Curry-Schrittweite gleich ∞ :

$$\tau_{\text{Curry}}(x, u) := \inf(\{t > 0 \mid \varphi'_{x,u}(t) = 0\}).$$

⁶Benannt nach dem amerikanischen Mathematiker und Logiker Haskell Brooks Curry, siehe https://en.wikipedia.org/wiki/Haskell_Curry. Interessanterweise sind auch die Programmiersprachen *Haskell*, *Brook* und *Curry* sowie die Methode des *Curryings* nach ihm benannt. Currying verwenden wir übrigens, wenn wir zweite Ableitungen $DD\Phi: M \rightarrow L(X; L(X; Y))$ mit (punktabhängigen) Bilinearformen $D^2\Phi: M \rightarrow L^{(2)}(X \times X; Y)$ identifizieren.

Alle diese Schrittweiten fasst man unter dem Begriff *exakte Schrittweiten* zusammen.

Sowohl *endliche* Cauchy- als auch *endliche* Curry-Schrittweiten braucht es nicht notwendigerweise zu geben; in gewissen Situationen kann man ihre Existenz aber mit dem Satz von Weierstrass herleiten:

Lemma 3.36 Sei $F: X \rightarrow \mathbb{R}$ eine stetig differenzierbare Funktion und $c \in \mathbb{R}$ sodass $L_c^-(F)$ beschränkt ist. Dann gilt für jeden Punkt $x \in L_c^-(F)$ mit Abstiegsrichtung $u \in X$ für Curry- und Cauchy- Schrittweiten:

$$0 < \tau_{\text{Curry}}(x, u) < \tau_{\text{eq}}(x, u) := \min(K) \quad \text{und} \\ 0 < \tau_{\text{Curry}}(x, u) \leq \tau_{\text{Cauchy}}(x, u) < T_{\text{eq}}(x, u) := \max(K).$$

Hierbei ist $K := \{ t > 0 \mid \varphi_{x,u}(t) = \varphi_{x,u}(0) \}$.

BEWEIS. Wir kürzen ab $\varphi := \varphi_{x,u}$. Weil $L_c^-(F)$ beschränkt ist, ist auch die Menge K beschränkt. Weil u eine Abstiegsrichtung ist, gilt $\varphi'(0) < 0$. Deswegen gibt es ein $\varepsilon > 0$ mit $\varphi(t) < \varphi(0)$ für alle $t \in]0, \varepsilon[$. Dies zeigt, dass K abgeschlossen ist. Außerdem ist K beschränkt, denn es liegt in der Subniveaumenge $L_c^-(F)$. Wegen $\lim_{t \rightarrow \infty} \|\Gamma_x(tu)\| = \infty$ gibt es ein $T \in \mathbb{R}$ mit $\Gamma_x(Tu) \notin L_c^-(F)$ und wir haben

$$\varphi(\varepsilon/2) < \varphi(0) \leq c < \varphi(T).$$

Nach dem Zwischenwertsatz ist die Menge K also auch nicht die leere Menge und $\tau_{\text{eq}}(x, u) := \min(K)$ ist wohldefiniert. Nach dem Satz von Rolle⁷ muss es dann wenigstens einen kritischen Punkt im Intervall $]0, \tau_{\text{eq}}(x, u)[$ geben. Damit haben wir $0 \leq \tau_{\text{Curry}}(x, u) < \tau_{\text{eq}}(x, u)$. Weil φ stetig differenzierbar ist, ist die Menge der kritischen Punkte abgeschlossen; deswegen muss auch $\tau_{\text{Curry}}(x, u)$ ein kritischer Punkt sein. Folglich gilt $\tau_{\text{Curry}}(x, u) \neq 0$.

Analog impliziert der Satz von Weierstrass die Existenz eines Minimums von φ in $[0, \max(K)]$; dieses muss dann auch ein Minimum von φ auf $[0, \infty[$ sein. Dies zeigt $\tau_{\text{Cauchy}}(x, u) < \max(K)$. Wegend der notwendigen Bedingung erster Ordnung für lokale Minima muss aber auch $\tau_{\text{Curry}}(x, u) \leq \tau_{\text{Cauchy}}(x, u)$ gelten. $\square$

Definition 3.37 Sei $F: X \rightarrow \mathbb{R}$ eine stetig differenzierbare Funktion und sei $c \in \mathbb{R}$ so dass $L_c^-(F)$ beschränkt ist. Wir sagen, F und Γ erfüllen die *Lipschitz-Bedingung* mit Konstante $L \geq 0$ genau dann, wenn für alle $x \in L_c^-(F)$ mit Abstiegsrichtung $u \in X$ gilt

$$|\varphi'_{x,u}(s) - \varphi'_{x,u}(t)| \leq |s - t| L \|u\|^2 \quad \text{für alle } s, t \in [0, \tau_{\text{eq}}(x, u)]. \quad (3.26)$$

Lemma 3.38 Die Funktion $F: X \rightarrow \mathbb{R}$ sei stetig differenzierbar mit beschränkter Subniveaumenge $L_c^-(F)$ und erfülle die Lipschitz-Bedingung mit Konstante $L \geq 0$. Dann gilt für alle $x \in L_c^-(F)$ mit Abstiegsrichtung $u \in X$:

⁷Der Satz von Rolle ist übrigens auch eine Folge des Satzes von Weierstrass.

1. $\varphi_{x,u}(t) \leq \varphi_{x,u}(0) + \varphi'_{x,u}(0)t + \frac{L}{2} \|u\|^2 t^2$ für alle $t \in [0, \tau_{\text{eq}}(x, u)]$.
2. $\tau_{\text{eq}}(x, u) \geq -\frac{2\varphi'_{x,u}(0)}{L\|u\|^2}$.
3. $\tau_{\text{Curry}}(x, u) \geq -\frac{\varphi'_{x,u}(0)}{L\|u\|^2}$.
4. Die Curry-Schrittweitenregel erfüllt auf $L_c^-(F)$ die Effizienzbedingung (3.4) mit Konstante $\vartheta = \frac{1}{2L}$:

$$F(\Gamma_x(\tau_{\text{Curry}}(x, u)u)) \leq F(x) - \frac{1}{2L} \frac{\langle DF(x), u \rangle^2}{\|u\|^2}.$$

BEWEIS. Wir kürzen ab $\varphi := \varphi_{x,u}$.

Zu 1.:

$$\begin{aligned} \varphi(t) - \varphi(0) &= \int_0^t \varphi'(s) \, ds = t\varphi'(0) + \int_0^t (\varphi'(s) - \varphi'(0)) \, ds \\ &\leq t\varphi'(0) + \int_0^t sL\|u\|^2 \, ds = t\varphi'(0) + \frac{L}{2} \|u\|^2 t^2. \end{aligned}$$

Zu 2.: Einsetzen von $\tau_{\text{eq}}(x, u)$ in Punkt 1. und Ausnutzen von $\varphi(0) = \varphi(\tau_{\text{eq}})$.

Zu 3.: Wegen $0 < \tau_{\text{Curry}}(x, u) < \tau_{\text{eq}}(x, u)$ gilt für $t := \tau_{\text{Curry}}(x, u)$ nach Lemma 3.38:

$$0 = \varphi(t) = \varphi(0) + (\varphi(t) - \varphi(0)) \leq \varphi(0) + \tau_{\text{Curry}}(x, u) C \|u\|^2.$$

Zu 4.: Weil φ auf dem Intervall $[0, \tau_{\text{Curry}}(x, u)]$ monoton fallend ist, folgt mit $s := -\frac{\varphi'(0)}{L\|u\|^2}$ aus Punkt 3 und Punkt 1:

$$\varphi(\tau_{\text{Curry}}(x, u)) \leq \varphi(s) \leq \varphi(0) + s\varphi'(0) + \frac{L}{2} \|u\|^2 s^2 = \varphi(0) - \frac{1}{2L} \left(\frac{\varphi'(0)}{\|u\|} \right)^2.$$

Startet man nun den allgemeinen Abstiegsalgorithmus (siehe Algorithmus 1) mit der Curry-Schrittweitenregel und mit Startpunkt $x_0 \in L_c^-(F)$, so verbleiben die Iterierten x_k in $L_c^-(F)$. Per vollständiger Induktion sieht man nun leicht ein, dass dann die Effizienzbedingung (3.4) stets erfüllt ist. $\square$

Die Lipschitz-Bedingung ist nicht allzu künstlich, denn man hat:

Lemma 3.39 Die Funktion $F: X \rightarrow \mathbb{R}$ Funktion sei stetig differenzierbar mit beschränkter Subniveaumenge $L_c^-(F)$. Es gebe es Konstanten $C(F)$, $C(DF)$, $C(\Gamma)$ und $C(D\Gamma) \geq 0$ so dass für alle $x, y \in L_c^-(F)$ und alle $v \in X$ gilt:

1. $\|DF(x)\| \leq C(F)$,
2. $\|DF(x) - DF(y)\| \leq C(DF) \|x - y\|$,

$$3. \quad \|D\Gamma_x(t v) v\| \leq C(\Gamma) \|v\|.$$

$$4. \quad \|D\Gamma_x(s v) v - D\Gamma_x(t v) v\| \leq C(D\Gamma) \|v\|^2 |s - t| \text{ für alle } s, t \in [0, \tau_{\text{eq}}(x, u)].$$

Dann ist die Lipschitz-Bedingung (3.26) mit Konstante $L := C(D\Gamma) C(\Gamma)^2 + C(F) C(D\Gamma)$ erfüllt.

BEWEIS. Seien $s, t \in [0, \tau_{\text{eq}}(x, u)]$. Wir kürzen ab $\gamma(t) := \Gamma_x(t u)$ und beobachten $\gamma'(t) = D\Gamma_x(t u) u$. Es gilt $\gamma(s), \gamma(t) \in L_c^-(F)$. Aus $\gamma(s) - \gamma(t) = \int_s^t \gamma'(r) dr$ erhalten wir mit Bedingung 3 die Abschätzung $\|\gamma(s) - \gamma(t)\| \leq C(\Gamma) \|u\| |s - t|$. Daher können wir wie folgt abschätzen:

$$\begin{aligned} |\varphi'_{x,u}(s) - \varphi'_{x,u}(t)| &= |\langle DF(\gamma(s)), \gamma'(s) \rangle - \langle DF(\gamma(t)), \gamma'(t) \rangle| \\ &\leq |\langle DF(\gamma(s)) - DF(\gamma(t)), \gamma'(s) \rangle| + |\langle DF(\gamma(t)), \gamma'(s) - \gamma'(t) \rangle| \\ &\leq C(D\Gamma) \|\gamma(s) - \gamma(t)\| \|\gamma'(s)\| + C(F) \|\gamma'(s) - \gamma'(t)\| \\ &\leq C(D\Gamma) C(\Gamma) \|u\| |s - t| C(\Gamma) \|u\| + C(F) C(D\Gamma) \|u\|^2 |s - t| \\ &= |s - t| L \|u\|^2. \end{aligned} \quad \square$$

Bemerkung 3.40 Für die Standard-Aufdatierungsabbildung $\Gamma_x(v) = x + v$ in normierten Räumen gilt $C(\Gamma) = 1$ und $C(D\Gamma) = 0$, so dass man die folgende, sehr einfache Abschätzung erhält:

$$|\varphi'_{x,u}(s) - \varphi'_{x,u}(t)| \leq |s - t| C(D\Gamma) \|u\|^2 \quad \text{für alle } s, t \in [0, \tau_{\text{eq}}(x, u)]. \quad (3.27)$$

Natürlich kann man versuchen, exakte Schrittweiten durch Nullstellen-Algorithmen mehr oder weniger exakt zu bestimmen, z. B. mit der regula falsi oder mit dem Newton-Verfahren. Dafür werden aber etliche Auswertungen $\varphi'_{x,u}$ und $\varphi''_{x,u}$ benötigt. Dazu muss man wiederum Richtungsableitungen $DF(y) v$ und $D^2F(y)(v, v)$ der Zielfunktion berechnen. Das kann für komplizierte Zielfunktionen teuer werden (und oft kann oder möchte man auch keine zweiten Ableitungen implementieren). Deswegen behandeln wir als nächstes Schrittweitenstrategien, die mit einigen Auswertungen von $\varphi_{x,u}$ und nur wenigen Auswertungen von $\varphi'_{x,u}$ auskommen und trotzdem zu effizienten Suchstrategien führen.

3.6.2 Armijo-Schrittweiten

Definition 3.41 Seien $\sigma \in]0, 1[$. Wir sagen, eine Schrittweite $t > 0$ erfüllt die *Armijo-Bedingung* mit Konstante σ genau dann, wenn

$$\varphi_{x,u}(t) \leq \varphi_{x,u}(0) + t \sigma \varphi'_{x,u}(0). \quad (3.28)$$

Lemma 3.42 Sei $M \subset X$ eine offene Menge, $F: M \rightarrow \mathbb{R}$ eine differenzierbare Funktion und sei $\sigma \in]0, 1[$ gegeben. Sei $x \in M$ ein Punkt mit Suchrichtung $u \in X$. Dann gibt es ein $\delta = \delta(x, u) > 0$ so dass alle $t \in]0, \delta[$ die Armijo-Bedingung mit Konstante σ erfüllen.

BEWEIS. Wir kürzen ab $\varphi := \varphi_{x,u}$.

Auf dem nicht-trivialen Intervall $I = \{t \geq 0 \mid \forall s \in [0, t]: \Gamma_x(su) \in M\}$ betrachten wir die affine Funktion $\psi: I \rightarrow \mathbb{R}$, $\psi(t) := \varphi(0) + t \sigma \varphi'(0)$ und die Menge $A := \{t \in I \mid \psi(t) < \varphi(t)\}$. Ist A die leere Menge, so können wir $\delta \in I$ beliebig groß wählen. Andernfalls besitzt die Menge A ein Infimum $\delta \geq 0$. Wäre $\delta = 0$ so gäbe es eine Folge $t_n \in A$ mit $\lim_{n \rightarrow \infty} t_n = 0$ und man bekäme folgenden Widerspruch

$$\sigma \varphi'(0) > \varphi'(0) = \lim_{n \rightarrow \infty} \frac{\varphi(t_n) - \varphi(0)}{t_n} \geq \lim_{n \rightarrow \infty} \frac{\psi(t_n) - \psi(0)}{t_n} = \psi'(0) = \sigma \varphi'(0).$$

Also muss $\delta > 0$ sein. □

Wir brauchen einen Schrittweitschätzer τ_k also bloß sukzessive (z. B. durch Skalieren mit einem Faktor $0 < s < 1$ verkleinern, wenn die Armijo-Bedingung nicht erfüllt wird, und erhalten schließlich eine Armijo-Schrittweite. Diese kann man noch durch folgenden Interpolationstrick verbessern: Aus der Bestimmung der Suchrichtung hat man meistens $\varphi(0)$ und $\varphi'(0)$ schon vorliegen; zum Testen der Armijo-Bedingung wird $\varphi(\tau_k)$ ausgewertet. Dies genügt, um hieran ein quadratisches Polynom $\chi(t)$ mit $\chi(0) = \varphi(0)$, $\chi'(0) = \varphi'(0)$ und $\chi(\tau_k) = \varphi(\tau_k)$ zu fitten. Sei $\tau_k^* = -\frac{1}{2} \frac{\tau_k^2 \varphi'(0)}{\varphi(t) - \varphi(0) - \tau_k \varphi'(0)}$ (falls existent) der kritische Punkt von χ . Als verbesserten Schrittweitschätzer verwenden wir anschließend $\tau_{k+1} := \max(c \tau_k, \tau_k^*)$. Dies führt direkt zu [Algorithmus 4](#); dieser ist auch als *backtracking line search* bekannt. Verwendet man ihn in Verbindung mit Newton-artigen Verfahren, so ist $\tau_0 = 1$ eine geeignete Wahl eines Startwertes, denn für diese ist zumindest in der Nähe einer Lösung optimal. Ist die Zielfunktion konvex und quadratisch, so liefert τ_0^* immer die optimale Schrittweite.

Satz 3.43 Sei $F: X \rightarrow \mathbb{R}$ eine stetig differenzierbare Funktion und sei $x \in M$ ein Punkt mit Abstiegsrichtung $u \in X$. Seien Konstanten $\tau_0 > 0$, $\sigma \in]0, 1/2[$, $S := \frac{1}{2(1-\sigma)}$ und $s \in]0, S[$ gegeben.

Dann gilt $\tau_{k+1} \in [s \tau_k, S \tau_k]$ für die in [Algorithmus 4](#) erzeugten τ_k und nach endlich vielen Iterationen ist eine Armijo-Schrittweite gefunden.

BEWEIS. Sei $\varphi(t) := F(x + tu)$. Ist im k -ten Schritt die Armijo-Bedingung noch nicht erfüllt, so gilt $\varphi(\tau_k) > \varphi(0) + \tau_k \sigma \varphi'(0)$ und damit hat man

$$0 < (\sigma - 1) \tau_k \varphi'(0) < \varphi(\tau_k) - \varphi(0) - \tau_k \varphi'(0).$$

Algorithm 4 Armijo-Schrittweitenregel mit quadratischer Interpolation

Require: $\langle DF(x), u \rangle < 0$, $\tau_0 > 0$, $\sigma \in]0, 1/2[$, $s \in]0, (2(1 - \sigma))^{-1}[$.

```

1: function ARMIJOSTEPsize( $x, u, F, DF$ )
2:    $k \leftarrow 0$ 
3:   while  $F(\Gamma_x(\tau_k u)) > F(x) + \tau_k \sigma \langle DF(x), u \rangle$  do                                ▶ Prüfe Armijo-Bedingung.
4:      $\tau_k^* \leftarrow -\frac{1}{2} \frac{\tau_k^2 \langle DF(x), u \rangle}{F(\Gamma_x(\tau_k u)) - F(x) - \tau_k \langle DF(x), u \rangle}$                 ▶ Krit. Punkt d. quadr. Interpolation
5:      $\tau_{k+1} \leftarrow \max(s \tau_k, \tau_k^*)$                                                 ▶ Backtracking
6:      $k \leftarrow k + 1$ 
7:   end while
8:   return  $\tau_k$ 
9: end function
    
```

Folglich gilt

$$\tau_k^* = -\frac{1}{2} \frac{\tau_k^2 \varphi'(0)}{\varphi(\tau_k) - \varphi(0) - \tau_k \varphi'(0)} < -\frac{1}{2} \frac{\tau_k^2 \varphi'(0)}{(\sigma - 1) \tau_k \varphi'(0)} = \frac{1}{2(1 - \sigma)} \tau_k = S \tau_k.$$

Wegen $s \leq S < 1$ gilt $\tau_{k+1} \in [s \tau_k, S \tau_k]$ und $\tau_k \leq S^k \tau_0$. Für hinreichend großes k gilt aber $S^k \tau_0 < \delta$ mit dem $\delta > 0$ aus Lemma 3.42 und spätestens ab dann muss τ_k eine Armijo-Schrittweite sein. $\square$

Armijo-Schrittweitenregeln brauchen in der Regel eine vorgegebene Effizienzbedingung nicht zu erfüllen: Schließlich erfüllen hinreichend kleine Schrittweiten ja immer eine Armijo-Bedingung.⁸ Die in Algorithmus 4 implementierte Schrittweitenregel stellt dies unter geeigneten Zusatzbedingungen aber auf geschickte Weise doch wieder sicher:

Satz 3.44 Sei $F: X \rightarrow \mathbb{R}$ eine stetig differenzierbare Funktion mit beschränkter Subniveaumenge $L_c^-(F)$ und F, Γ füllen die Lipschitz-Bedingung (3.26) mit Konstante $L \geq 0$. Seien Konstanten $\tau_0 > 0$, $\sigma \in]0, 1/2[$, $S := \frac{1}{2(1-\sigma)}$ und $s \in]0, S]$ gegeben.

Dann erfüllt die von Algorithmus 4 gefundene Schrittweite $t = \tau_{\text{Armijo}}(x, u)$ die Ungleichung

$$F(\Gamma_x(tu)) \leq F(x) - \vartheta \min \left(-\langle DF(x), u \rangle, \left(\frac{\langle DF(x), u \rangle}{\|u\|} \right)^2 \right) \quad (3.29)$$

mit $\vartheta := \sigma \min \left(\tau_0, \frac{2s(1-\sigma)}{L} \right)$. Insbesondere heißt das, dass die Suchstrategie $(p, \tau_{\text{Armijo}}, \Gamma)$ die die Effizienzbedingung (3.4) mit Konstante ϑ auf $L_c^-(F)$ erfüllt.

BEWEIS. Sei $x \in L_c^-(F)$ ein Punkt mit Abstiegsrichtung $u \in X$. Sei $t = \tau_k$ die ermittelte Schrittweite. Wir kürzen ab $\varphi := \varphi_{x,u}$. Ist $k = 0$, so haben wir $t = \tau_0$ und daher

$$F(\Gamma_x(tu)) = F(\Gamma_x(\tau_0 u)) = \varphi(t) \leq \varphi(0) + \tau_0 \sigma \varphi'(0) \leq F(x) + \tau_0 \sigma \langle DF(x), u \rangle.$$

⁸Ist zusätzlich noch der zulässige Bereich eingeschränkt ($M \neq X$), so gibt es nahe des Randes von M möglicherweise Punkte mit Suchrichtungen, die nach außen zeigen. Dort kann man dann auch keine unteren Schranken für Schrittweiten erwarten.

Das ist aber schon (3.29) für $\vartheta = \sigma$. Ist $k > 1$, so hat man

$$\varphi(\tau_k) \leq \varphi(0) + \tau_k \sigma \varphi'(0), \quad \varphi(\tau_{k-1}) < \varphi(0) + \tau_{k-1} \sigma \varphi'(0) \quad \text{und} \quad s \tau_{k-1} \leq \tau_k.$$

Wir unterscheiden zwei Fälle:

Fall 1: $\tau_{k-1} \leq \tau_{\text{eq}}(x, u)$. Wir kombinieren die verfehlte Armijo-Bedingung mit Lemma 3.38 und erhalten

$$\varphi(0) + \tau_{k-1} \sigma \varphi'(0) < \varphi(\tau_{k-1}) \leq \varphi(0) + \varphi'(0) \tau_{k-1} + \frac{L}{2} \|u\|^2 \tau_{k-1}^2$$

Umgestellt ergibt dies

$$t = \tau_k \geq s \tau_{k-1} \geq -\frac{2s(1-\sigma)}{L} \frac{\varphi'(0)}{\|u\|^2}.$$

Fall 2: $\tau_{k-1} > \tau_{\text{eq}}(x, u)$. Nach Lemma 3.38 gilt dann $\tau_{k-1} > -\frac{2\varphi'(0)}{L\|u\|^2}$ und folglich

$$t = \tau_k \geq s \tau_{k-1} \geq -\frac{2s}{L} \frac{\varphi'(0)}{\|u\|^2}.$$

In beiden Fällen ist aber die Armijo-Bedingung erfüllt; deswegen gilt

$$\varphi(t) \leq \varphi(0) + t \sigma \varphi'(0) \leq \varphi(0) - \sigma \frac{2s(1-\sigma)}{L} \left(\frac{\varphi'(0)}{\|u\|} \right)^2.$$

Mit $\vartheta = \sigma \min\left(\tau_0, \frac{2s(1-\sigma)}{L}\right)$ erhalten wir

$$F(\Gamma_x(tu)) \leq F(x_k) - \vartheta \min\left(-\langle DF(x), u \rangle, \frac{\langle DF(x), u \rangle^2}{\|u\|^2}\right).$$

Dass auch die Effizienzbedingung (3.4) erfüllt sein muss, sieht man schließlich mit vollständiger Induktion ein, denn die Iterierten x_k des allgemeinen Abstiegsverfahrens zur Anfangsbedingung $x_0 \in L_c^-(F)$ können die Menge $L_c^-(F)$ wegen der Armijo-Bedingung nicht verlassen. $\square$

3.6.3 Wolfe-Powell-Schrittweiten

Definition 3.45 Seien $0 < \sigma < \varrho < 1$, $\varphi := \varphi_{x,u}$. Wir sagen, eine Schrittweite $t > 0$ erfüllt

1. die Krümmungsbedingung mit Konstante ϱ genau dann, wenn

$$\varphi'(t) \geq \varrho \varphi'(0). \quad (3.30)$$

2. die Wolfe-Powell-Bedingungen mit Konstanten (σ, ϱ) genau dann, wenn sowohl die Armijo- als auch die Krümmungsbedingung erfüllt sind, d. h., wenn

$$\varphi(t) \leq \varphi(0) + t \sigma \varphi'(0) \quad \text{und} \quad \varphi'(t) \geq \varrho \varphi'(0). \quad (3.31)$$

3. die starke Wolfe-Powell-Bedingungen mit Konstanten (σ, ϱ) genau dann, wenn

$$\varphi(t) \leq \varphi(0) + t \sigma \varphi'(0) \quad \text{und} \quad |\varphi'(t)| \leq \varrho |\varphi'(0)|. \quad (3.32)$$

Satz 3.46 Sei $F: X \rightarrow \mathbb{R}$ eine stetig differenzierbare Funktion mit beschränkter Subniveaumenge $L_c^-(F)$. Sei (p, τ, Γ) eine Suchstrategie, so dass die Lipschitz-Bedingung (3.26) mit Konstante $L \geq 0$ auf $L_c^-(F)$ erfüllt ist und so dass die generierten Schrittweiten die Wolfe-Powell-Bedingung mit Konstanten (σ, ϱ) erfüllen.

Dann erfüllt (p, τ, Γ) die Effizienzbedingung (3.4) mit Konstante $\vartheta = \sigma \frac{(1-\varrho)}{L}$ auf $L_c^-(F)$. Dann gilt für alle $x \in L_c^-(F)$ mit Abstiegsrichtung $u = p(x)$ und Schrittweite $t = \tau(x, u)$ die Ungleichung

$$F(\Gamma_x(tu)) \leq F(x) - \sigma \frac{(1-\varrho)}{L} \left(\frac{\langle DF(x), u \rangle}{\|u\|} \right)^2. \quad (3.33)$$

Insbesondere heißt das, dass die Suchstrategie (p, τ, Γ) auf $L_c^-(F)$ die Effizienzbedingung (3.4) mit Konstante $\vartheta := \sigma \frac{(1-\varrho)}{L}$ erfüllt.

BEWEIS. Wieder stellt die Armijo-Bedingung sicher, dass die Iterierten des allgemeinen Abstiegsalgorithmus die Subniveaumenge $L_c^-(F)$ nicht verlassen.

Zu einem festen Punkt $x \in L_c^-(F)$ und einer festen Abstiegsrichtung $u \in X$ betrachten wir wieder die Funktion $\varphi = \varphi_{x,u}: [0, \delta] \rightarrow \mathbb{R}$, wobei $\delta > 0$ wie in Lemma 3.42 gewählt sei. Nun kombinieren wir die Lipschitz-Bedingung $|\varphi'(t) - \varphi'(0)| \leq t L \|u\|^2$ mit der Krümmungsbedingung $\varphi'(t) - \varphi'(0) \geq (\varrho - 1) \varphi'(0)$ und erhalten wegen $\varphi(t) \leq \varphi(0)$ die Ungleichung

$$(\varrho - 1) \varphi'(0) \leq \varphi'(t) - \varphi'(0) \leq t L \|u\|^2.$$

Division durch $L \|u\|^2$ und Einsetzen von $\varphi'(0) = \langle DF(x), u \rangle$ ergibt $t \geq -\frac{(1-\varrho)}{L} \frac{\langle DF(x), u \rangle}{\|u\|^2}$. Analog zum Beweis von Satz 3.44 setzt man dies wieder in die Armijo-Bedingung ein und erhält

$$F(\Gamma_x(tu)) = \varphi(t) \leq \varphi(0) + t \sigma \varphi'(0) \leq F(x) - \sigma \frac{(1-\varrho)}{L} \frac{\langle DF(x), u \rangle^2}{\|u\|^2}.$$

Vollständige Induktion zeigt schließlich die Effizienzbedingung (3.4) für alle Startwerte $x_0 \in L_c^-(F)$. $\square$

Satz 3.47 Sei $F: X \rightarrow \mathbb{R}$ eine stetig differenzierbare Funktion, Γ eine stetig differenzierbare Aufdatierungsabbildung, $x \in X$ ein Punkt mit Abstiegsrichtung $u \in X$ und seien $0 < \sigma < \varrho < 1$ gegeben. Ist $\varphi_{x,u}$ nach unten beschränkt, so besitzt die Menge $S_{\text{WP}} := \{t > 0 \mid t \text{ erfüllt (3.32)}\}$ der starken Wolfe-Powell-Schrittweiten ein nichtleeres Inneres.

BEWEIS. Seien $\psi(t) := \varphi(0) + t \sigma \varphi'(0)$ und $A := \{t \in [0, \infty[\mid \psi(t) < \varphi(t)\}$ wie im Beweis von Lemma 3.42. Diesmal kann diese Menge nicht die leere Menge sein, denn nun ist φ von unten beschränkt, ψ aber nicht. Für $\delta := \inf(A)$ gilt deshalb $\delta > 0$ und $\varphi(\delta) = \psi(\delta)$.

Nach dem Mittelwertsatz gibt es ein $t \in]0, \delta[$, so dass gilt:

$$\varphi'(t) = \frac{\varphi(\delta) - \varphi(0)}{\delta - 0}.$$

Wegen $t < \delta$ ist auch die Armijo-Bedingung erfüllt. Nun erhält man

$$|\varphi'(t)| = \frac{|\varphi(\delta) - \varphi(0)|}{\delta} = \frac{|\psi(\delta) - \psi(0)|}{\delta} = \sigma |\varphi'(0)| < \varrho |\varphi'(0)|.$$

Das ist aber gerade die starke Wolfe-Powell-Bedingung. Es gilt also $t \in S_{\text{WP}}$. Weil t die starken Wolfe-Powell-Bedingungen aber mit strikten Ungleichungen erfüllt und weil F und Γ und damit φ stetig differenzierbar sind, gibt es auch eine Umgebung von t , die auch in S_{WP} liegt. $\square$

Damit enthält die Menge der (starken) Wolfe-Powell-Schrittweiten ein Intervall endlicher Länge. Das bedeutet, dass Intervallschachtelungsverfahren für die Suche nach (starken) Wolfe-Powell-Schrittweiten nach endlich viele Iterationen determinieren. Ein solches Verfahren für das finden einer (schwachen!) Wolfe-Powell-Schrittweite wird etwa in Algorithmus 5 umgesetzt. Um zu verstehen, wie er funktioniert betrachten wir zunächst das folgende Lemma:

Lemma 3.48 Sei $F: X \rightarrow \mathbb{R}$ eine stetig differenzierbare Funktion, Γ eine stetig differenzierbare Aufdatierungsabbildung, $x \in X$ ein Punkt mit Abstiegsrichtung $u \in X$ und seien $0 < \sigma < \varrho < 1$ gegeben. Ferner seien $0 \leq a \leq b$ gegeben so dass

$$\varphi(a) \leq \varphi(0) + a \sigma \varphi'(0), \quad \varphi'(a) < \varrho \varphi'(0) \quad (3.34)$$

und

$$\varphi(b) > \varphi(0) + b \sigma \varphi'(0) \quad (3.35)$$

gelten. Dann enthält $]a, b[$ eine offene Menge von (schwachen) Wolfe-Powell-Schrittweiten.

BEWEIS. Sei wieder $\psi(t) := \varphi(0) + t \sigma \varphi'(0)$ und $A := \{t \in [a, \infty[\mid \psi(t) < \varphi(t)\}$. Wegen $b \in A$ ist A nicht leer. Wegen $\varphi(a) \leq \psi(a)$ und $\varphi'(a) < \varrho \varphi'(0) < \sigma \varphi'(0) = \psi'(a)$ gilt

$$\delta := \inf(A) > 0.$$

Wie oben sieht man auch ein, dass $\varphi(\delta) = \psi(\delta)$ gelten muss. Abermals mit dem Mittelsatz finden wir ein $t \in]a, \delta[$ so dass

$$\varphi'(t) = \frac{\varphi(\delta) - \varphi(a)}{\delta - a}.$$

Und nun rechnen wir analog zu oben mit $\varphi(a) \leq \psi(a)$ und $\varphi'(0) < 0$:

$$\varphi'(t) = \frac{\varphi(\delta) - \varphi(a)}{\delta - a} = \frac{\psi(\delta) - \varphi(a)}{\delta - a} \geq \frac{\psi(\delta) - \psi(a)}{\delta - a} = \sigma \varphi'(0) > \varrho \varphi'(0).$$

Also ist t eine schwache Wolfe-Powell-Schrittweite. Wegen der strikten Ungleichung $\varphi'(t) > \varrho \varphi'(0)$ muss es noch eine offene Umgebung von t in $]a, \delta[$ geben, die nur Wolfe-Powell-Schrittweiten enthält. $\square$

Algorithmus 5 macht sich dieses Lemma zu Nutzen, indem es vorgegebene Startschrittweite τ_0 in einer Expansionsphase so lange streckt, bis ein $b = \tau$ gefunden wird, das (3.35) erfüllt. Wird auf dem Weg dorthin bereits eine Wolfe-Powell-Schrittweite gefunden kann auch einfach abgebrochen werden. Ist $b = \tau$ mit (3.35) gefunden, so setzt man $a = 0$ und beginnt eine Intervallschachtelung (Kontraktionsphase).

Die Wahl von $\tau \in [a + \lambda_1(b - a), b - \lambda_2(b - a)]$ kann man einfach durch $\tau = \frac{1}{2}(a + b)$ bewerkstelligen. Alternativ kann man ein Polynom χ dritten Grades bestimmen mit $\chi(a) = \varphi(a)$, $\chi'(a) = \varphi'(a)$, $\chi(b) = \varphi(b)$ und $\chi'(b) = \varphi'(b)$ und wählt einen kritischen Punkt von χ in $[a + \lambda_1(b - a), b - \lambda_2(b - a)]$ falls vorhanden und $\tau = \frac{1}{2}(a + b)$ sonst. Die Werte $\varphi(a)$, $\varphi'(a)$, $\varphi(b)$, $\varphi'(b)$ liegen zu diesem Zeitpunkt schon vor, da zuvor schon Armijo- und Krümmungsbedingung mit ihrer Hilfe überprüft werden mussten.

Algorithmus 5 kann noch dahingehend verbessert werden, dass Armijo und Krümmungsbedingung nicht doppelt überprüft werden und dass die (vergleichsweise teure) Überprüfung der Krümmungsbedingung bei den zusammengesetzten logischen Ausdrücken in Zeile 5 und 18 nur dann durchgeführt wird, wenn es unbedingt nötig ist, indem die Armijo-Bedingung immer zuerst überprüft wird.

Im Idealfall findet man noch eine kostengünstige Strategie, wie man τ_0 wählen kann, sodass es meistens die Wolfe-Bedingungen erfüllt.⁹ Dann benötigt **Algorithmus 5** je nur eine Auswertung von $F(\Gamma_{x_k}(\tau_0 u_k))$ und $DF(\Gamma_{x_k}(\tau_0 u_k))$. Wegen $x_{k+1} = \Gamma_{x_k}(\tau_0 u_k)$ kann man dann diese Werte für die nächste Iteration des allgemeinen Abstiegsalgorithmus **Algorithmus 1** nutzen; d. h. **Algorithmus 5** fungiert dann nur als safe guard und schlägt ohne zusätzliche Kosten zu Buche.

Eine andere Frage ist, wie σ und ϱ zu wählen sind. Nocedal und Wright schlagen etwa $\sigma = 10^{-4}$ und $\varrho = 0.9$ bei Newton-artige Suchrichtungsstrategien vor (siehe [8]). Letztlich befinden wir uns bei solchen Parameterwahlen aber in einem Bereich der Optimierung, in der sie aufhört, eine exakte Wissenschaft zu sein, und dazu übergeht, eine Kunst zu sein.

⁹z. B. $\tau_0 = 1$ für Newton-artige Verfahren oder man versucht es mit der Schrittweite der vorherigen Iteration des allgemeinen Abstiegsalgorithmus.

Algorithm 5 Wolfe-Powell-Schrittweitenregel

Require: $\langle DF(x), u \rangle < 0$, $\varphi(t) := F(I_x(tu))$, $0 < \sigma < \varrho < 1$, $\lambda_1, \lambda_2 \in]0, 1/2[$, $S > 1$.

```

1: function WOLFEPOWELLSTEPSize( $x, u, F, DF, \tau_0$ )
2:    $\tau \leftarrow \tau_0$ 
3:   if  $\varphi(\tau) \leq \varphi(0) + \tau \sigma \varphi'(0)$  then
4:     while  $\varphi(\tau) \leq \varphi(0) + \tau \sigma \varphi'(0)$  and  $\varphi'(\tau) < \varrho \varphi'(0)$  do           ▷ Expansionsphase
5:        $\tau \leftarrow S \tau$ 
6:     end while
7:     if  $\varphi(\tau) \leq \varphi(0) + \tau \sigma \varphi'(0)$  then
8:       return  $\tau$            ▷ while-Schleife hat mit  $\varphi'(\tau) \geq \varrho \varphi'(0)$  terminiert
9:     end if
10:  end if
11:                                     ▷ Kontraktionsphase
12:   $a \leftarrow 0$            ▷  $a$  erfüllt (3.34)
13:   $b \leftarrow \tau$            ▷  $b$  erfüllt (3.35)
14:  choose  $c \in [a + \lambda_1(b - a), b - \lambda_2(b - a)]$ 
15:  while  $\varphi(c) > \varphi(0) + c \sigma \varphi'(0)$  or  $\varphi'(c) < \varrho \varphi'(0)$  do
16:    if  $\varphi(c) > \varphi(0) + c \sigma \varphi'(0)$  then
17:       $b \leftarrow c$            ▷  $b$  erfüllt (3.35)
18:    else
19:       $a \leftarrow c$            ▷  $a$  erfüllt (3.34)
20:    end if
21:    choose  $c \in [a + \lambda_1(b - a), b - \lambda_2(b - a)]$ 
22:  end while
23:  return  $c$ 
24: end function

```

3.7 Newton-artige Verfahren

Die bisher betrachteten Newton-Verfahren hatten die Form

$$x_{k+1} = x_k - t_k DDF(x_k)^{-1} (DF(x_k))$$

mit geeignet gewählten Schrittweiten $t_k \geq 0$. In der Praxis begegnet man häufig Zielfunktionen F deren zweite Ableitungen nur sehr teuer zu berechnen sind: Ist F hinreichend komplex, so kann schon die Implementierung äußerst schwierig und zeitaufwendig sein. Bei einem Optimierungsproblem im $\mathbb{R}^n$ ist $DDF(x)$ außerdem durch eine $(n \times n)$ -Matrix repräsentiert. Für die Berechnung der Einträge von $DDF(x)$ sind also in der Regel schon mindestens $O(n^2)$ Rechenoperationen nötig; für das Invertieren (bzw. zum Faktorisieren) schlagen dann schon $O(n^3)$ Rechenoperationen zu Buche. Einfache Gradientenverfahren benötigen aber nur $O(n)$ Operationen pro Iteration. Auf der anderen Seite benötigen Gradientenverfahren wie die Methode des steilsten Abstiegs oder die Konjugierte-Gradienten-Methode aber sehr viele Iterationen, weil sie auch bei optimalen Schrittweiten nur linear konvergieren mit einer Konstante, die empfindlich von der Konditionszahl von $D^2F(x_*)$ abhängt (siehe Übungen).

Newton-artige Verfahren versuchen daher, die Abbildungen $DDF(x_k)$ durch geschickt gewählte Abbildungen $B_k \in L(X; X')$ zu ersetzen und erzeugen Iterierte der Form

$$x_{k+1} = x_k - t_k B_k^{-1} (DF(x_k)).$$

Wir werden zwei Verfahren kennenlernen: Zum einen das Gauss-Newton-Verfahren für Ausgleichsprobleme, das seinen Vorteil aus der besonderen Form der Zielfunktion zieht. Zum anderen ein Quasi-Newton-Verfahren, das versucht, nützliche Information über $DDF(x_k)$ aus den bereits bekannten Iterierten $x_0, \dots, x_k$ und den Differentialen $DF(x_0), \dots, DF(x_k)$ zu ziehen.

In beiden Fällen ist $B_k = J_{b_k}$ eine Riesz-Abbildung bezüglich eines Skalarproduktes b_k .

Lemma 3.49 *Sei (X, g) ein Hilbertraum, $U \subset X$ eine offene Menge und $F: U \rightarrow \mathbb{R}$ stetig differenzierbar. Sei $(x_k)_{k \in \mathbb{N}}$ eine Folge von Punkten in U , $(b_k)_{k \in \mathbb{N}}$ eine Folge von symmetrischen, stetigen Bilinearformen und $(t_k)_{k \in \mathbb{N}}$ eine Folge von positiven Zahlen mit*

$$x_{k+1} = x_k + t_k u_k \quad \text{und} \quad u_k = -J_{b_k}^{-1} DF(x_k).$$

Gilt $\lambda g(u, u) \leq b_k(u, u) \leq \Lambda g(u, u)$ für alle $u \in X$ und alle $k \in \mathbb{N}$, so sind die Winkelbedingung und die Geschwindigkeitsbedingung aus Definition 3.2 erfüllt mit den Konstanten $\alpha = \frac{\lambda}{\Lambda}$ und $\beta = \frac{1}{\Lambda}$.

BEWEIS. Zunächst beobachten wir, dass die Geschwindigkeitsbedingung mit Konstante $\beta = \frac{1}{\Lambda}$ erfüllt ist:

$$\begin{aligned} \|DF(x_k)\| &= \sup_{v \in X \setminus \{0\}} \frac{\langle DF(x_k), v \rangle}{\|v\|} = \sup_{v \in X \setminus \{0\}} \frac{b_k(u_k, v)}{\|v\|} \leq \sup_{v \in X \setminus \{0\}} \frac{\sqrt{b_k(u_k, u_k)} \sqrt{b_k(v, v)}}{\|v\|} \\ &\leq \Lambda \sup_{v \in X \setminus \{0\}} \frac{\sqrt{g(u_k, u_k)} \sqrt{g(v, v)}}{\|v\|} = \Lambda \|u_k\|. \end{aligned}$$

Damit rechnen wir weiter:

$$-\langle DF(x_k), u_k \rangle = b_k(u_k, u_k) \geq \lambda g(u_k, u_k) = \lambda \|u_k\| \|u_k\| \geq \frac{\lambda}{\Lambda} \|DF(x_k)\| \|u_k\|.$$

Dies ist aber genau die Winkelbedingung mit Konstante $\alpha = \frac{\lambda}{\Lambda}$. □

Folgerung 3.50 Sei (X, g_0) ein Hilbertraum und $F: X \rightarrow \mathbb{R}$ stetig differenzierbar. Sei $g: X \rightarrow L^2(X; \mathbb{R})$ eine Riemannsche Metrik, d. h. eine stetige Familie von Skalarprodukten und es gebe $0 < \lambda \leq \Lambda < \infty$ mit $\lambda g_0(u, u) \leq g(x)(u, u) \leq \Lambda g_0(u, u)$ für alle $x \in X$ und alle $u \in X$. Dann erfüllt die Methode des steilsten Abstiegs bezüglich g gegeben durch

$$x_{k+1} = x_k + t_k u_k \quad \text{und} \quad u_k = -J_{g(x_k)}^{-1} DF(x_k)$$

jeweils Winkel- und Geschwindigkeitsbedingungen. Entsteht zusätzlich $t_k = \tau(x_k, u_k)$ einer effizienten Schrittweitenregel τ , dann liegt globale Konvergenz vor (siehe [Satz 3.4](#)).

3.8 Gauß-Newton-Verfahren

In diesem Abschnitt behandeln wir allgemeine Ausgleichprobleme, das sind Optimierungsprobleme mit Zielfunktionen der Form

$$F(x) := \frac{1}{2} \|\Phi(x) - y\|^2,$$

wobei eine möglicherweise nichtlineare, stetig differenzierbare Abbildung $\Phi: X \rightarrow Y$ in einen Hilbertraum (Y, h) und $y \in Y$ gegeben sind. Ein Minimum löst die möglicherweise gar nicht lösbare Gleichung $\Phi(x) = b$ “so gut es eben geht”, das heißt, der Punkt $\Phi(x)$ hat minimalen Abstand zu y unter allen Punkten im Bild von Φ . Der Faktor $\frac{1}{2}$ in der Zielfunktion ist rein kosmetischer Natur und ändert nichts an den Minimierern.

Solche Optimierungsprobleme nennt man *Ausgleichsprobleme*. Sie kommen insbesondere dort vor, wo man empirische, fehlerbehaftete Messungen an ein gegebenes Modell fitten möchte; also insbesondere in den empirischen Naturwissenschaften. Wenn die Messfehler in etwa einer Normalverteilung folgen, dann sind die Kleinsten-Quadrate-Schätzer gerade die Maximum-Likelihood-Schätzer.

3.8.1 Anwendungen und Motivation

Beispiel 3.51 Seien T eine Menge, (X, g) , (Y, h) Hilberträume, und $U \subset X$ eine offene Menge. Ein *Modell* ist im Wesentlichen eine Abbildung der Form $f: T \times U \rightarrow Y$, wobei wir U als *Parameterbereich* bezeichnen. Zu jeder Parameterwahl $x \in U$ erhält man ein quantifiziertes Modell $f_x: T \rightarrow Y$. So kann f für ein physikalisches Gesetz stehen: Bei gegebenen Parametern (z. B. physikalischen Naturkonstanten) $x \in U$ macht es für gewisse Eingangsdaten $t \in T$ Voraussagen über die Ausgangsdaten $y = f_x(t)$ eines physikalischen Vorgangs (z. B. eines Experiments). Oft trifft man auf die Situation, dass die Form des Modells f bekannt ist (z. B. aus theoretischen Überlegungen heraus), die Parameter $x \in U$ aber durch empirische Experimente ermittelt werden müssen.

Sei x_+ der tatsächliche, zunächst noch unbekannte Parameterwert. Führt man nun $N \in \mathbb{N}$ Versuche mit Eingangsdaten $t_1, \dots, t_N \in U$ durch, so erhält man Messungen $y_1, \dots, y_N \in Y$. Gibt es keine Messfehler, dann hat man also

$$y_i = f_{x_+}(t_i) \quad \text{für } i \in \{1, \dots, N\}.$$

Dann ist x_+ sicherlich ein Minimierer der Funktion

$$F: U \rightarrow Y, \quad F(x) := \frac{1}{2} \sum_{i=1}^N \|f_x(t_i) - y_i\|^2 = \frac{1}{2} \|\Phi(x) - y\|^2,$$

wobei $\Phi(x) := (f_x(t_1), \dots, f_x(t_N))$ und $b := (y_1, \dots, y_N)$. Ist die Abbildung Φ sogar injektiv, so ist x_+ der einzige Minimierer von F . Wie man durch zählen der Freiheitsgrade leicht einsieht, muss dann im Fall $X = \mathbb{R}^n$ und $Y = \mathbb{R}^m$ aber $Nm \geq n$ gelten.

Beispiel 3.52 (Methode der kleinsten Quadrate)

In Wirklichkeit werden die Messungen aber immer fehlerbehaftet sein, das heißt man hat

$$y_i = f_{x_+}(t_i) + \varepsilon_i \quad \text{für } i \in \{1, \dots, N\}$$

mit zufällig verteilten Fehlern $\varepsilon_i \in Y$. Deswegen wird man sehr viel mehr Messungen ($Nm \gg n$) durchführen und versuchen, die zufälligen Fehler durch geeignete Ausgleichrechnung “herauszurechnen”, um einen zuverlässigen Schätzer für x_+ zu bekommen.

Von nun an seien $X = \mathbb{R}^n$ und $Y = \mathbb{R}^m$ und $\langle \cdot, \cdot \rangle$ bezeichne die jeweiligen Euklidischen Skalarprodukte. Gängige und oftmals vernünftige Annahme ist, dass $\varepsilon_1, \dots, \varepsilon_N$ Realisierungen von $\mathbb{R}^m$ -wertigen, paarweise unabhängigen, identisch verteilten und normalverteilten Zufallsvariablen $Z_1, \dots, Z_N$ ist. Insbesondere heißt das, dass Z_i die folgende *Gauß-Glockenfunktion* als Wahrscheinlichkeitsdichtefunktion besitzt:

$$\varrho_i(z) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2} \langle (z - \mu_i), \Sigma^{-1}(z - \mu_i) \rangle\right).$$

Hierbei ist $\mu \in \mathbb{R}^m$ der gemeinsame *Erwartungswert* und Σ die gemeinsame *Kovarianzmatrix* (eine symmetrische und positiv-definite Abbildung). Für eine offene Menge $A \subset \mathbb{R}^m$ ist daher die Wahrscheinlichkeit $P(Z_i \in A)$, einen Fehlerwert in der Menge A zu erhalten, gegeben durch $P(Z_i \in A) = \int_A \varrho_i(z) d\mathcal{L}_m(z)$, wobei $\mathcal{L}_m$ das Lebesgue-Maß bedeute.

Hat man weder bei der Modellierung noch bei den Messungen *systematische Fehler* gemacht, so werden die Z_i im Mittel verschwinden, d. h. $\mu = 0$. Außerdem verwenden wir das innere Produkt $h(z_1, z_2) := \langle z_1, \Sigma^{-1} z_2 \rangle$ für $z_1, z_2 \in Y$ und schreiben $\|z\|_h := \sqrt{h(z, z)}$. Die Dichtefunktionen sehen dann so aus:

$$\varrho_i(z) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2} \|z\|_h^2\right).$$

Wegen der stochastischen Unabhängigkeit ist die gemeinsame Dichtefunktion von $Z := (Z_1, \dots, Z_N)$ am Punkt $z := (z_1, \dots, z_N) \in (\mathbb{R}^m)^N$ gegeben durch

$$\varrho(z) = \prod_{i=1}^N \varrho_i(z_i).$$

Unter der Annahme, dass x der richtige Parameterwert ist, ist die *Likelihood*, genau das Ergebnis $y := (y_1, \dots, y_N)$ zu erhalten, gegeben durch

$$L(x) = \varrho\left((y_1, \dots, y_N) - (f_x(t_1), \dots, f_x(t_N))\right) = \prod_{i=1}^N \varrho_i(y_i - f_x(t_i)).$$

Logarithmiert man dies, so erhält man als *Log-Likelihood-Funktion* $\Lambda(x)$ gerade

$$\Lambda(x) = \log(L(x)) = -\frac{1}{2} \sum_{i=1}^N \log(\det(2\pi\Sigma)) - \underbrace{\frac{1}{2} \sum_{i=1}^N \|y_i - f_x(t_i)\|_h^2}_{F(x)}.$$

Da die erste Summe konstant ist, maximiert man die Likelihood, indem man F minimiert. Das heißt, Minima von F sind *Maximum-Likelihood-Schätzer* für x_+ . Im Allgemeinen wird man Σ aus empirischen Daten heraus, z. B. durch Messwiederholungen, schätzen müssen. Im Fall $m = 1$ vereinfacht sich $\Lambda(x)$ aber zu

$$\Lambda(x) = \log(L(x)) = -\frac{1}{2} \sum_{i=1}^N \log(2\pi\Sigma) - \underbrace{\frac{1}{\Sigma^2} \frac{1}{2} \sum_{i=1}^N |y_i - f_x(t_i)|^2}_{F(x)},$$

d. h., die Zielfunktion $F(x) = \frac{1}{2} \sum_{i=1}^N |y_i - f_x(t_i)|^2$ kann unabhängig von der Kovarianzmatrix gewählt werden. In diesem Fall ist auch leicht ersichtlich, warum diese Methode der Ausgleichsrechnung klassisch als *Methode der kleinsten Quadrate* bezeichnet wird.

Bemerkung 3.53 Für einen historischen Abriss der Methode der kleinsten Quadrate sei verwiesen auf [7].

3.8.2 Konvergenzanalyse

Kommen wir nun auf das allgemeine Ausgleichproblem

$$F(x) := \frac{1}{2} \|\Phi(x) - y\|^2$$

zurück. In den Beispielen haben wir gesehen, dass im Falle endlichdimensionaler Räume X und Y in der Praxis oft $\dim(X) \gg \dim(Y)$ gilt. Das Newton-Verfahren direkt für die Gleichung $\Phi(x) = y$ zu benutzen, fällt also flach. Wir können aber versuchen, unsere bisher erlernten Optimierungsmethoden hier einzusetzen. Wir nehmen an, dass Φ mindestens zweimal stetig differenzierbar ist. Für $x \in U$ und $u \in X$ rechnen wir recht schnell nach:

$$DF(x)u = h(\Phi(x) - y, D\Phi(x)u) = g(D\Phi(x)^*(\Phi(x) - y), u).$$

Inbesondere gilt $\text{grad}_g(F)(x) = D\Phi(x)^*(\Phi(x) - y)$. Für $u, v \in X$ erhalten wir für die zweite Ableitung von F :

$$D^2F(x)(u, v) = h(D\Phi(x)v, D\Phi(x)u) + h(\Phi(x) - y, D^2\Phi(x)(u, v)).$$

Wir sehen, dass wir auch die zweite Ableitung $D^2\Phi(x)$ benötigen. Selbst, wenn wir es kennen würden, wäre noch nicht garantiert, dass die Bilinearform $D^2F(x)$ koerziv ist (dass also $DDF(x): X \rightarrow X'$ invertierbar wäre). Für jeden Punkt $x_+ \in X$ mit $\Phi(x_+) = y$, gilt aber

$$D^2F(x_+)(u, v) = h(D\Phi(x_+)v, D\Phi(x_+)u).$$

Wir führen daher die punktabhängige Bilinearform

$$b(x)(u, v) := h(D\Phi(x)v, D\Phi(x)u) \tag{3.36}$$

ein. Diese ist wenigstens positiv-semidefinit. Gilt zusätzlich noch $\sqrt{\lambda} \|u\| \leq \|D\Phi(x)u\| \leq \sqrt{\Lambda} \|u\|$ für alle $x \in X$ und $u \in X$, so gilt $\lambda g(u, u) \leq b(x)(u, u) \leq \Lambda g(u, u)$ und die Abbildung $B(x) := J_{b(x)}: X \rightarrow X'$ ist für alle $x \in X$ stetig invertierbar. Statt den negativen Gradienten $-\text{grad}_g(F)(x)$ bezüglich g können wir auch

$$-\text{grad}_b(F)(x) := -J_{b(x)}^{-1}(DF(x)) = -B(x)^{-1}(DF(x))$$

als Abstiegsrichtung verwenden. Das (gedämpfte) *Gauß-Newton-Verfahren* besteht nun eben in der Methode des steilsten Abstiegs bezüglich b , so dass man Iterierte von der Form

$$x_{k+1} = x_k - t_k \text{grad}_b(F)(x_k) \quad (3.37)$$

erhält. Aus (3.36) erhalten wir ohne Mühe die Identität $J_{b(x)} = J_g D\Phi(x)^* D\Phi(x)$ und daher gilt

$$\begin{aligned} \text{grad}_b(F)(x) &= J_{b(x)}^{-1} DF(x) \\ &= (D\Phi(x)^* D\Phi(x))^{-1} J_g^{-1} J_g (DF(x)) \\ &= (D\Phi(x)^* D\Phi(x))^{-1} D\Phi(x)^* (D\Phi(x) - y) = D\Phi(x)^\dagger (\Phi(x) - y). \end{aligned}$$

Die Abbildung $D\Phi(x)^\dagger := (D\Phi(x)^* D\Phi(x))^{-1} D\Phi(x)^*$ ist auch als *Moore-Penrose-Pseudoinverse* von $D\Phi(x)$ bekannt.¹⁰ Schreiben wir (3.37) damit um, so erhalten wir

$$x_{k+1} = x_k - t_k D\Phi(x)^\dagger (\Phi(x) - y).$$

Das sieht den Iterierten eines (gedämpften) Newton-Verfahrens für die Gleichung $\Phi(x) - y = 0$ verdächtig ähnlich; die (nicht existente) Inverse von $D\Phi(x)$ wurde lediglich durch die Pseudoinverse $D\Phi(x)^\dagger$ ersetzt.

Folgerung 3.54 (Globale Konvergenz des Gauß-Newton-Verfahrens)

Seien (X, g) , (Y, h) Hilberträume, $\Phi \in C^1(X; Y)$ eine Abbildung mit $\lim_{\|x\| \rightarrow \infty} \|\Phi(x)\| = \infty$. Für alle $r > 0$ gebe es $0 < \lambda(r) \leq \Lambda(r) < \infty$ und $L(r) \geq 0$ mit

1. $\sqrt{\lambda(r)} \|u\| \leq \|D\Phi(x)u\| \leq \sqrt{\Lambda(r)} \|u\|$ für alle $x \in B(0; r)$ und $u \in X$.
2. $\|D\Phi(z_1) - D\Phi(z_2)\| \leq L(r) \|z_1 - z_2\|$ für alle $z_1, z_2 \in B(0; r)$.

Sei $x_0 \in X$ beliebig und $(x_k)_{k \in \mathbb{N}}$ eine Folge, die durch das Gauß-Newton-Verfahren mit Schrittweitenregel [Algorithmus 4](#) oder [Algorithmus 5](#) erzeugt wird.

Dann gilt $\lim_{k \rightarrow \infty} \|DF(x_k)\| = 0$ und jeder Häufungspunkt ist ein kritischer Punkt von F . Ist X endlichdimensional, so gibt es auch mindestens einen Häufungspunkt.

BEWEIS. Sei $x_0 \in X$. Wegen $\lim_{\|x\| \rightarrow \infty} \|\Phi(x)\| = \infty$ ist die Subniveaumengen $L_c^-(F)$ für $c = F(x_0)$ beschränkt. Es gibt also ein $r > 0$ mit $L_c^-(F) \subset \bar{B}(0; r)$. Für alle $x \in \bar{B}(0; r)$ ist aber $D\Phi(x)^\dagger$ wohldefiniert (s. o.). Liegt x_k im Ball $\bar{B}(0; r)$, so sorgt die Armijo-Bedingung dafür,

¹⁰Mehr dazu in den Übungen.

dass $x_{k+1} \in L_c^-(F) \subset \bar{B}(0; r)$. Die Folge $(x_k)_{k \in \mathbb{N}}$ ist also wohldefiniert und beschränkt. Ist X endlichdimensional, so folgt die Existenz eines Häufungspunktes direkt aus dem Satz von Weierstrass und dem Satz von Heine-Borel (siehe [Satz 2.5](#) und [Satz 2.6](#)).

Für $\lim_{k \rightarrow \infty} \|DF(x_k)\| = 0$ verwenden wir [Satz 3.4](#). Aus [Lemma 3.49](#) folgt, dass das Gauß-Newton-Verfahren Abstiegsrichtungen erzeugt, die sowohl die Winkelbedingung (3.2) als auch die Geschwindigkeitsbedingung (3.3) erfüllen. Weil F durch 0 von unten beschränkt ist, müssen wir nur noch die Effizienzbedingung (3.4) nachweisen. Dazu erinnern wir uns an [Satz 3.44](#) und [Satz 3.46](#); ihretwegen brauchen wir nur die Lipschitzbedingung (3.26) nachweisen. Unsere Aufdatierungsabbildung ist hier $\Gamma_x(v) = x + v$. Wegen [Bemerkung 3.40](#) reicht es daher, zu zeigen, dass DF Lipschitz-stetig ist. Das folgt aber aus folgender Rechnung für alle $z_1, z_2 \in L_c^-(F)$:

$$\begin{aligned} \|DF(z_1) - DF(z_2)\| &= \|D\Phi(z_1)^*(\Phi(z_1) - y) - D\Phi(z_2)^*(\Phi(z_2) - y)\| \\ &\leq \|D\Phi(z_1)^*(\Phi(z_1) - \Phi(z_2))\| + \|(D\Phi(z_1) - D\Phi(z_2))^*(\Phi(z_2) - y)\| \\ &\leq \sqrt{\Lambda} \|\Phi(z_1) - \Phi(z_2)\| + L(r) \|z_1 - z_2\| \|\Phi(z_2) - y\| \\ &\leq \Lambda \|z_1 - z_2\| + L(r) \|z_1 - z_2\| \|\Phi(x_0) - y\| \\ &\leq (\Lambda + L(r) \|\Phi(x_0) - y\|) \|z_1 - z_2\|. \end{aligned}$$

Hierbei gilt $\|\Phi(z_1) - \Phi(z_2)\| \leq \sqrt{\Lambda} \|z_1 - z_2\|$ wegen des Mittelwertsatzes. $\square$

Satz 3.55 (Lokale Konvergenz des Gauß-Newton-Verfahrens)

Seien (X, g) , (Y, h) Hilberträume, $U \subset X$ eine offene Menge, $\Phi \in C^1(U; Y)$ eine Abbildung, so dass die Moore-Penrose-Pseudoinverse $D\Phi^\dagger: U \rightarrow L(Y; X)$, $z \mapsto D\Phi(z)^\dagger$ existent und stetig ist. Sei $x_* \in U$ ein Minimum von F und es gebe Konstanten $0 \leq c < 1$ und $L > 0$ so dass die folgenden Bedingungen erfüllt sind:

1. $\|D\Phi(z)^\dagger(\Phi(x_*) - y)\| \leq c \|z - x_*\|$ für alle $z \in U$.
2. $\|D\Phi(z_2)^\dagger(D\Phi(z_1) - D\Phi(z_2))\| \leq L \|z_1 - z_2\|$ für alle $z_1, z_2 \in U$.

Sei $r > 0$ mit $\frac{Lr}{2} + c < 1$ und $B(x_*; r) \subset U$. Dann gilt: Für jeden Startwert $x_0 \in B(x_*; r)$ konvergiert die Folge der ungedämpften Gauß-Newton-Iterierten

$$x_{k+1} = x_k - D\Phi(x_k)^\dagger(\Phi(x_k) - y)$$

gegen x_* und man hat die Fehlerabschätzungen

$$\|x_{k+1} - x_*\| \leq \frac{L}{2} \|x_k - x_*\|^2 + c \|x_k - x_*\| \quad \text{und} \quad \|x_k - x_*\| \leq \left(\frac{L\|x_0 - x_*\|}{2} + c \right)^k \|x_0 - x_*\|.$$

Gilt sogar $\Phi(x_*) = y$, so kann man $c = 0$ wählen und erhält quadratische Konvergenz wie beim Newtonverfahren:

$$\|x_{k+1} - x_*\| \leq \frac{L}{2} \|x_k - x_*\|^2.$$

BEWEIS. Sei $x \in U$. Wir definieren die Abbildung $\Theta(x) := x - D\Phi(x)^\dagger(\Phi(x) - y)$. Wegen $D\Phi(x)^\dagger D\Phi(x) = \text{id}_X$ können wir wie folgt rechnen:

$$\begin{aligned} \Theta(x) - x_* &= x - x_* - D\Phi(x)^\dagger(\Phi(x) - y) \\ &= x - x_* - D\Phi(x)^\dagger(\Phi(x) - \Phi(x_*)) - D\Phi(x)^\dagger(\Phi(x_*) - y) \\ &= D\Phi(x)^\dagger(\Phi(x_*) - \Phi(x) - D\Phi(x)(x_* - x)) - D\Phi(x)^\dagger(\Phi(x_*) - y) \\ &= \int_0^1 D\Phi(x)^\dagger(D\Phi(x + t(x_* - x)) - D\Phi(x))(x_* - x) dt - D\Phi(x)^\dagger(\Phi(x_*) - y). \end{aligned}$$

Mit der Dreiecksungleichung und mit Punkt 1 und Punkt 2 erhalten wir daher

$$\begin{aligned} \|\Theta(x) - x_*\| &\leq \|x_* - x\| \int_0^1 \|D\Phi(x)^\dagger(D\Phi(x + t(x_* - x)) - D\Phi(x))\| dt + c \|x - x_*\| \\ &\leq \frac{L}{2} \|x_* - x\|^2 + c \|x - x_*\|. \end{aligned}$$

Nun überlegt man sich leicht durch vollständige Induktion, dass wegen $\frac{Lr}{2} + c < 1$ die Folgenglieder x_k stets in $B(x_*; r)$ verbleiben müssen und die oben genannten Abschätzungen erfüllen. $\square$

3.8.3 Varianten

Ist $D\Phi$ die erste Ungleichung in Bedingung 1 von [Folgerung 3.54](#) verletzt, so kann man zur Bilinearform $b(x)$ aus (3.36) noch Terme dazu addieren. Einfachste Methode ist, ein Vielfaches des auf X verwendeten Skalarproduktes g zu addieren:

$$b_\alpha(x)(u, v) := h(D\Phi(x)u, D\Phi(x)v) + \alpha g(u, v) \quad \text{mit } \alpha > 0. \quad (3.38)$$

Man spricht in diesem Fall auch von *Tikhonov-Regularisierung* und nennt $\alpha > 0$ den *Regularisierungsparameter*. Als Iterierte des *regularisierten Gauß-Newton-Verfahrens* erhält man dann

$$x_{k+1} = x_k - t_k (D\Phi(x_k)^* D\Phi(x_k) + \alpha \text{id}_X)^{-1} D\Phi(x_k)^* (\Phi(x_k) - y).$$

Die (auch nur in Ausnahmefällen gegebene) quadratische Konvergenz aus [Satz 3.55](#) geht dabei natürlich verloren; globale Konvergenz aber zeigt man ganz analog wie in [Folgerung 3.54](#).

Bemerkung 3.56 Tikhonov-Regularisierung wirkt gewissermaßen wie ein Tiefpassfilter. Das sieht man in etwa so ein: Sind $0 < \sigma_1 \leq \sigma_2 \leq \sigma_3 \leq \dots$ die Singulärwerte von $A = D\Phi(x)$, so besitzt $A^\dagger$ die Singulärwerte $\dots \leq \frac{1}{\sigma_3} \leq \frac{1}{\sigma_2} \leq \frac{1}{\sigma_1}$. Sind die ersten Singulärwerte besonders klein, so werden Anteile von $\Phi(x) - y$, die zu diesen Singulärwerten gehören, übergebührlich verstärkt. Die Tikhonov-Regularisierung $(A^*A + \alpha \text{id}_X)^{-1}A^*$ hingegen hat die Singulärwerte $\dots \leq \frac{\sigma_3}{\sigma_3^2 + \alpha} \leq \frac{\sigma_2}{\sigma_2^2 + \alpha} \leq \frac{\sigma_1}{\sigma_1^2 + \alpha}$. Für im Vergleich zu α große σ_i gilt $\frac{\sigma_i}{\sigma_i^2 + \alpha} = \frac{1}{\sigma_i + \alpha/\sigma_i} \approx \frac{1}{\sigma_i}$; für Beiträge

bezüglich solcher Singulärwerte ändert sich also fast nichts. Für kleine σ_i dominiert aber der Regularisierungsparameter α . Das ist besonders interessant, wenn inexacte Messdaten $y = \Phi(x_+) + \varepsilon$ vorliegen und man Grund zu der Annahme hat, dass die Komponenten von ε ihre wesentlichen Beiträge für kleine Singulärwerte σ_i aufweisen.

Wählt man sogar den Regularisierungsparameter in jeder Iteration neu dann haben die Iterierten die Form

$$x_{k+1} = x_k - t_k (D\Phi(x_k)^* D\Phi(x_k) + \alpha_k \text{id}_X)^{-1} D\Phi(x_k)^* (\Phi(x_k) - y).$$

und man spricht von einem *Levenberg-Marquardt-Verfahren*.

Für die dazu zu treffenden Parameterwahlen gibt es einen ganzen Zoo eher heuristischer Methoden, die man als *educated guessing* bezeichnen kann. Das heißt nicht, dass diese Methoden schlecht wären, sondern dass der Nutzer dieser Methoden nicht umhin kommen wird, einige dieser Parameterwahlstrategien auszuprobieren und die für sein konkretes Problem erfolgreichste auszuwählen. Schließlich werden diese Methoden sehr erfolgreich bei Inversen Problemen¹¹, z. B. aus der Bildgebung eingesetzt.

3.8.4 Übungsaufgaben

Aufgabe 3.57 (Moore-Penrose-Pseudoinverse)

Seien (X, g) , (Y, h) endlichdimensionale Hilberträume und $A: X \rightarrow Y$ eine stetige, lineare Abbildung. Eine lineare Abbildung $A^\dagger: Y \rightarrow X$ nennt man *Moore-Penrose-Pseudoinverse*, wenn sie folgende Bedingungen erfüllt:

$$A A^\dagger A = A, \quad A^\dagger A A^\dagger = A^\dagger, \quad (A A^\dagger)^* = A A^\dagger \quad \text{und} \quad (A^\dagger A)^* = A^\dagger A.$$

Zeigen Sie:

1. Falls $A^\dagger$ existiert, so gilt $(A^*)^\dagger = (A^\dagger)^*$.
2. Falls $A^\dagger$ existiert, so ist $P = \text{id}_X - A^\dagger A$ ist der orthogonale Projektor auf $\ker(A)$.
3. Falls $A^\dagger$ existiert, so ist $Q = A A^\dagger$ ist der orthogonale Projektor auf $\text{im}(A)$.
4. Ist A injektiv, dann gilt $A^\dagger = (A^* A)^{-1} A^*$.
5. Ist A surjektiv, dann gilt $A^\dagger = A^* (A A^*)^{-1}$.

Bemerkung: Alle Aussagen bis auf [Punkt 4](#) sind auch für unendlichdimensionale Hilberträume richtig. [Punkt 4](#) bleibt auch korrekt, wenn man zusätzlich noch verlangt, dass das Bild von A abgeschlossen ist.¹² Dies ist aber etwas schwieriger. Einfacher wird es, wenn man voraussetzt, dass es ein $c > 0$ gibt, sodass $\|A u\| \geq c \|u\|$ für alle $u \in X$.¹³

¹¹Das sind eben Gleichungssysteme der Form $\Phi(x) = y$, bei denen Φ stetig-differenzierbar und injektiv ist, es aber kein $\lambda > 0$ gibt, so dass $\|D\Phi(x_*) u\| \geq \sqrt{\lambda} \|u\|$ gilt.

¹²Weil $\text{id}_Y - A A^\dagger$ gerade $\text{im}(A)$ als Kern besitzen muss, ist ein abgeschlossenes Bild generell notwendig für die Existenz einer Moore-Penrose-Pseudoinversen. Tatsächlich ist das auch hinreichend.

¹³Tatsächlich sind diese Aussagen äquivalent, siehe [Aufgabe 4.67](#). Um das einzusehen benötigt man allerdings den Satz von der offenen Abbildung [Folgerung 4.63](#) (siehe auch [Aufgabe 4.67](#)).

3.9 Inexakte Newton-Verfahren

Als nächstes möchten wir uns kurz überlegen, ob wir überhaupt eine Chance haben, bessere als lineare Konvergenzraten zu erreichen, ohne das exakte Newton-Verfahren zu benutzen.

Lemma 3.58 (Superlineare Konvergenz)

Seien X und Y Banachräume, $U \subset X$ eine offene Menge, $\Phi: U \rightarrow Y$ eine stetig differenzierbare Abbildung. An der Nullstelle $x_* \in X$ von Φ sei $D\Phi(x_*)$ stetig invertierbar und in einer Umgebung von x_* sei $D\Phi$ Lipschitz-stetig. Seien $(x_k)_{k \in \mathbb{N}_0}$ eine Folge in U , $(t_k)_{k \in \mathbb{N}_0}$ eine Folge positiver Zahlen und $(u_k)_{k \in \mathbb{N}_0}$ eine Folge von Suchrichtungen, die folgende Bedingungen erfüllen:

1. Für jedes $k \in \mathbb{N}_0$ gilt $x_{k+1} = x_k + t_k u_k$.
2. Die Folge $(x_k)_{k \in \mathbb{N}_0}$ konvergiert gegen x_* .
3. Die Folge $(t_k)_{k \in \mathbb{N}_0}$ konvergiert gegen 1.
4. Es gelte $c_k := \frac{\|u_k - u_k^N\|}{\|u_k\|} \xrightarrow{k \rightarrow \infty} 0$ mit der Newton-Suchrichtung $u_k^N := -D\Phi(x_k)^{-1} \Phi(x_k)$.

Dann konvergiert $(x_k)_{k \in \mathbb{N}}$ sogar superlinear gegen x_* , d. h., es gilt $\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} = 0$.

BEWEIS. Da $D\Phi(x_*)$ invertierbar ist und da $D\Phi$ Lipschitz-stetig in einer Umgebung von x_* ist, findet man ein $r > 0$ und ein $L \geq 0$, sodass $B(x_*; r) \subset U$ gilt und so dass

$$\|D\Phi(x)^{-1}(D\Phi(y) - D\Phi(x))\| \leq L \|y - x\|$$

gilt für alle $x, y \in B(x_*; r)$ und alle $u \in X$.

Ohne Einschränkung nehmen wir an dass $\|x_k - x_*\| < r$ gilt. Die Dreiecksungleichung impliziert

$$\begin{aligned} \|x_{k+1} - x_*\| &= \|x_k + t_k u_k - x_*\| \leq \|x_k + u_k^N - x_*\| + \|u_k - u_k^N\| + |t_k - 1| \|u_k\| \\ &= \|x_k + u_k^N - x_*\| + (c_k + |t_k - 1|) \|u_k\|. \end{aligned}$$

Für den ersten Term auf der rechten Seite rechnen wir

$$\begin{aligned} x_k + u_k^N - x_* &= D\Phi(x_k)^{-1}(\Phi(x_*) - \Phi(x_k) - D\Phi(x_k)(x_* - x_k)) \\ &= \int_0^1 D\Phi(x_k)^{-1} (D\Phi(x_k + s(x_* - x_k)) - D\Phi(x_k)) (x_* - x_k) ds \end{aligned}$$

und erhalten daher

$$\|x_k + u_k^N - x_*\| \leq L \|x_* - x_k\|^2 \quad \text{und} \quad \|u_k^N\| \leq (1 + Lr) \|x_k - x_*\|.$$

Aus Letzterem leiten wir ab

$$\|u_k\| \leq \|u_k^N\| + \|u_k - u_k^N\| \leq (1 + Lr) \|x_k - x_*\| + c_k \|u_k\|.$$

Wegen $c_k \rightarrow 0$ ergibt dies

$$\|u_k\| \leq \frac{1}{1-c_k} (1 + Lr) \|x_k - x_*\| \quad (3.39)$$

für hinreichend große k . Das setzen wir oben ein und erhalten

$$\|x_{k+1} - x_*\| \leq \left(L\|x_* - x_k\| + \frac{c_k + |t_k - 1|}{1 - c_k} (1 + Lr) \right) \|x_k - x_*\|, \quad (3.40)$$

was schließlich zu $\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} = 0$ führt. $\square$

Folgerung 3.59 Man wähle $t_k = 1$ und eine monoton steigende und bei 0 stetige Funktion $\omega: [0, \infty[\rightarrow [0, \infty[$ mit $\omega(0) = 0$, $\omega(\|u_k\|) \leq \frac{1}{2}$. Man bestimme u_k so dass

$$\|u_k - u_k^N\| \leq \|u_k\| \omega(\|u_k\|). \quad (3.41)$$

Für späte k gilt wegen (3.39) $\|u_k\| \leq 2(1 + Lr) \|x_k - x_*\|$ und man erhält die Konvergenzrate

$$\|x_{k+1} - x_*\| \leq (L\|x_* - x_k\| + 2(1 + Lr) \omega(2(1 + Lr) \|x_k - x_*\|)) \|x_k - x_*\|.$$

Ist insbesondere $\omega(s) = \min(\alpha s, \frac{1}{2})$ im Wesentlichen linear, so hat man wieder quadratische Konvergenz:

$$\|x_{k+1} - x_*\| \leq (L + 4\alpha(1 + Lr)^2) \|x_k - x_*\|^2.$$

Bemerkung 3.60 Hieran sehen wir auch, wie wichtig es ist, bei Newton-artigen Verfahren Schrittweitenregeln zu verwenden, die zumindest gegen Ende Schrittweiten der Länge 1 auswählen: Passt man nicht darauf auf, so wird das Verfahrens möglicherweise allein durch die ungenügende Schrittweitenregel abgebremst. Dass für späte k aber die Schrittweite 1 möglich ist, folgt im Wesentlichen aus den Erkenntnissen des Satzes von Kantorovich (siehe Satz 3.27) und aus $u_k \approx u_k^N$.

Beispiel 3.61 Verwendet man iterative lineare Löser, um eine approximative Lösung u_k der linearen Gleichung $D\Phi(x_k) u_k^N = -\Phi(x_k)$ zu erhalten und hat man eine obere Schranke für $\|D\Phi(x_k)^{-1}\| \leq C$, so kann man die Abschätzung

$$\frac{\|u_k - u_k^N\|}{\|u_k\|} \leq \frac{\|D\Phi(x_k)^{-1}\| \|D\Phi(x_k) (u_k - u_k^N)\|}{\|u_k\|} \leq C \frac{\|D\Phi(x_k) u_k + \Phi(x_k)\|}{\|u_k\|}$$

verwenden, um die Güte der Approximation $\|u_k - u_k^N\|$ zu kontrollieren. Man beachte, dass viele iterative lineare Löser gerade das Residuum $\|D\Phi(x_k) u_k + \Phi(x_k)\|$ und das relative Residuum $\frac{\|D\Phi(x_k) u_k + \Phi(x_k)\|}{\|u_k\|}$ für ihr Abbruchkriterium verwenden.

Solche iterativen Löser sind zum Beispiel das CG-Verfahren für selbstadjungiertes $D\Phi(x_k)$, GMRES oder BiCGStab. Sie benötigen in der Regel nur Matrix-Vektor-Produkte der Form $A v$ und $A^T w$ für Vektoren v und $w \in \mathbb{R}^n$. Kann man diese also effizient auswerten (und hat man einen passablen Vorkonditionierer für $D\Phi(x_k)$ zur Hand), so erhält man effiziente Lösungsverfahren mit adaptiv regelbarer Genauigkeit. Die zu $D\Phi(x_k)$ gehörende Matrix muss dabei nicht einmal assembliert werden, was zu so genannten *Matrix-freien Methoden* führt.

3.10 Quasi-Newton-Verfahren

Für die algorithmische Minimierung einer zweimal stetig differenzierbaren Funktion $F: U \subset X \rightarrow \mathbb{R}$ untersuchen wir hier *Quasi-Newton-Verfahren* von der Form

$$x_{k+1} = x_k - t_k B_k^{-1}(DF(x_k)).$$

Dabei ist $B_k = J_{b_k}$ stets der Riesz-Isomorphismus eines Skalarproduktes b_k . [Lemma 3.58](#) lehrt uns, dass wir für superlineare Konvergenz des Verfahrens “lediglich” $\lim_{k \rightarrow \infty} x_k = x_*$, $\lim_{k \rightarrow \infty} \frac{\|u_k - u_k^N\|}{\|u_k\|} = 0$ und $\lim_{k \rightarrow \infty} t_k = 1$ garantieren müssen.

Lemma 3.62 *Seien (X, g) ein Hilbertraum, $U \subset X$ eine offene Menge und $F: U \rightarrow \mathbb{R}$ eine zweimal stetig differenzierbare Abbildung. Sei $x_* \in U$ ein lokales Minimum von F , das die hinreichenden Bedingungen zweiter Ordnung erfüllt (siehe [Satz 2.23](#) und auf einer Umgebung von x_* sei DDF Lipschitz-stetig. Sei $(b_k)_{k \in \mathbb{N}_0}$ eine Folge von Skalarprodukten auf X mit Riesz-Isomorphismen $B_k := J_{b_k}$ und sei $(x_k)_{k \in \mathbb{N}_0}$ eine Folge von Punkten in U mit $\lim_{k \rightarrow \infty} x_k = x_*$. Für $u_k := -B_k^{-1}(DF(x_k))$ sind dann folgende Aussagen äquivalent*

1. $\lim_{k \rightarrow \infty} \frac{\|u_k - u_k^N\|}{\|u_k\|} = 0.$
2. $\lim_{k \rightarrow \infty} \frac{\|(B_k - DDF(x_*))u_k\|}{\|u_k\|} = 0.$

In diesem Fall gibt es ein $k_0 \in \mathbb{N}_0$ so dass für alle $k \geq k_0$ die Schrittweiten $t_k := 1$ die starken Wolfe-Powell-Bedingungen für $\sigma < \frac{1}{2}$ erfüllen.

BEWEIS. Seien $\lambda, \Lambda > 0$, so dass $\lambda \|u\|^2 \leq D^2F(x_*)(u, u) \leq \Lambda \|u\|^2$. Man findet $r > 0$ und $L \geq 0$, so dass für alle $x, y \in B(x_*; r)$ gelten:¹⁴

$$\|DDF(x) - DDF(y)\| \leq L \|x - y\| \quad \text{und} \quad \frac{\lambda}{2} \|u\|^2 \leq \|D^2F(x)(u, u)\| \leq 2\Lambda \|u\|^2$$

Für hinreichend große k gilt stets $x_k \in B(x_*; r)$. Zunächst rechnen wir

$$\begin{aligned} u_k - u_k^N &= (-B_k^{-1} + DDF(x_k)^{-1}) DF(x_k) \\ &= DDF(x_k)^{-1} (-DDF(x_k) + B_k) B_k^{-1} DF(x_k) \\ &= DDF(x_k)^{-1} (DDF(x_k) - B_k) u_k \\ &= DDF(x_k)^{-1} (DDF(x_k) - DDF(x_*)) u_k + DDF(x_k)^{-1} (DDF(x_*) - B_k) u_k. \end{aligned}$$

Einerseits erhalten wir nun

$$\begin{aligned} \|u_k - u_k^N\| &\leq \frac{2}{\lambda} \|(DDF(x_k) - DDF(x_*)) u_k\| + \frac{2}{\lambda} \|(DDF(x_*) - B_k) u_k\| \\ &\leq \frac{2L}{\lambda} \|x_k - x_*\| \|u_k\| + \frac{2}{\lambda} \|(DDF(x_*) - B_k) u_k\|. \end{aligned}$$

¹⁴Im Beweis von [Satz 2.23](#) haben wir schon einmal gezeigt, dass das geht.

Andererseits können wir auch wie folgt abschätzen:

$$\begin{aligned} \|(DDF(x_*) - B_k) u_k\| &\leq \|(DDF(x_k) - DDF(x_*)) u_k\| + \|DDF(x_k)(u_k - u_k^N)\| \\ &\leq L \|x_k - x_*\| \|u_k\| + 2\Lambda \|u_k - u_k^N\| \|u_k\|. \end{aligned}$$

Das liefert uns Äquivalenz von Punkt 1 und Punkt 2.

Die Aussage über die Wolfe-Powell-Bedingungen betreffend berechnen wir zunächst für $\varphi := \varphi_{x_k, u_k^N}$:

$$\varphi'(t) = DF(x_k + t u_k^N) u_k^N \quad \text{und} \quad \varphi''(t) = D^2F(x_k + t u_k^N)(u_k^N, u_k^N).$$

Wegen $\varphi'(0) + \varphi''(0) = -DF(x_k) u_k^N + DF(x_k) u_k^N = 0$ gilt

$$|\varphi'(1)| = |\varphi'(1) - \varphi'(0) - \varphi''(0)| \leq \left| \int_0^1 (\varphi''(s) - \varphi''(0)) ds \right| \leq \int_0^1 |\varphi''(s) - \varphi''(0)| ds \leq \frac{L}{2} \|u_k^N\|^3.$$

Außerdem haben $|\varphi'(0)| = -DF(x_k) u_k^N = D^2F(x_k)(u_k^N, u_k^N) \geq \frac{\lambda}{2} \|u_k^N\|^2$, so dass wir $\frac{|\varphi'(1)|}{|\varphi'(0)|} \leq \frac{L}{\lambda} \|u_k^N\|$ erhalten. Das ist für beliebiges, gegebenes $0 < \varrho < 1$ und hinreichend große k schon die starke Krümmungsbedingung $|\varphi'(1)| \leq \varrho |\varphi'(0)|$.

Analog erhalten wir

$$\varphi(t) = \varphi(0) + t \varphi'(0) + \frac{t^2}{2} \varphi''(0) + O(t^3 \|u_k^N\|^3) = \varphi(0) + t(1 - \frac{t}{2}) \varphi'(0) + O(t^3 \|u_k^N\|^3).$$

Hieraus folgt für $\sigma < \frac{1}{2}$ und hinreichend großes k die Armijo-Bedingung:

$$\varphi(1) = \varphi(0) + \frac{1}{2} \varphi'(0) + O(\|u_k^N\|^3) \leq \varphi(0) + \sigma \varphi'(0). \quad \square$$

Wir brauchen mit den B_k die Abbildung $DDF(x_*)$ also nur in Richtung später Suchrichtungen, also nur *entlang* des Suchpfades, zu approximieren. Das ist zum Glück eine deutlich schwächere Bedingung als $\lim_{k \rightarrow \infty} \|DDF(x_*) - B_k\| = 0$. Mit einer Schrittweitenregel, die immer zunächst prüft, ob $t_k = 1$ die Wolfe-Powell-Bedingungen erfüllt, können wir dann auch $\lim_{k \rightarrow \infty} t_k = 1$ garantieren.

Könnten wir noch $\lambda \|u\|^2 \leq b_k(u, u) \leq \Lambda \|u\|^2$ für alle $u \in X$ und alle $k \in \mathbb{N}_0$ garantieren, so könnten wir direkt [Lemma 3.49](#) anwenden, um globale Konvergenz zu erhalten; mit hinreichenden Bedingungen zweiter Ordnung (siehe [Satz 2.23](#)) könnte man dann auch $\lim_{k \rightarrow \infty} x_k = x_*$ erzielen. In der Tat sind aber globale Konvergenzbeweise für Quasi-Newton-Verfahren etwas schwieriger gelagert und sehr technisch.¹⁵ Das bringt uns wenige neue Erkenntnisse. Deswegen wollen wir uns damit begnügen, einige Algorithmen kennenzulernen, die

$$\lim_{k \rightarrow \infty} \frac{\|(B_k - DDF(x_*))u_k\|}{\|u_k\|} = 0$$

erzwingen, indem sie entlang des Suchpfades eingesammelte Information über F verwenden, um den Schätzer B_k sukzessive zu verbessern.

¹⁵Zum Beispiel muss man die Winkelbedingung (3.2) abschwächen zur sogenannten *Zoutendijk-Bedingung*: $\sum_{k=1}^{\infty} \cos^2(\vartheta_k) = \infty$ mit $\cos(\vartheta_k) := -\frac{\langle DF(x_k), u_k \rangle}{\|DF(x_k)\| \|u_k\|}$.

3.10.1 Die Broyden-Klasse

Sei (X, g) ein Hilbertraum und sei $F: X \rightarrow \mathbb{R}$ eine zweimal stetig differenzierbare Funktion mit Lipschitz-stetiger zweiten Ableitung.

$$u_k := -B_k^{-1} DF(x_k) \quad \text{und} \quad x_{k+1} = x_k + t_k u_k$$

Wegen

$$DF(x_{k+1}) - DF(x_k) - DDF(x_{k+1})(x_{k+1} - x_k) = O(\|x_{k+1} - x_k\|^2)$$

wählt man B_{k+1} so, dass die *Sekantengleichung* $B_{k+1}(x_{k+1} - x_k) = DF(x_{k+1}) - DF(x_k)$ erfüllt ist. Mit $s_k := x_{k+1} - x_k$ und $\eta_k := DF(x_{k+1}) - DF(x_k)$ schreiben wir dies als

$$B_{k+1} s_k = \eta_k \quad (\text{Sekantengleichung})$$

Eine wesentliche Eigenschaft von $DDF(x_{k+1})$ ist aber auch, dass $D^2F(x_{k+1})$ eine symmetrische Bilinearform ist. Außerdem hätten wir gerne, dass b_{k+1} positiv-definit ist, um 1.) Invertierbarkeit und 2.) Abstiegsrichtungen zu garantieren. Dann muss aber notwendig $\langle \eta_k, s_k \rangle > 0$ gelten, was man durch geeignete Liniensuche erreichen kann (siehe [Aufgabe 3.72](#)). Wegen $DDF(x_{k+1}) - DDF(x_k) = O(\|x_{k+1} - x_k\|)$ versucht man, die Differenz $B_{k+1} - B_k$ so “klein” wie möglich zu machen. Was “klein” bedeutet, müssen wir noch herausarbeiten; je nach Begriff erhalten wir andere Formeln für B_{k+1} in Abhängigkeit von s_k , η_k und B_k .

Definition 3.63 Sei (X, g) ein Hilbertraum. Wir definieren den *Raum der Skalarprodukte* durch

$$\text{Met}(X) := \left\{ b \mid \begin{array}{l} b \text{ ist stetige, symmetrische Bilinearform auf } X, \\ J_b: X \rightarrow X' \text{ ist stetig invertierbar} \end{array} \right\}.$$

Insbesondere ist der Raum $\text{Met}(X)$ unabhängig von der Wahl von g .

Proposition 3.64 (Broyden-Klasse)

Sei X ein Hilbertraum, sei $b \in \text{Met}(X)$, $s \in X$ und $\eta \in X'$ mit $\langle \eta, s \rangle > 0$. Für $\varphi \in \mathbb{R}$ definieren wir

$$b^+ = b - \frac{(J_b s) \otimes (J_b s)}{b(s, s)} + \frac{\eta \otimes \eta}{\langle \eta, s \rangle} + \varphi b(s, s) \left(\frac{\eta}{\langle \eta, s \rangle} - \frac{(J_b s)}{b(s, s)} \right) \otimes \left(\frac{\eta}{\langle \eta, s \rangle} - \frac{(J_b s)}{b(s, s)} \right) \quad (3.42)$$

Es gilt

1. b^+ ist eine stetige, symmetrische Bilinearform.
2. b^+ erfüllt die Sekantengleichung $J_{b^+} s = \eta$.
3. $J_{b^+} u = J_b u$ für alle $u \in \ker(\eta) \cap \ker(J_b s) = \text{span}(\{J_b^{-1} \eta, s\})^{\perp_b}$.
4. Für $\varphi > \varphi_{\text{crit}}$ gilt $b^+ \in \text{Met}(X)$. Hierbei ist $\varphi_{\text{crit}} = \frac{\langle \eta, s \rangle^2}{\langle \eta, s \rangle^2 - b(s, s) \langle \eta, J_b^{-1} \eta \rangle} \leq 0$.

Außerdem ist jedes b^+ , dass Bedingungen 1 bis 3 erfüllt, von der Form (3.42).

BEWEIS. Wir kürzen ab $\sigma := J_b s$ und $\xi := \frac{1}{\langle \eta, s \rangle} \eta - \frac{1}{b(s, s)} \sigma$. Weiters benutzen wir $y := J_b^{-1} \eta$.

$$b^+ = b - \frac{1}{b(s, s)} \sigma \otimes \sigma + \frac{1}{\langle \eta, s \rangle} \eta \otimes \eta + \varphi b(s, s) \xi \otimes \xi$$

Stetigkeit folgt aus der Tatsache, dass es sich bei b^+ um eine endliche Summe stetiger Bilinearformen handelt. Auch die Symmetrie ist augenfällig. Das ist schon [Punkt 1](#). Für [Punkt 2](#) beobachten wir zunächst

$$\langle \xi, s \rangle = \frac{\langle \eta, s \rangle}{\langle \eta, s \rangle} - \frac{\langle \sigma, s \rangle}{b(s, s)} = \frac{\langle \eta, s \rangle}{\langle \eta, s \rangle} - \frac{b(s, s)}{b(s, s)} = 0.$$

Nun rechnen wir für beliebiges $v \in X$ nach

$$\langle J_{b^+} s, v \rangle = b^+(s, v) = b(s, v) - \frac{b(s, s)}{b(s, s)} b(s, v) + \frac{\langle \eta, s \rangle}{\langle \eta, s \rangle} \langle \eta, v \rangle + 0 = \langle \eta, v \rangle.$$

Also gilt $J_{b^+} s = \eta$. [Punkt 3](#) ist nun auch offensichtlich.

Für [Punkt 4](#) definieren wir $w := \frac{1}{b(y, s)} y - \frac{1}{b(s, s)} s = J_b^{-1} \xi$. Dann haben wir $\langle \sigma, w \rangle = b(w, s) = \langle \xi, s \rangle = 0$ und $Y := \text{span}(\{y, s\}) = \text{span}(\{w, s\})$. Ein allgemeines $v \in X$ können wir b -orthogonal zerlegen in

$$v = \alpha s + \beta w + u \quad \text{mit } u \in Y^{\perp_b} = \ker(\sigma) \cap \ker(\eta)$$

und wir erhalten

$$b^+(v, v) = \alpha^2 b^+(s, s) + 2\alpha\beta b^+(s, w) + \beta^2 b^+(w, w) + b(u, u).$$

Um zu überprüfen, ob b^+ in $\text{Met}(X)$ liegt, reicht es daher zu zeigen, dass b^+ auf dem endlich-dimensionalen Untervektorraum Y positiv definit ist. Bezüglich der b -Orthogonalbasis $\{s, w\}$ von Y betrachten wir daher die Gramsche Matrix

$$\mathbf{A} := \begin{pmatrix} b^+(s, s) & b^+(s, w) \\ b^+(w, s) & b^+(w, w) \end{pmatrix} = \begin{pmatrix} b(y, s) & b(y, w) \\ b(w, y) & b^+(w, w) \end{pmatrix}.$$

Wegen $\mathbf{A}_{11} > 0$ reicht es nach dem Hurwitz-Kriterium, diejenigen φ zu finden, für die $\det(\mathbf{A}) = b(y, s) b^+(w, w) - b(y, w)^2 > 0$ gilt. Wegen $\langle \sigma, w \rangle = 0$ haben wir

$$b^+(w, w) = b(w, w) + \frac{\langle \eta, w \rangle^2}{\langle \eta, s \rangle} + \varphi b(s, s) \langle \xi, w \rangle^2 = b(w, w) + \frac{b(y, w)^2}{b(y, s)} + \varphi b(s, s) b(w, w)^2$$

und deswegen gilt

$$\det(\mathbf{A}) = b(y, s) b(w, w) + \cancel{b(y, w)^2} + \varphi b(s, s) b(y, s) b(w, w)^2 - \cancel{b(y, w)^2}.$$

Dies ist positiv, falls $\varphi > \varphi_{\text{crit}} := -\frac{1}{b(s, s) b(w, w)}$ und mit $b(w, w) = \frac{b(y, y)}{b(y, s)^2} - \frac{1}{b(s, s)}$ erhalten wir

$$\varphi_{\text{crit}} = \frac{b(y, s)^2}{b(y, s)^2 - b(s, s) b(y, y)} \leq 0. \quad \square$$

Die Mitglieder der Broyden-Klasse sind also in gewissem Sinne sehr ökonomisch: Sie erfüllen die Sekantengleichung, weichen von der Bilinearform b aber nur auf einem zweidimensionalen Unterraum ab.

3.10.2 Davidon-Fletcher-Powell-Update-Formel (DFP)

Das älteste der Quasi-Newton-Verfahren ist das von Davidon 1959 erfundene und von Fletcher und Powell analysierte und populär gemachte *DFP-Verfahren*. Dabei wählt man für das Update von b_k nach b_{k+1} dasjenige Mitglied der Broyden-Klasse mit $\varphi = 1$: Ist also $s_k := x_{k+1} - x_k$ und $\eta_k := DF(x_{k+1}) - DF(x_k)$, so wählt man mit $B_k := J_{b_k}$:

$$b_{k+1} = b_k - \frac{(B_k s_k) \otimes (B_k s_k)}{b_k(s_k, s_k)} + \frac{\eta_k \otimes \eta_k}{\langle \eta_k, s_k \rangle} + b_k(s_k, s_k) \left(\frac{\eta_k}{\langle \eta_k, s_k \rangle} - \frac{(B_k s_k)}{b_k(s_k, s_k)} \right) \otimes \left(\frac{\eta_k}{\langle \eta_k, s_k \rangle} - \frac{(B_k s_k)}{b_k(s_k, s_k)} \right)$$

Mit $\varrho_k := \frac{1}{\langle \eta_k, s_k \rangle}$ und dem kanonischen Isomorphismus $\psi_X: X \rightarrow X''$, können wir das auch etwas kompakter schreiben:

$$B_{k+1} = \left(\text{id}_{X'} - \varrho_k \eta_k \otimes \psi_X(s_k) \right) B_k \left(\text{id}_X - \varrho_k s_k \otimes \eta_k \right) + \varrho_k \eta_k \otimes \eta_k. \quad (\text{DFP})$$

Noch etwas bündiger wird es, wenn wir die Nennung des kanonischen Isomorphismus unterdrücken und für die Identitäten suggestiv eine 1 schreiben:

$$B_{k+1} = (1 - \varrho_k \eta_k \otimes s_k) B_k (1 - \varrho_k s_k \otimes \eta_k) + \varrho_k \eta_k \otimes \eta_k \quad (\text{DFP})$$

Für die Bestimmung der neuen Suchrichtung $u_{k+1} = -B_{k+1}^{-1}(DF(x_k))$ benötigen wir eigentlich die Inverse $H_{k+1} := B_{k+1}^{-1}$.

Diese lässt sich aber mit der *Sherman-Morrison-Formel* relativ leicht direkt durch $H_k := B_k^{-1}$ ausdrücken (siehe [Aufgabe 3.71](#)):

$$H_{k+1} = H_k - \frac{(H_k \eta_k) \otimes \psi_X(H_k \eta_k)}{\langle \eta_k, H_k \eta_k \rangle} + \varrho_k s_k \otimes \psi_X(s_k). \quad (\text{DFP})$$

in exakter Notation und etwas schlampiger, aber schöner

$$H_{k+1} = H_k - \frac{(H_k \eta_k) \otimes (H_k \eta_k)}{\langle \eta_k, H_k \eta_k \rangle} + \varrho_k s_k \otimes s_k. \quad (\text{DFP})$$

Lemma 3.65 (Sherman-Morrison-Formel)

Seien X, Y Banachräume, $A \in L(X; Y)$ stetig invertierbar, $\xi \in X'$ und $y \in Y$ mit $1 + \langle \xi, A^{-1} y \rangle \neq 0$. Dann ist auch $A + y \otimes \xi$ stetig invertierbar und die Inverse ist gegeben durch

$$(A + y \otimes \xi)^{-1} = A^{-1} - \frac{(A^{-1} y) \otimes (\xi \circ A^{-1})}{1 + \langle \xi, A^{-1} y \rangle}.$$

BEWEIS. Dies ist der Spezialfall der *Sherman-Morrison-Woodbury-Formel* (siehe [Aufgabe 3.70](#) unten) für $Z = \mathbb{R}$, $U = y$ und $V = \xi$. $\square$

In endlichdimensionalen Räumen (und nur in diesen) gibt es folgende Charakterisierung:

Proposition 3.66 (Charakterisierung des DFP-Updates)

Sei (X, g) ein endlichdimensionaler Hilbertraum. Seien $s \in X$ und $\eta \in X'$ mit $\langle \eta, s \rangle > 0$ und $g \in \text{Met}(X)$ mit $J_g s = \eta$. Für gegebenes $b \in \text{Met}(X)$ ist

$$J_{b^+} = (1 - \varrho \eta \otimes s) J_b (1 - \varrho_k s \otimes \eta) + \varrho \eta \otimes \eta \quad \text{mit } \varrho := \langle \eta, s \rangle^{-1}$$

die eindeutige Lösung des Optimierungsproblems

$$\text{Minimiere } \|J_{b^+} - J_b\|_g \quad \text{über alle } b^+ \in \text{Met}(X) \text{ mit } J_{b^+} s = \eta.$$

Hierbei bezeichnet $\|\cdot\|_g$ die Norm auf $L(X; X')$, die durch folgendes inneres Produkt induziert wird:

$$\langle A_1, A_2 \rangle_g := \text{tr}((J_g^{-1} A_1)^{*g} (J_g^{-1} A_2)) \quad \text{für } A_1, A_2 \in L(X; X').$$

BEWEIS. Tatsächlich ist J_{b^+} optimal unter allen lediglich symmetrischen Bilinearformen. Zuerst beobachtet man, dass

$$Y_\eta := \{a \in L^2(X; \mathbb{R}) \mid a \text{ ist symmetrisch, } J_a s = \eta\}$$

ein affiner Raum über dem Vektorraum

$$Y_0 := \{a \in L^2(X; \mathbb{R}) \mid a \text{ ist symmetrisch, } J_a s = 0\}$$

ist. Wegen [Folgerung 2.32](#) muss man nun noch lediglich nachrechnen, dass

$$J_{b^+} - J_b = \frac{\langle \eta, s \rangle + b(s, s)}{\langle \eta, s \rangle^2} \eta \otimes \eta - \frac{1}{\langle \eta, s \rangle} (\sigma \otimes \eta + \eta \otimes \sigma) \quad \text{mit } \sigma := J_b s$$

im orthogonalen Komplement $Y_0^\perp$ von Y_0 bezüglich $\langle \cdot, \cdot \rangle_g$ liegt. Dazu zeigen wir, dass

$$\{\eta \otimes \xi + \xi \otimes \eta \mid \xi \in X'\} \subset Y_0^{\perp g}$$

gilt.¹⁶ Sei $c = \eta \otimes \xi + \xi \otimes \eta$. Wir nutzen

$$(J_g^{-1} J_c)^{*g} = J_g^{-1} J_c \quad \text{und} \quad J_g^{-1} J_c = s \otimes \xi + (J_g^{-1} \xi) \otimes \eta.$$

und rechnen

$$\begin{aligned} \langle J_c, J_a \rangle_g &= \text{tr}((s \otimes \xi + (J_g^{-1} \xi) \otimes \eta) J_g^{-1} J_a) \\ &= \text{tr}(J_g^{-1} J_a (s \otimes \xi)) + \text{tr}(((J_g^{-1} \xi) \otimes \eta) J_g^{-1} J_a) = \text{tr}(((J_g^{-1} \xi) \otimes (\eta \circ J_g^{-1} J_a))). \end{aligned}$$

Schließlich beobachten wir noch für beliebiges $v \in X$:

$$\langle \eta, J_g^{-1} J_a v \rangle = g(J_g^{-1} \eta, J_g^{-1} J_a v) = \langle J_a v, s \rangle = a(v, s) = a(s, v) = \langle J_a s, v \rangle = 0.$$

Dies zeigt endlich $\langle J_c, J_a \rangle_g = 0$. □

¹⁶Tatsächlich gilt sogar $Y_0^{\perp g} = \{\eta \otimes \xi + \xi \otimes \eta \mid \xi \in X'\}$, aber das ist uns hier nicht so wichtig.

Bemerkung 3.67 Bei dem Skalarprodukt g darf man gerne an die gemittelte zweite Ableitung $g_k = \int_0^1 D^2 F(x_k + t(x_{k+1} - x_k)) dt$ denken. Wegen

$$\eta_k = DF(x_{k+1}) - DF(x_k) = \int_0^1 DDF(x_k + t(x_{k+1} - x_k))(x_{k+1} - x_k) dt = J_{g_k} s_k$$

erfüllt diese nämlich die Sekantengleichung.

Bemerkung 3.68 Auf unendlichdimensionalen Hilberträumen ist die Charakterisierung [Proposition 3.66](#) nicht ohne Weiteres anwendbar, denn nicht für alle $A \in L(X; X')$ ist $\|A\|_g$ endlich. Das sieht man schon an $A = J_g$; man hat

$$\langle A, A \rangle_g := \text{tr}((J_g^{-1} J_g)^* (J_g^{-1} J_g)) = \text{tr}(\text{id}_X) = \infty.$$

Vielmehr sind die Operatoren $A \in L(X; X')$ von der Form $A = J_g K$ mit einem Hilbert-Schmidt Operator $K \in L(X; X)$ und Hilbert-Schmidt-Operatoren sind stets kompakt. Die DFP-Updates haben nur endlichen Rang und sind deshalb auch Hilbert-Schmidt-Operatoren (modulo Multiplikation mit J_g^{-1}). Man kann also die DFP-Updates noch charakterisieren als das kleinste Hilbert-Schmidt-Update bezüglich der von g induzierten Hilbert-Schmidt-Norm.

3.10.3 Broyden-Fletcher-Goldfarb-Shanno-Update-Formel (BFGS)

Für unseren allgemeinen Abstiegsalgorithmus sind ja eigentlich die Abbildungen $H_k = J_{b_k}^{-1}$ entscheidend. Wendet man die Charakterisierung [Proposition 3.66](#) des DFP-Update analog auf die Inversen der Rieszabbildungen an, so kommt man zur *BFGS-Update-Formel*; sie entspricht dem Parameter $\varphi = 0$ in der Broyden-Klasse.

Proposition 3.69 (Charakterisierung des BFGS-Updates)

Sei (X, g) ein endlichdimensionaler Hilbertraum. Seien $s \in X$ und $\eta \in X'$ mit $\langle \eta, s \rangle > 0$ und $g \in \text{Met}(X)$ mit $J_g s = \eta$. Für gegebenes $b \in \text{Met}(X)$ ist

$$J_{b^+}^{-1} = (1 - \varrho s \otimes \eta) J_b^{-1} (1 - \varrho \eta \otimes s) + \varrho s \otimes s \quad \text{mit } \varrho := \langle \eta, s \rangle^{-1} \quad (3.43)$$

die eindeutige Lösung des Optimierungsproblems

$$\text{Minimiere } \|J_{b^+}^{-1} - J_b^{-1}\|_g \quad \text{über alle } b^+ \in \text{Met}(X) \text{ mit } J_{b^+} s = \eta.$$

Hierbei bezeichnet $\|\cdot\|_g$ die Norm auf $L(X'; X)$, die durch folgendes inneres Produkt induziert wird:

$$\langle A_1, A_2 \rangle_g := \text{tr}((A_1 J_g^{-1})^* (A_2 J_g^{-1})) \quad \text{für } A_1, A_2 \in L(X'; X).$$

Das *BFGS-Verfahren* besteht nun im Wesentlichen aus dem allgemeinen Abstiegsalgorithmus mit Suchrichtungen $u_k = -B_k^{-1}(DF(x_k)) = u_k = -H_k(DF(x_k))$, der BFGS-Update-Formel

$$H_{k+1} = (1 - \varrho_k s_k \otimes \eta_k) H_k (1 - \varrho_k \eta_k \otimes s_k) + \varrho_k s_k \otimes s_k \quad (\text{BFGS})$$

und einer Wolfe-Powell-Schrittweitenregel (letztere dient dazu, $\langle \eta_k, s_k \rangle > 0$ zu erzwingen, siehe [Aufgabe 3.72](#)). Auch, wenn sie eigentlich nicht gebraucht werden, kann man die B_k auch wieder mit der Sherman-Morrison-Formel erhalten:

$$B_{k+1} = B_k - \frac{(B_k s_k) \otimes (B_k s_k)}{b_k(s_k, s_k)} + \frac{\eta_k \otimes \eta_k}{\langle \eta_k, s_k \rangle}. \quad (\text{BFGS})$$

Das BFGS-Verfahren gilt als das erfolgreichste unter den bisher bekannten Verfahren der unrestringierten Optimierung glatter Funktionen. Weil es nur erste Ableitungen benötigt und fast ohne das Invertieren von Matrizen auskommt, ist es unter Umständen sogar in der Laufzeit schneller als das Newton-Verfahren. Da ein Abstieg immer garantiert wird, ist es auch robuster unter der Wahl “schlechter” Startbedingungen.

Aus diesen Gründen ist das BFGS-Verfahren Bestandteil vieler Optimierungs-Bibliotheken.¹⁷ Es ist so mächtig, dass man sogar versucht, restringierte Probleme auf unrestringierte zurückzuführen, um BFGS einsetzen zu können. Damit werden wir uns aber erst im nächsten Kapitel beschäftigen.

3.10.4 Implementierung

Für die Implementierung auf Computern muss man an Stelle von Bilinearformen ihre Gramsche Matrizen bezüglich einer gegebenen Basis verwenden. Selbstredend muss dann X ein endlichdimensionaler Hilbertraum sein. Ziel ist es, die Update-Formeln der Broyden-Klassen aus [Proposition 3.64](#) in eine numerische effiziente Form zu bringen.

Wir wählen dazu eine Basis $e = (e_1, \dots, e_n)$ von X . Sei $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ die dazu duale Basis von X' , d. h., wir haben $\langle \varepsilon_i, e_j \rangle = \delta_{ij}$ mit dem Kroneckerdelta δ_{ij} .

Ist g ein Skalarprodukt auf X , so ist die dazugehörige *Gramsche Matrix* bezüglich e gegeben durch

$$\mathbf{G}_{ij} := g(e_i, e_j).$$

Bezeichnenderweise gilt dann für den Riesz-Isomorphismus $\langle J_g e_i, e_j \rangle = g(e_i, e_j) = \mathbf{G}_{ij}$ und also

$$J_g e_i = \sum_{j=1}^n \mathbf{G}_{ij} \varepsilon_j \quad \text{und} \quad J_g^{-1} \varepsilon_j = \sum_{i=1}^n e_i (\mathbf{G}^{-1})_{ij}.$$

Elemente $u \in X$ und $\alpha \in X'$ können wir *beide* als *Spaltenvektoren* $\mathbf{u}$ und $\mathbf{a} \in \mathbb{R}^n$ repräsentieren mit $\mathbf{u}_i := \langle \varepsilon_i, u \rangle$ und $\mathbf{a}_j := \langle \alpha, e_j \rangle$, d. h., wir haben

$$u = \sum_{i=1}^n \mathbf{u}_i e_i \quad \text{und} \quad \alpha = \sum_{j=1}^n \mathbf{a}_j \varepsilon_j.$$

¹⁷Zum Beispiel enthält MATLAB sowohl das DFP- als auch das BFGS-Verfahren als backend für die Funktion `fminunc`.

Gilt $J_g u = \alpha$, so haben wir

$$\mathbf{a} = \mathbf{G} \mathbf{u} \quad \text{und} \quad \mathbf{u} = \mathbf{G}^{-1} \mathbf{a}.$$

Wir müssen uns dabei also immer bewusst machen, ob ein Spaltenvektor einen primalen Vektor oder einen dualen Vektor repräsentiert.

Seien $\mathbf{B}_k$ und $\mathbf{B}_{k+1}$ die Gramschen Matrizen von b_k bzw. b_{k+1} , $\mathbf{s}_k$ der zu s_k gehörende Spaltenvektor und $\mathbf{y}_k$ der zu η_k gehörende Spaltenvektor. Die Update-Formeln der Broyden-Klasse haben dann die Form

$$\mathbf{B}_{k+1} = \mathbf{B}_k - \frac{(\mathbf{B}_k \mathbf{s}_k)(\mathbf{B}_k \mathbf{s}_k)^\top}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^\top}{\mathbf{y}_k^\top \mathbf{s}_k} + \varphi_k (\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k) \left(\frac{\mathbf{y}_k}{\mathbf{y}_k^\top \mathbf{s}_k} - \frac{(\mathbf{B}_k \mathbf{s}_k)}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k} \right) \left(\frac{\mathbf{y}_k^\top}{\mathbf{y}_k^\top \mathbf{s}_k} - \frac{(\mathbf{B}_k \mathbf{s}_k)^\top}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k} \right).$$

Außerdem haben wir $\varrho_k = \frac{1}{\mathbf{y}_k^\top \mathbf{s}_k}$ und können die DFP- und BFGS-Update-Formeln in folgender Tabelle zusammenfassen:

	DFP-Update	BFGS-Update
$\mathbf{B}_{k+1}$	$(\mathbf{I} - \varrho_k \mathbf{y}_k \mathbf{s}_k^\top) \mathbf{B}_k (\mathbf{I} - \varrho_k \mathbf{s}_k \mathbf{y}_k^\top) + \varrho_k \mathbf{y}_k \mathbf{y}_k^\top$	$\mathbf{B}_k - \frac{(\mathbf{B}_k \mathbf{s}_k)(\mathbf{B}_k \mathbf{s}_k)^\top}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k} + \varrho_k \mathbf{y}_k \mathbf{y}_k^\top$
$\mathbf{H}_{k+1}$	$\mathbf{H}_k - \frac{(\mathbf{H}_k \mathbf{y}_k)(\mathbf{H}_k \mathbf{y}_k)^\top}{\mathbf{y}_k^\top \mathbf{H}_k \mathbf{y}_k} + \varrho_k \mathbf{s}_k \mathbf{s}_k^\top$	$(\mathbf{I} - \varrho_k \mathbf{s}_k \mathbf{y}_k^\top) \mathbf{H}_k (\mathbf{I} - \varrho_k \mathbf{y}_k \mathbf{s}_k^\top) + \varrho_k \mathbf{s}_k \mathbf{s}_k^\top$

Man beachte, dass $\mathbf{s}_k \mathbf{y}_k^\top$ eine $n \times n$ -Matrix ist und man dafür $O(n^2)$ Multiplikation (und eben soviel Speicher) benötigen würde, um diese Matrix aufzustellen. Jede Auswertung der Form $\mathbf{H}_k \mathbf{v}$ schlägt dann noch einmal mit $O(n^2)$ Rechenoperationen zu buche.

Andererseits kann man die Matrix-Vektor-Produkte wie folgt umschreiben $(\mathbf{s}_k \mathbf{y}_k^\top) \mathbf{v} = \mathbf{s}_k (\mathbf{y}_k^\top \mathbf{v})$ und kommt dann auf lediglich $O(n)$ Rechenoperationen und $O(n)$ Speicherplatz. In der Tat braucht man auch nur Matrix-Vektor-Produkte $\mathbf{H}_k \mathbf{v}$ für das BFGS-Verfahren. Geschickte Ausnutzung dieser Überlegung führt zum *L-BFGS-Verfahren*. Das “L” im Namen steht dabei für *limited memory*. Bei diesem speichert man lediglich die Vektoren $\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_{k-1}$, $\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{k-1}$ und die Zahlen $\varrho_0, \varrho_1, \varrho_2, \dots, \varrho_{k-1}$. $\mathbf{H}_k \mathbf{v}$ und verwendet einen Algorithmus wie [Algorithmus 6](#) zur Bestimmung der Suchrichtungen $\mathbf{u} = -\mathbf{H}_k (DF(x_k))$.¹⁸ Bei diesem werden für ein gegebenes $m \in \mathbb{N} \cup \{\infty\}$ nur die letzten $l := \min(k, m)$ Daten

$$\mathbf{y}_{k-l}, \dots, \mathbf{y}_{k-1}, \quad \mathbf{s}_{k-l}, \dots, \mathbf{s}_{k-1} \quad \text{und} \quad \varrho_{k-l}, \dots, \varrho_{k-1}$$

der Suchhistorie beibehalten; alle anderen werden einfach “vergessen”.¹⁹ Für $m = \infty$ benötigt man in der k -ten Iteration $O(nk)$ Speicher und $O(nk)$ Rechenoperationen, das heißt, der

¹⁸Man beachte, dass [Algorithmus 6](#) basisfrei notiert wurde.

¹⁹De-facto wird man die alten Daten durch die neuen Daten überschreiben, um so unnötige Speicherallokation und -deallokation zu vermeiden. Bei diesem Vorgehen benötigt man nur minimalen Mehraufwand, um die Lesereihenfolge aus dem Speicher festzulegen.

Aufwand pro Iteration wächst mit der Anzahl der benötigten Iteration an. In der Praxis ist es aber überhaupt nicht kontraproduktiv, $m < \infty$ (z. B. $m \in \{5, 30\}$) zu wählen: Am Anfang des L-BFGS-Verfahren werden in der Regel recht große Schritte gemacht; die Daten, die man dabei für die Approximation H_k von $DF(x_k)$ einsammelt, sind also recht schnell veraltet. So verwendet das L-BFGS-Verfahren nur die aktuellsten Informationen, um H_k zu approximieren und man spart überdies noch Speicher und Rechenzeit. Schließlich fassen wir das L-BFGS-Verfahren noch einmal in [Algorithmus 7](#) zusammen.

Algorithm 6 L-BFGS-Suchrichtung

Require: $m \in \mathbb{N} \cup \{\infty\}$, $l = \min(k, m)$, $\langle \eta_i, s_i \rangle > 0$, $\varrho_i = \langle \eta_i, s_i \rangle^{-1}$

```

1: function LBFGSSearchDirection( $x, DF, H_0, s_{k-l}, \dots, s_{k-1}, \eta_{k-l}, \dots, \eta_{k-1}, \varrho_{k-l}, \dots, \varrho_{k-1}$ )
2:    $\xi \leftarrow DF(x)$ 
3:   if  $l > 0$  then
4:     for  $i = k-1, k-2, \dots, k-l$  do
5:        $\alpha_i \leftarrow \rho_i \langle \xi, s_i \rangle$ 
6:        $\xi \leftarrow \xi - \alpha_i \eta_i$ 
7:     end for
8:   end if
9:    $r \leftarrow H_0 \xi$ 
10:  if  $l > 0$  then
11:    for  $i = k-l, k-l+1, \dots, k-1$  do
12:       $\beta \leftarrow \rho_i \langle \eta_i, r \rangle$ 
13:       $r \leftarrow r + (\alpha_i - \beta) s_i$ 
14:    end for
15:  end if
16:  return  $-r$ 
17: end function

```

Algorithm 7 L-BFGS-Verfahren

Require: $m \in \mathbb{N} \cup \{\infty\}$, $H_0 \in L(X'; X)$ symmetric and invertible

```

1: function LBFGSDESCENT( $x_0, F, DF$ )
2:    $k \leftarrow 0$ 
3:   while  $\neg \text{crit}(x_k, F, DF)$  do                                      $\triangleright$  Abbruchkriterium
4:      $u_k \leftarrow \text{LBFGSSearchDirection}(x_k, DF, H_0, \dots)$ 
5:      $t_k \leftarrow \text{WolfePowellStepSize}(x_k, u_k, F, DF, \tau_0 = 1)$         $\triangleright$  Wolfe-Powell Schrittweite
6:      $x_{k+1} \leftarrow x_k + t_k u_k$                                       $\triangleright$  Aufdatierung
7:     if  $k > m$  then
8:       discard  $s_{k-m}$  and  $\eta_{k-m}$  from storage                          $\triangleright$  Speicherbereinigung
9:     end if
10:     $s_k \leftarrow x_{k+1} - x_k$                                           $\triangleright$  Speichern der Suchhistorie
11:     $\eta_k \leftarrow DF(x_{k+1}) - DF(x_k)$ 
12:     $k \leftarrow k + 1$ 
13:  end while
14:  return  $x_k$ 
15: end function

```

3.10.5 Übungsaufgaben

Aufgabe 3.70 (Sherman-Morrison-Woodbury-Formel)

Seien X, Y und Z Banachräume, $A \in L(X; Y)$ stetig invertierbar, $U \in L(Z; Y)$ und $V \in L(X; Z)$ so dass $\text{id}_Z + V A^{-1} U$ stetig invertierbar ist. Dann ist auch $A + U V$ stetig invertierbar mit

$$(A + U V)^{-1} = A^{-1} - A^{-1} U (\text{id}_Z + V A^{-1} U)^{-1} V A^{-1}.$$

Aufgabe 3.71

Sei X ein Hilbertraum, $b \in \text{Met}(X)$, $s \in X$ und $\eta \in X'$ mit $\langle \eta, s \rangle > 0$. Für $B := J_b$ und $B^+ := J_{b^+}$ lautet die DFP-Update-Formel

$$B^+ = (1 - \varrho \eta \otimes s) B (1 - \varrho s \otimes \eta) + \varrho \eta \otimes \eta. \quad (\text{DFP})$$

Zeigen Sie für $H^+ := (B^+)^{-1}$ und $H := B^{-1}$ mit Hilfe der Sherman-Morrison-Formel die Identität

$$H^+ = H - \frac{(H \eta) \otimes (H \eta)}{\langle \eta, H \eta \rangle} + \varrho s \otimes s. \quad (\text{DFP})$$

Aufgabe 3.72

Sei u_k eine Abstiegsrichtung bei x_k und sei t_k eine Wolfe-Powell-Schrittweite bezüglich der Aufdatierungsabbildung $\Gamma_x(v) = x + v$. Zeigen Sie, dass dann $\langle \eta_k, s_k \rangle > 0$ gilt für $s_k := t_k x_k$ und $\eta_k := DF(x_k + t_k u) - DF(x_k)$.

Aufgabe 3.73

Implementieren Sie den L-BFGS-Algorithmus mit Wolfe-Powell-Schrittweiten und testen Sie ihn an der Rosenbrock-Funktion. Verwenden Sie für die Länge der Suchhistorie die Werte $m = 0$ (Gradientenabstieg), $m = 5$, $m = 10$, $m = 20$ und $m = 100 \approx \infty$.

Aufgabe 3.74

Sei (X, g) ein Hilbertraum. Auf dem Raum der Skalarprodukte

$$\text{Met}(X) := \left\{ b \mid \begin{array}{l} b \text{ ist stetige, symmetrische Bilinearform auf } X, \\ J_b: X \rightarrow X' \text{ ist stetig invertierbar} \end{array} \right\}$$

definieren wir die Abbildung $d_\infty: \text{Met}(X) \times \text{Met}(X) \rightarrow \mathbb{R} \cup \{\infty\}$ durch

$$d_\infty(b, h) := \log \inf \{ \Lambda \geq 1 \mid \forall u \in X: \Lambda^{-1} b(u, u) \leq h(u, u) \leq \Lambda b(u, u) \}.$$

1. Zeigen Sie, dass $d_\infty(b, h)$ stets endlich ist für beliebige $b, h \in \text{Met}(X)$.
2. Zeigen Sie, dass $(\text{Met}(X), d_\infty)$ ein metrischer Raum ist.
3. Zeigen Sie, dass $(\text{Met}(X), d_\infty)$ vollständig ist.

Hinweis: Führen sie [Punkt 3](#) durch ein Skalierungsargument und die Polarisationsformel auf den (bekannterweise) vollständigen Raum $(C_b^0(S; \mathbb{R}), \|\cdot\|_\infty)$ der beschränkten stetigen Funktionen auf der Einheitssphäre $S := \{ u \in X \mid g(u, u) = 1 \}$ mit der Supremumsnorm $\|f\|_\infty := \sup_{u \in S} \|f(u)\|$ zurück.

Aufgabe 3.75

Wir möchten die stabile Lage einer an ihren Enden befestigten Gummikette unter ihrer eigenen Last berechnen. Die Kette bestehe aus n identischen Gliedern, habe eine Ruhelänge $L > 0$ und eine Gesamtmasse $m > 0$. Die Lage der Kette können wir durch die Positionen $x^{(0)}, x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^3$ der Endpunkte ihrer Glieder beschreiben, wobei wir annehmen, dass jedes Glied zwar gestreckt, aber nicht verbogen werden kann. Die stabile Lage der Kette wird durch minimale Gesamtenergie charakterisiert. Die durch Streckung gespeicherte Energie des i -ten Kettengliedes betrage $\ell w(\|x^{(i)} - x^{(i-1)}\|/\ell)$ mit der Ruhelänge $\ell = \frac{L}{n}$, wobei $w(t) = \frac{1}{4} \left(\frac{1+t^4}{2t^2} - 1 \right)$ sei. Die gespeicherte Gravitationsenergie des i -ten Gliedes beträgt $-\frac{m}{2n} (x^{(i-1)} + x^{(i)})^\top \vec{g}$, wobei wir der Einfachheit halber $\vec{g} = (0, 0, -1)^\top$ ansetzen.

1. Formulieren Sie die Aufgabe, die stabile Lage der Kette zu finden, als ein restringiertes Optimierungsproblem in den Variablen

$$(x_1^{(0)}, x_1^{(1)}, \dots, x_1^{(n)}, x_2^{(0)}, x_2^{(1)}, \dots, x_2^{(n)}, x_3^{(0)}, x_3^{(1)}, \dots, x_3^{(n)}).$$

Berechnen Sie das Differential der Zielfunktion.

2. Für die Funktion $w(t) = \frac{1}{4} \left(\frac{1+t^4}{2t^2} - 1 \right)$ implementieren Sie die Zielfunktion und ihren Gradienten.
3. Verwenden Sie den allgemeinen Abstiegsalgorithmus mit unterschiedlichen Suchstrategien (Richtung des steilsten Abstiegs, L-BFGS-Verfahren, ...), um dieses Minimierungsproblem z. B. mit $L = 2$, $m = 1$, $x_0 = (0, 0, 0)^\top$ und $x_n = (1, 0, 1)^\top$ näherungsweise zu lösen. Vergleichen Sie die Performanz für $n = 2^k$ Glieder und mit $\text{TOL} = \sqrt{n} \cdot 10^{-6}$ für $k = 4, \dots, 10$ (beschränken Sie die maximale Anzahl an Iterationen auf 100). Probieren Sie dabei die Skalarprodukte mit den Gramschen Matrizen $B_0 = I$ und

$$B_0 = \begin{pmatrix} A & 0 & 0 \\ 0 & A & 0 \\ 0 & 0 & A \end{pmatrix} \quad \text{mit} \quad A = \frac{1}{\ell} \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 2 & -1 & 0 & \ddots & 0 & 0 \\ 0 & -1 & 2 & -1 & \ddots & 0 & 0 \\ 0 & 0 & -1 & 2 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{(n+1) \times (n+1)}$$

aus!

4. Schaffen Sie es auch, die zweite Ableitung der Zielfunktion zu implementieren? Dann können Sie auch das Newton-Verfahren versuchen.

4 Restringierte Optimierung

4.1 Optimalitätskriterien der restringierten Optimierung

4.1.1 Tangentialkegel

In einem Banachraum X und für eine auf der offenen Menge $U \subset X$ differenzierbare Funktion $F: U \subset X \rightarrow \mathbb{R}$ untersuchen wir das Optimierungsproblem

$$\begin{array}{ll} \text{Minimiere} & F(x) \\ \text{u. d. N.} & x \in M \end{array} \quad (4.1)$$

mit einem Zulässigkeitsbereich $M \subset X$ mit $M \subset U$.

Definition 4.1 Sei Z ein Vektorraum. Eine Menge $K \subset Z$ nennen wir *Kegel*, wenn für alle $s > 0$ gilt $sK \subset K$, wobei $sK := \{sz \mid z \in K\}$.

Einen Kegel K nennen wir *spitz* genau dann, wenn $(-K) \cap K \subset \{0\}$ gilt; andernfalls nennen wir K *stumpf*. Ist $0 \in K$, so nennen wir K *punktiert*.

Proposition 4.2 Sei $M \subset X$ eine nichtleere Teilmenge im Banachraum X . Einen Vektor $u \in X$ nennt man tangential an M im Punkt $x \in X$ oder auch auch (zulässigen) Tangentialvektor an x , falls es Folgen $(x_k)_{k \in \mathbb{N}}$ in M und $(t_k)_{k \in \mathbb{N}}$ in $]0, \infty[$ gibt mit

$$\lim_{k \rightarrow \infty} x_k = x, \quad \lim_{k \rightarrow \infty} t_k = 0 \quad \text{und} \quad \lim_{k \rightarrow \infty} \frac{x_k - x}{t_k} = u.$$

Die Menge aller Tangentialvektoren an M in x bezeichnet man mit $T_x M$. Sie bildet einen abgeschlossenen Kegel, den sogenannten Tangentialkegel.

BEWEIS. Sei $u \in T_x M$ ein Tangentialvektor und $s > 0$. Dann gibt es eine Folge $(x_k)_{k \in \mathbb{N}}$ von Punkten in M und eine Folge $(t_k)_{k \in \mathbb{N}}$ von positiven Zahlen mit $x_k \rightarrow x$, $t_k \rightarrow 0$ und $\frac{x_k - x}{t_k} \rightarrow u$. Dann gilt aber auch $x_k \rightarrow x$, $\frac{t_k}{s} \rightarrow 0$ und $s \frac{x_k - x}{t_k} \rightarrow su$ und damit ist $su \in T_x M$ auch ein Tangentialvektor.

Sei $(u^{(k)})_{k \in \mathbb{N}}$ eine Folge von Tangentialvektoren am Punkt x mit Grenzwert u . Dann gibt es Folgen $(x_m^{(k)})_{m \in \mathbb{N}}$ von Punkten in M und $(t_m^{(k)})_{m \in \mathbb{N}}$ von positiven Zahlen, sodass

$$x_m^{(k)} \xrightarrow{m \rightarrow \infty} x, \quad t_m^{(k)} \xrightarrow{m \rightarrow \infty} 0 \quad \text{und} \quad \frac{x_m^{(k)} - x}{t_m^{(k)}} \xrightarrow{m \rightarrow \infty} u^{(k)}.$$

Für jedes $l \in \mathbb{N}$ gibt es daher ein $m(k)$ mit

$$\left\| x_{m(k)}^{(k)} - x \right\| \leq \frac{1}{k}, \quad \left| t_{m(k)}^{(k)} \right| \leq \frac{1}{k} \quad \text{und} \quad \left\| \frac{x_{m(k)}^{(k)} - x}{t_{m(k)}^{(k)}} - u^{(k)} \right\| \leq \frac{1}{k}.$$

Folglich haben wir

$$\left\| \frac{x_{m(k)}^{(k)} - x}{t_{m(k)}^{(k)}} - u \right\| \leq \left\| \frac{x_{m(k)}^{(k)} - x}{t_{m(k)}^{(k)}} - u^{(k)} \right\| + \|u^{(k)} - u\| \leq \frac{1}{k} + \|u^{(k)} - u\| \xrightarrow{k \rightarrow \infty} 0.$$

Das zeigt $u \in T_x M$. □

4.1.2 Notwendige Optimalitätsbedingung erster Ordnung

Satz 4.3 (Notwendige Optimalitätsbedingung erster Ordnung)

Sei $x_* \in M$ ein lokales Minimum des Optimierungsproblems (4.2). Dann gilt $DF(x_*)u \geq 0$ für alle zulässigen Tangentialvektoren $u \in T_{x_*} M$.

BEWEIS. Sei $u \in T_{x_*} M$. Dann gibt es Folgen $x_n \in M$ und $t_n > 0$ mit

$$x_* = \lim_{n \rightarrow \infty} x_n \quad \text{und} \quad u = \lim_{n \rightarrow \infty} \frac{x_n - x_*}{t_n}.$$

Weil x_* ein lokales Minimum ist, gilt auf der einen Seite $F(x_n) - F(x_*) \geq 0$. Auf der anderen Seite gilt mit der Definition der Fréchet-Ableitung eine Funktion ϱ mit $\varrho(r) \xrightarrow{r \rightarrow 0} 0$, sodass

$$F(x_n) = F(x_*) + DF(x_*)(x_n - x_*) + \|x_n - x_*\| \varrho(\|x_n - x_*\|).$$

Folglich gilt

$$\begin{aligned} DF(x_*)(x_n - x_*) &= F(x_n) - F(x_*) - \|x_n - x_*\| \varrho(\|x_n - x_*\|) \\ &\geq 0 - \|x_n - x_*\| \varrho(\|x_n - x_*\|). \end{aligned}$$

Division durch $t_k > 0$ führt zu

$$DF(x_*)u = \lim_{n \rightarrow \infty} DF(x_*) \frac{x_n - x_*}{t_n} \geq - \lim_{n \rightarrow \infty} \left\| \frac{x_n - x_*}{t_n} \right\| \varrho(\|x_n - x_*\|) \xrightarrow{k \rightarrow \infty} 0. \quad \square$$

Lemma 4.4 (Hinreichende Optimalitätsbedingung erster Ordnung) Zusätzlich nehmen wir noch an, dass $M \subset X$ eine konvexe Menge und dass $F: U \rightarrow \mathbb{R}$ eine differenzierbare konvexe Funktion ist. Ist $x_* \in M$ ein Punkt mit $DF(x_*)u \geq 0$ für alle zulässigen Tangentialvektoren $u \in T_{x_*} M$, dann ist ein globales Minimum des Optimierungsproblems (4.2).

BEWEIS. Sei $y \in M$. Da M konvex ist, liegt auch $x_* + t(y - x_*)$ in M für alle $t \in]0, 1[$ und $u := y - x_*$ ist somit ein Tangentialvektor an x_* . Nach Aussage 2 von Satz 2.39 gilt nun aber

$$F(y) \geq F(x_*) + \langle DF(x_*), y - x_* \rangle \geq F(x_*). \quad \square$$

4.1.3 Hinreichende Optimalitätsbedingung zweiter Ordnung

Satz 4.5 (Hinreichende Optimalitätsbedingung zweiter Ordnung)

Sei $U \subset X$ offen, $M \subset U$ eine zulässige Menge und $F: U \rightarrow \mathbb{R}$ eine zweimal stetig differenzierbare Funktion. Sei $x_* \in M$ ein Punkt mit folgenden Eigenschaften:

1. x_* besitzt eine Umgebung V in U , so dass $V \cap M$ bezüglich x_* sternförmig ist.
2. Es gelte $DF(x_*)u \geq 0$ für alle $u \in T_{x_*}M$.
3. Es gebe ein $\lambda > 0$ so dass $DDF(x_*)(u, u) \geq \lambda \|u\|^2$ für alle $u \in T_{x_*}M$.

Dann ist x_* ein isoliertes, lokales Minimum des Optimierungsproblems (4.2).

BEWEIS. Durch Verkleinern von V können wir erreichen, dass $\|D^2F(x) - D^2F(x_*)\| < \frac{\lambda}{2}$ für alle $x \in V$ gilt. Dann haben wir für alle $u \in T_{x_*}M$ und alle $x \in V$:

$$\begin{aligned} D^2F(x)(u, u) &= D^2F(x_*)(u, u) - (D^2F(x_*)(u, u) - D^2F(x)(u, u)) \\ &\geq \lambda \|u\|^2 - \|D^2F(x_*) - D^2F(x)\| \|u\|^2 \geq \frac{\lambda}{2} \|u\|^2. \end{aligned}$$

Jedes $y \in (V \cap M) \setminus \{0\}$ können wir schreiben als $y = x_* + u$ mit einem $u \in (T_{x_*}M) \setminus \{0\}$. Nach dem Satz von Taylor gibt es ein $t \in]0, 1[$ mit

$$F(y) - F(x_*) = DF(x_*)u + \frac{1}{2}D^2F(x_* + tu)(u, u).$$

Da $V \cap M$ sternförmig bezüglich x_* ist, ist aber auch $x := x_* + tu$ darin enthalten. Daher gilt

$$F(y) - F(x_*) = DF(x_*)u + \frac{1}{2}D^2F(x_* + tu)(u, u) \geq 0 + \frac{\lambda}{4} \|y - x_*\|^2. \quad \square$$

Beispiel 4.6 Für nicht-konvexe Mengen ist dies nicht anwendbar. Um das einzusehen betrachten wir $M := \{x \in \mathbb{R}^2 \mid x_2 \geq 0, \|x\|_2^2 \geq 1\}$. Die Funktion $F: M \rightarrow \mathbb{R}$, $F(x) = x_1 + \frac{\lambda}{2}\|x\|_2^2$ erfüllt im Punkt $x_* = (1, 0)^\top$ die notwendigen Bedingungen erster Ordnung: Man hat

$$T_{x_*}M = \{u \in \mathbb{R}^2 \mid u_1 \geq 0, u_2 \geq 0\} \quad \text{und}$$

Damit gilt $DF(x_*)u = 2u_1 \geq 0$ für alle $u = (u_1, u_2) \in T_{x_*}M$. Die zweite Ableitung berechnet sich zu

$$D^2F(x_*)(u, u) = \lambda \|u\|_2^2 \quad \text{für alle } u \in T_{x_*}M.$$

Entlang des runden Randes von M haben wir aber

$$F((\cos(t), \sin(t))) = \frac{\lambda}{2} + \cos(t)$$

und dies ist für $t \in]0, \pi/2[$ stets kleiner als $F(1, 0) = F(\cos(0), \sin(0))$.

Später werden wir auch hinreichende Bedingungen zweiter Ordnung für allgemeinere zulässige Mengen kennenlernen.

4.1.4 Karush-Kuhn-Tucker-Bedingungen

Möchte man die zulässige Menge M in einer für Computer verständlichen Weise beschreiben, so ist man oft auf Gleichungen und Ungleichungen angewiesen. In diesem Abschnitt behandeln wir daher den folgenden Fall:

Seien X und Z Banachräume und sei $U \subset X$ eine offene Menge. Für (stetig) differenzierbare Abbildungen $F: U \rightarrow \mathbb{R}$ und $\Psi: U \rightarrow Z$ und einen abgeschlossenen, konvexen Kegel $K \subset Z$ betrachten wir das Optimierungsproblem

$$\begin{aligned} &\text{Minimiere} && F(x) \\ &\text{u. d. N.} && \Psi(x) \in K, \\ &&& x \in U. \end{aligned} \tag{4.2}$$

Die zulässige Menge ist also

$$M := \{x \in U \mid \Psi(x) \in K\} = \Psi^{-1}(K).$$

Für die bisherigen Optimalitätsbedingungen müssen wir für $a \in M$ den Tangentialkegel $T_a M$ kennen. Das ist in der Regel gar nicht so einfach. Deswegen wäre es erstrebenswert, wenn man $T_a M$ mittels $D\Psi(a)$ mit $T_{\Psi(a)}K$ in Verbindung bringen könnte.

Proposition 4.7 Sei Ψ differenzierbar bei $a \in M = \Psi^{-1}(K)$. Den linearisierten Tangentialkegel an M am Punkt $a \in X$ definieren wir durch

$$T_a^{\text{lin}} M := \{u \in X \mid D\Psi(a)u \in T_{\Psi(a)}K\} = D\Psi(a)^{-1}(T_{\Psi(a)}K).$$

Es gilt stets

$$T_a M \subset T_a^{\text{lin}} M.$$

Bei Gleichheit, also bei $T_a M = T_a^{\text{lin}} M$, sagt man, bei a ist die Abadie-Bedingung erfüllt.¹

BEWEIS. Sei $u \in T_a M$. Dann gibt es eine Folge $(x_k)_{k \in \mathbb{N}}$ in M und eine Folge $(t_k)_{k \in \mathbb{N}}$ positiver Zahlen, so dass

$$x_k \rightarrow a, \quad t_k \rightarrow 0 \quad \text{und} \quad \frac{x_k - a}{t_k} \rightarrow u.$$

Wegen der Differenzierbarkeit von Ψ bei a gilt

$$D\Psi(a)(x_k - a) = \Psi(x_k) - \Psi(a) + \omega(x_k) \|x_k - a\|$$

mit einer Funktion $\omega: X \rightarrow Z$ mit $\omega(x) \rightarrow 0$ für $x \rightarrow a$. Nun hat man aber $\Psi(x_k) \in K$ und daher

$$D\Psi(a)u = \lim_{k \rightarrow \infty} \left(\frac{\Psi(x_k) - \Psi(a)}{t_k} + \omega(x_k) \frac{\|x_k - a\|}{t_k} \right) = \lim_{k \rightarrow \infty} \frac{\Psi(x_k) - \Psi(a)}{t_k} \in T_{\Psi(a)}K.$$

Also ist $u \in T_{\Psi(a)}K$. □

¹Die Abadie-Bedingung wird in der Literatur in der Regel als *Abadie constraint qualification* (Abadie CQ) bezeichnet.

Beispiel 4.8 (Fehlende Abadie CQ)

Man könnte nun auf die Idee kommen, dass die Abadie-Bedingung stets erfüllt sein muss. Dem ist aber nicht so. Auf $X = \mathbb{R}^2$ betrachten wir die Abbildung $\Psi: X \rightarrow Z := \mathbb{R}$, $\Psi(x_1, x_2) = x_1 \cdot x_2$ und den Kegel $K := [0, \infty[\subset Z$. Der Zulässigkeitsbereich besteht in diesem Fall aus dem ersten und dritten Quadranten des Koordinatensystems, die Koordinatenachsen mit eingeschlossen. Am Punkt $a := (0, 0)^\top$ gilt $T_a M = M$ aber wegen $D\Psi(a) = (0, 0)$ gilt $T_a^{\text{lin}} M = \mathbb{R}^2$.

Proposition 4.9 Sei Z ein normierter, reeller Vektorraum und sei $A \subset Z$ eine beliebige Teilmenge. Wir definieren den polaren Kegel $\text{pol}(A)$ durch

$$\text{pol}(A) := \{\mu \in Z' \mid \forall z \in A: \langle \mu, z \rangle \leq 0\}.$$

Dann ist $\text{pol}(A)$ ein abgeschlossener und konvexer Kegel. Für Mengen $A_1, A_2 \subset Z$ gilt stets

$$A_1 \subset A_2 \quad \Rightarrow \quad \text{pol}(A_1) \supset \text{pol}(A_2).$$

BEWEIS. Siehe [Aufgabe 4.37](#). □

Mit dem Begriff des polaren Kegels können wir [Satz 4.3](#) wie folgt interpretieren: Ist $x_* \in M$ eine Lösung des Optimierungsproblems (4.1), so gilt

$$-DF(x_*) \in \text{pol}(T_{x_*} M).$$

Ziel dieses Abschnittes ist es, diese notwendigen Bedingungen erster Ordnung wenn auch nicht in allen, dann doch wenigstens in vielen relevanten Fällen in etwas “Maschinenlesbares” zu übersetzen. Genauer werden wir unter gewissen Regularitätsbedingungen zeigen, dass lokale Minima auch sogenannte Karush-Kuhn-Tucker-Punkte sind.

Definition 4.10 Wir nennen einen Punkt $x_* \in X$ einen *Karush-Kuhn-Tucker-Punkt* oder kurz *KKK-Punkt* genau dann, wenn es einen sogenannten *Lagrange-Multiplikator* $\mu_* \in Z'$ gibt, so dass die *Karush-Kuhn-Tucker-Bedingungen* erfüllt sind:

$$DF(x_*) + \mu_* D\Psi(x_*) = 0, \tag{4.3}$$

$$\Psi(x_*) \in K, \tag{4.4}$$

$$\mu_* \in \text{pol}(K), \tag{4.5}$$

$$\langle \mu_*, \Psi(x_*) \rangle = 0. \tag{4.6}$$

Dies wird schon unter der folgenden etwas schwächeren Bedingung gelten.

Proposition 4.11 (Abadie CQ $\Rightarrow$ Guignard CQ)

Gilt $\text{pol}(T_a M) = \text{pol}(T_a^{\text{lin}} M)$, so sagt man, bei a ist die Guignard-Bedingung erfüllt.² Die Abadie-Bedingung impliziert immer auch die Guignard-Bedingung.

²Guignard constraint qualification, auch (Guignard CQ).

Beispiel 4.12 (Guignard CQ $\Rightarrow$ Abadie CQ)

Mit [Beispiel 4.8](#) haben wir schon ein Beispiel gefunden, bei dem die Guignard-Bedingung sehr wohl, die Abadie-Bedingung nicht erfüllt ist. Für den Punkt $a = (0, 0)^T$ haben wir oben schon $T_a^{\text{lin}} M = \mathbb{R}^2 \neq M = T_a M$ herausbekommen. Deswegen gilt $\text{pol}(T_a M) = \{0\}$. Ist $\xi = (\xi_1, \xi_2) \in \text{pol}(T_a^{\text{lin}} M)$, so muss wegen $\pm e_1 \in$ und $\pm e_2 \in T_a^{\text{lin}} M$ auch

$$\xi_1 = \langle \xi, e_1 \rangle \leq 0, \quad -\xi_1 = \langle \xi, -e_1 \rangle \leq 0, \quad \xi_2 = \langle \xi, e_1 \rangle \leq 0 \quad \text{und} \quad -\xi_2 = \langle \xi, -e_2 \rangle \leq 0$$

gelten. Wir erhalten also $\text{pol}(T_a^{\text{lin}} M) = \{0\} = T_a M$ und die Guignard-Bedingung ist erfüllt.

Exemplarisch können wir hier auch nachweisen, dass a ein KKT-Punkt sein muss, wenn es ein lokales Minimum von F auf M ist: Schließlich muss $DF(a) \in -\text{pol}(T_a M) = \{0\}$ gelten und wir können als Lagrange-Multiplikator $\mu = 0$ wählen.

Abgeschlossene, konvexe Kegel kann man exakt durch ihren polaren Kegel beschreiben. Das ist Gegenstand des Bipolarensatzes, den wir im nächsten Abschnitt angehen wollen, denn dies wird uns später helfen, die notwendigen Bedingungen erster Ordnung in die Karush-Kuhn-Tucker-Bedingungen zu übersetzen.

4.1.5 Kegelgeometrie

Proposition 4.13 *Sei X ein Vektorraum und sei P eine Menge von Halbnormen auf X . Wir definieren die lokalkonvexe Topologie $\mathcal{T}(P)$ als die feinste Topologie, die alle $p \in P$ stetig macht. Insbesondere sind alle Mengen von der folgenden Form offen:*

$$B(x; p, \varepsilon) := \{y \in X \mid p(y - x) \leq \varepsilon\}, \quad x \in X, \quad p \in P, \quad \varepsilon > 0.$$

Der topologische Raum $(X, \mathcal{T}(P))$ ist ein topologischer Vektorraum, das heißt, die Additionsabbildung $+: X \times X \rightarrow X$, $(x, y) \mapsto x + y$ und die Skalarmultiplikation $\cdot: \mathbb{R} \times X \rightarrow X$, $(t, x) \mapsto tx$ sind stetig.

Mit $(X, P)'$ bezeichnen wir dann den Vektorraum der bezüglich P stetigen linearen Funktione auf X .

Einen topologischen Raum $(X, \mathcal{T})$ nennen wir lokalkonvex, falls es eine Halbnormfamilie P gibt, so dass $\mathcal{T} = \mathcal{T}(P)$ gilt.

Neben normierten Räumen sind die einzigen beiden für uns relevanten Beispiele sind die folgenden:

Beispiel 4.14 Zu einem normierten Raum X betrachten wir die Halbnormfamilie

$$P := \{x \mapsto |\langle \xi, x \rangle| \mid \xi \in X'\}.$$

Dann ist $\mathcal{T}(P)$ die schwache Topologie auf X . Es gilt (vgl. [13, Korollar VIII.3.4]):

$$(X, P)' = X'.$$

Beispiel 4.15 Auf dem Dualraum X' eines normierten Raumes X betrachten wir die Halbnormfamilie

$$P := \{ \xi \mapsto |\langle \xi, x \rangle| \mid x \in X \}.$$

Dann ist $\mathcal{T}(P)$ die *schwach-*-Topologie* auf X' . Es gilt (vgl. [13, Korollar VIII.3.4]):

$$(X, P)' = \psi_X(X) = \{ \xi \mapsto \langle \xi, x \rangle \mid x \in X \}.$$

Der Hahn-Banachsche Trennungssatz [Satz 2.34](#) lässt sich auf lokalkonvexe Mengen wie folgt erweitern:

Satz 4.16 (Hahn-Banachscher Trennungssatz)

Sei (X, P) ein reeller, lokalkonvexer Raum, $C \subset X$ eine abgeschlossene, konvexe Menge und $x_0 \in X$ ein Punkt. Dann liegt der Punkt x_0 entweder in C oder er wird durch eine Hyperebene von ihm getrennt, d. h. es gibt ein $\xi \in X'$ und ein $s \in \mathbb{R}$ mit $\langle \xi, x_0 \rangle > s$ und $\langle \xi, x \rangle \leq s$ für alle $x \in C$.

BEWEIS. Im Wesentlichen kann man genauso vorgehen wie beim Trennungssatz für normierte Räume. Die Details sind in [13, Kapitel VIII], insbesondere in Theorem VIII.2.11 und Theorem VIII.2.12 zu finden. $\square$

Diese Verallgemeinerung benötigen wir, um das folgende Lemma zu zeigen, dass nicht ganz von [Satz 2.34](#) abgedeckt wird:

Lemma 4.17 (Hahn-Banachscher Trennungssatz im Dualraum) Sei X ein reeller, normierter Raum, $C \subset X'$ eine schwach-*-abgeschlossene, konvexe Menge und $\xi_0 \in X'$ ein Punkt. Dann liegt der Punkt ξ_0 entweder in C oder er wird durch ein $x \in X$ von ihm getrennt, d. h. es gibt ein $x \in X$ und ein $s \in \mathbb{R}$ mit $\langle \xi_0, x \rangle > s$ und $\langle \xi, x \rangle \leq s$ für alle $\xi \in C$.

BEWEIS. Wir verwenden den Raum (Y, P) aus [Beispiel 4.15](#). Da C schwach-*-abgeschlossen ist, ist es auch P -abgeschlossen. Nach [Satz 4.16](#) gibt es ein $\eta \in (Y, P)' = \psi_X(X)$ und ein $s \in \mathbb{R}$, so dass $\langle \eta, \xi_0 \rangle > s$ und $\langle \eta, \xi \rangle \leq s$ für alle $\xi \in C$. Also gibt es ein $x \in X$ mit $\eta = \psi_X(x)$ und wir haben

$$\langle \xi_0, x \rangle = \langle \eta, \xi_0 \rangle > s \quad \text{und} \quad \langle \xi_0, x \rangle = \langle \eta, \xi \rangle \leq s$$

für alle $\xi \in C$. $\square$

Den polaren Kegel für Mengen in einem lokalkonvexen Raum definieren wir ganz analog:

Proposition 4.18 Sei Z ein reeller, lokalkonvexer Vektorraum und sei $A \subset Z$ eine beliebige Teilmenge. Wir definieren den polaren Kegel $\text{pol}(A)$ durch

$$\text{pol}(A) := \{ \mu \in Z' \mid \forall z \in A: \langle \mu, z \rangle \leq 0 \} = \bigcap_{z \in A} L_0^-(\psi_Z(z)),$$

wobei $\psi_Z: Z \rightarrow Z'', \langle \psi_Z(z), \mu \rangle := \langle \mu, z \rangle$ die kanonische Einbettung sei. Dann ist $\text{pol}(A)$ also ein abgeschlossener und sogar schwach-*-abgeschlossener und konvexer Kegel. Für Mengen $A_1, A_2 \subset Z$ gilt stets

$$A_1 \subset A_2 \quad \Rightarrow \quad \text{pol}(A_1) \supset \text{pol}(A_2).$$

Wesentliches Werkzeug ist der folgende Spezialfall des Hahn-Banachschen Trennungssatzes.

Lemma 4.19 (Konischer Trennungssatz)

Sei Z ein reeller, lokalkonvexer Raum, sei $K \subset Z$ ein abgeschlossener, konvexer Kegel und $z_0 \in Z$. Dann liegt der Punkt z_0 entweder im Kegel K oder er wird durch eine Hyperebene von ihm getrennt, d. h. es gibt ein $\mu \in \text{pol}(K)$ mit $\langle \mu, z_0 \rangle > 0$.

BEWEIS. Sei $z_0 \in Z \setminus K$ ein Punkt außerhalb des Kegels K . Da K abgeschlossen und konvex ist, gibt es nach dem Satz Hahn-Banach [Satz 4.16](#) ein $\mu \in Z'$ und ein $s \in \mathbb{R}$ mit

$$\langle \mu, z_0 \rangle > s \quad \text{und} \quad \langle \mu, z \rangle \leq s \quad \text{für alle } z \in K.$$

Wir müssen nun lediglich noch zeigen, dass wir in diesen Ungleichungen s durch 0 ersetzen dürfen, denn dann wissen wir, dass $\mu \in \text{pol}(K)$ ist. Zunächst beobachten wir, dass $0 \in K$ ist und folgern $s \geq \langle \mu, 0 \rangle = 0$. Ist $s = 0$, so sind wir schon fertig. Im Falle $s > 0$ argumentieren wir so:

Angenommen, es gäbe ein $z \in K$ mit $t := \langle \mu, z \rangle > 0$. Weil K ein Kegel ist, liegt dann auch der Punkt $\tilde{z} := 2s/t z$ in K . Dann hat man aber

$$\langle \mu, \tilde{z} \rangle = 2s/t \langle \mu, z \rangle = 2s > s$$

und dies wäre ein Widerspruch zur Wahl von μ . Dies beweist die behauptete Aussage. $\square$

Definition 4.20 Sei Z ein reeller, lokalkonvexer Vektorraum und sei $A \subset Z$ eine beliebige Teilmenge. Wir definieren die *konische Hülle* $\text{cone}(A)$ durch

$$\text{cone}(A) := \left\{ \sum_{i=1}^m t_i a_i \mid m \in \mathbb{N}, a_i \in A, t_i \geq 0 \right\} = \bigcap_{\substack{K \text{ konv. Kegel} \\ A \subset K}} K.$$

Dann ist $\text{cone}(A)$ ein konvexer Kegel, der von A erzeugte konvexe Kegel.

Die *abgeschlossene konische Hülle* $\overline{\text{cone}}(A)$ von A definieren wir als den kleinsten abgeschlossenen, konvexen Kegel, der A enthält:

$$\overline{\text{cone}}(A) := \bigcap_{\substack{K \text{ abg. konv. Kegel} \\ A \subset K}} K.$$

Für eine Menge $B \subset X'$ definieren wir noch die *schwach-*-abgeschlossene, konische Hülle*

$$\widetilde{\text{cone}}(B) := \bigcap_{\substack{K \text{ schwach-*-abg. konv. Kegel} \\ B \subset K}} K.$$

Bemerkung 4.21 Sei X ein reeller, normierter Raum. Dann haben wir die verwirrende Situation, dass wir drei verschiedene Topologien auf X' haben:

1. die schwach-*-Topologie,
2. die schwache Topologie und
3. die starke Topologie, also die durch die Dualnorm induzierte Topologie.

Die Hahn-Banachschen Trennungssätze sagen uns dass für jede konvexe Menge $C \subset X'$ gilt:

$$C \text{ schwach-abgeschlossen} \Leftrightarrow C \text{ stark-abgeschlossen.}$$

Außerdem gilt für konvexe Mengen $C \subset X'$:

$$C \text{ schwach-*-abgeschlossen} \Rightarrow C \text{ stark-abgeschlossen.} \quad (4.7)$$

Inbesondere gibt es mehr stark-abgeschlossene, konvexe Kegel $K \supset B$ als schwach-*-abgeschlossene und folglich gilt

$$\overline{\text{cone}}(B) \subset \widetilde{\text{cone}}(B). \quad (4.8)$$

Man könnte jetzt versucht sein, Äquivalenz in (4.7) und somit Gleichheit in (4.8) anzunehmen. **Dem ist aber nicht so!** Vielmehr gilt:

$$C \text{ konvex und schwach-*-abg.} \Leftrightarrow C \text{ konvex und stark-abg.}$$

Fixme Note:
Überprüfe
das!

dann und nur dann, wenn X reflexiv ist!

Proposition 4.22 Sei Z ein reeller, lokalkonvexer Vektorraum und sei $B \subset Z'$ eine beliebige Teilmenge. Wir definieren den präpolaren Kegel $\text{prepol}(B)$ durch

$$\text{prepol}(B) := \{z \in Z \mid \forall \mu \in B: \langle \mu, z \rangle \leq 0\} = \bigcap_{\mu \in B} L_0^-(\mu).$$

Dann ist $\text{prepol}(B)$ ein abgeschlossener, konvexer Kegel. Für Mengen $B_1, B_2 \subset Z'$ mit gilt stets

$$B_1 \subset B_2 \Rightarrow \text{prepol}(B_1) \supset \text{prepol}(B_2).$$

BEWEIS. Siehe [Aufgabe 4.38](#). □

Satz 4.23 (Bipolarensatz)

Sei $A \subset Z$ eine Menge im reellen, lokalkonvexen Raum Z und sei $B \subset Z'$. Dann gelten

$$\text{prepol}(\text{pol}(A)) = \overline{\text{cone}}(A) \quad \text{und} \quad \text{pol}(\text{prepol}(B)) = \widetilde{\text{cone}}(B) \supset \overline{\text{cone}}(B)$$

Ist Z reflexiv, so gilt sogar

$$\text{pol}(\text{prepol}(B)) = \overline{\text{cone}}(B).$$

BEWEIS. Sei $a \in A$. Dann gilt für alle $\mu \in \text{pol}(A)$ schon $\langle \mu, a \rangle \leq 0$. Also ist $A \subset \text{prepol}(\text{pol}(A))$. Weil $\text{prepol}(\text{pol}(A))$ abgeschlossen und konvex ist, gilt auch $\overline{\text{cone}}(A) \subset \text{prepol}(\text{pol}(A))$.

Angenommen, es gibt $z_0 \in \text{prepol}(\text{pol}(A))$ mit $z_0 \notin \overline{\text{cone}}(A)$. Nach dem konischen Trennungssatz [Lemma 4.19](#) gibt es ein $\mu \in \text{pol}(\overline{\text{cone}}(A)) \subset \text{pol}(A)$ mit $\langle \mu, z_0 \rangle > 0$. Das ist aber ein Widerspruch zu $z_0 \in \text{prepol}(\text{pol}(A))$. Dies zeigt die erste Identität.

Für die zweite Identität zeigt man zuerst analog zu oben, dass $B \subset \widetilde{\text{cone}}(B) \subset \text{pol}(\text{prepol}(B))$ gilt.

Angenommen, es gäbe $\mu_0 \in \text{pol}(\text{prepol}(B))$ mit $\mu_0 \notin \widetilde{\text{cone}}(B)$. Nach dem konischen Trennungssatz [Lemma 4.19](#) im lokalkonvexen Raum Z' mit schwach-*-Topologie gibt es dann ein $z \in \text{prepol}(\widetilde{\text{cone}}(B)) \in Z$ mit $\langle \mu_0, z \rangle > 0$. Wegen $\mu_0 \in \text{pol}(\text{prepol}(B))$ führt dies aber zum Widerspruch $0 \geq \langle \mu_0, z \rangle = \langle M, \mu_0 \rangle > 0$.

Folgerung 4.24 *Ist Z reflexiv, so gilt*

$$\text{pol}(\text{pol}(A)) = \psi_Z(\text{prepol}(\text{pol}(A))) = \psi_Z(\overline{\text{cone}}(A)).$$

Unterdrückt man die Nennung der kanonischen Abbildung, so erhält man

$$\text{pol}(\text{pol}(A)) = \overline{\text{cone}}(A).$$

Dies erklärt den Namen “Bipolarensatz”.

Lemma 4.25 *Seien X und Z reelle, normierte Vektorräume. Sei $K \subset Z$ ein abgeschlossener, konvexer Kegel und $A \in L(X; Z)$ eine stetige, lineare Abbildung. Dann gilt*

$$A^{-1}(K) = \text{prepol}(A'(\text{pol}(K))) \quad \text{and} \quad \text{pol}(A^{-1}(K)) = \widetilde{\text{clos}}(A'(\text{pol}(K))).$$

BEWEIS. Die Menge $A^{-1}(K)$ ist ein abgeschlossener, konvexer Kegel. Für jedes $u \in A^{-1}(K)$ gilt $Au \in K$ und daher $\langle A'\mu, u \rangle = \langle \mu, Au \rangle \leq 0$ für alle $\mu \in \text{pol}(K)$. Folglich gilt $A'(\text{pol}(K)) \subset \text{pol}(A^{-1}(K))$. Der Bipolarensatz [Satz 4.23](#) impliziert nun

$$A^{-1}(K) = \overline{\text{cone}}(A^{-1}(K)) = \text{prepol}(\text{pol}(A^{-1}(K))) \subset \text{prepol}(A'(\text{pol}(K))).$$

Angenommen, es gibt ein $u_0 \in \text{prepol}(A'(\text{pol}(K)))$ mit $u_0 \notin A^{-1}(K)$. Dann ist $Au_0 \notin K$ und kann daher mit einem $\mu \in \text{pol}(K)$ von K getrennt werden, d. h., es gilt $\langle \mu, Au_0 \rangle > 0$. Dann gelangen wir aber zu

$$\langle A'\mu, u_0 \rangle = \langle \mu, Au_0 \rangle > 0$$

und das ist ein Widerspruch zu $u_0 \in \text{prepol}(A'(\text{pol}(K)))$. Folglich haben wir $A^{-1}(K) = \text{prepol}(A'(\text{pol}(K)))$.

Wendet man nun auf beiden Seiten $\text{pol}()$ an, so liefert der Bipolarensatz [Satz 4.23](#):

$$\text{pol}(A^{-1}(K)) = \text{pol}(\text{prepol}(A'(\text{pol}(K)))) = \widetilde{\text{cone}}(A'(\text{pol}(K))) = \widetilde{\text{clos}}(A'(\text{pol}(K))). \quad \square$$

Lemma 4.26 (Farkas-Alternative)

Seien X und Z reelle, normierte Räume. Sei $K \subset Z$ ein abgeschlossener, konvexer Kegel und $A \in L(X; Z)$ eine stetige, lineare Abbildung, sodass $A'(\text{pol}(K))$ schwach-*-abgeschlossen ist. Für jedes $\xi \in X'$ ist dann genau eine der beiden folgenden Aussagen wahr:

1. Es gibt ein $u \in X$ mit $Au \in K$ und $\langle \xi, u \rangle < 0$.
2. Es gibt ein $\mu \in \text{pol}(K)$ mit $\xi + A'\mu = 0$.

BEWEIS. Aussage 1 ist äquivalent zu $\xi \notin -\text{pol}(A^{-1}(K))$ und Aussage 2 ist äquivalent zu $\xi \in -A'(\text{pol}(K)) = -\overline{\text{clos}}(A'(\text{pol}(K)))$. Damit folgt das Lemma aus [Lemma 4.25](#). $\square$

Beispiel 4.27 Die Bedingung, dass $A'(\text{pol}(K))$ schwach-*-abgeschlossen sein soll ist ein lästiger Schönheitsfehler. Im Allgemeinen kann man sie aber nicht weglassen, denn das Bild eines schwach-*-abgeschlossenen, konvexen Kegels braucht selbst unter einer linearen Abbildung nicht abgeschlossen zu sein. Das soll dieses Beispiel aus [2] illustrieren: Da $\mathbb{R}^4$ ein Hilbertraum ist, können wir $\mathbb{R}^4$ mit seinem Dualraum identifizieren; da $\mathbb{R}^4$ reflexiv ist, sind in diesem Sinne die Begriffe “konvex und abgeschlossen” und “konvex und schwach-*-abgeschlossen” äquivalent. Wir betrachten nun den abgeschlossenen und konvexen Kegel

$$K := \left\{ (w, x, y, z) \in \mathbb{R}^4 \mid 0 \leq w, 0 \leq x \text{ und } y^2 + (z - x)^2 \leq x^2 \right\}.$$

Für $\lambda \geq 0$ betrachten wir die linearen Abbildungen

$$A_\lambda: \mathbb{R}^4 \rightarrow \mathbb{R}^3 \quad A_\lambda(w, x, y, z) := (x - \lambda w, y, z).$$

Für $\lambda = 0$ ist $A_\lambda(K)$ abgeschlossen; aber für $\lambda > 0$ nicht. Letzteres sehen wir so ein: Für $\varepsilon > 0$, $\lambda > 0$ definieren wir

$$\gamma(\varepsilon, \lambda) := \left(\frac{1+\varepsilon^2}{2\varepsilon}, \frac{(1-\varepsilon)^2}{2\varepsilon\lambda}, 1, \varepsilon \right) \in K.$$

Für alle $\varepsilon > 0$ gilt also $(1, 1, \varepsilon) = A_\lambda \gamma(\varepsilon, \lambda) \in A_\lambda(K)$. Der Punkt $(1, 1, 0) = \lim_{\varepsilon \rightarrow 0} (1, 1, \varepsilon)$ kann aber nicht in $A_\lambda(K)$ liegen, denn dazu müsste $1 + x^2 \leq x^2$ gelten.

Lohn unserer Mühen ist dieser wichtige Satz:

Satz 4.28 (Guinard CQ + $\varepsilon \Rightarrow$ KKT)

Seien X und Z reelle, normierte Räume. Sei $K \subset Z$ ein abgeschlossener, konvexer Kegel und sei $x_* \in U$ ein lokales Minimum des Optimierungsproblems (4.2), bei dem die Guinard-Bedingung erfüllt ist. Sei überdies $D\Psi(x_*)'(\text{pol}(T_{z_*}K))$ schwach-*-abgeschlossen, wobei $z_* := \Psi(x_*)$.

Dann ist x_* ein Karush-Kuhn-Tucker-Punkt (siehe [Definition 4.10](#)).

BEWEIS. Die Inklusion (4.4) ist natürlich erfüllt, da ja x_* ein zulässiger Punkt sein soll. Sei $\xi := DF(x_*)$ und $A := D\Psi(x_*)$. Wegen der notwendigen Bedingung erster Ordnung aus [Satz 4.3](#) und der Guinard-Bedingung erhalten wir

$$-\xi \in \text{pol}(T_{x_*}M) = \text{pol}(T_{x_*}^{\text{lin}}M) = \text{pol}(A^{-1}(T_{z_*}K)).$$

Es gilt also

$$\langle \xi, u \rangle \geq 0 \quad \text{für alle } u \in X \text{ mit } Au \in T_{z_*} K.$$

Das heißt, der erste Fall der Farkas-Alternative kommt nicht in Frage. Also muss der zweite Fall gelten. Es gibt also ein $\mu_* \in \text{pol}(T_{z_*} K)$ mit $\xi + \mu_* A = \xi + A'\mu_* = 0$. Das sind gerade die KKT-Bedingungen (4.3) und (4.5). Wie man als leichte Übungsaufgabe nachrechnet, gilt

$$\text{pol}(T_z K) = \{ \mu \in \text{pol}(K) \mid \langle \mu, z \rangle = 0 \}$$

für alle $z \in K$. Damit gilt dann auch $\langle \mu_*, \Psi(x_*) \rangle = 0$; das ist aber die letzte KKT-Bedingung (4.6). $\square$

4.1.6 Polyhedrische Kegel

In den Anwendungen sind die zu verwendenden Kegel sehr oft polyhedrisch. Für solche Kegel vereinfacht sich die Situation erheblich.

Definition 4.29 Sei $K \subset Z$ ein Kegel im reellen, lokalkonvexen Raum Z .

Wir nennen K *endlich erzeugt*, wenn es endlich viele Vektoren $z_1, \dots, z_n \in Z$ gibt mit $K = \text{cone}(z_1, \dots, z_n)$.

Wir nennen K *polyhedrisch*, wenn es endlich viele Funktionale $\zeta_1, \dots, \zeta_m \in Z'$ gibt mit

$$K = \bigcap_{j=1}^m L_0^-(\zeta_j) = \bigcap_{j=1}^m \{ z \in Z \mid \langle \zeta_j, z \rangle \leq 0 \}.$$

Beispiel 4.30 Sei $e_1, \dots, e_m$ die Standardbasis im $Z := \mathbb{R}^m$ sei und $\zeta_1, \dots, \zeta_m$ die dazugehörige duale Basis. Der Koordinatenorthant

$$K := \{ z = (z_1, \dots, z_m) \in \mathbb{R}^m \mid \forall j \in \{1, \dots, m\} : z_j \leq 0 \} = \bigcap_{j=1}^m L_0^-(\zeta_j)$$

ist offensichtlich polyhedrisch. Er ist aber auch endlich erzeugt durch $-e_1, \dots, -e_m$.

Bei sehr vielen in der Praxis vorkommenden Optimierungsproblemen kommt man mit Kegeln der Form $\{0\} \times \text{polyhedrischer Kegel}$ aus. Genauer sehen diese Probleme wie folgt aus: Seien X , Y und Z normierte Räume und sei $U \subset X$ offen. Seien $F: U \rightarrow \mathbb{R}$, $\Phi: U \rightarrow Y$ und $\Psi: U \rightarrow Z$ differenzierbare Abbildungen. Zu stetigen Linearformen $\zeta_1, \dots, \zeta_m \in Z'$ betrachten wir den polyhedrischen Kegel $K = \bigcap_{j=1}^m L_0^-(\zeta_j) \subset Z$ und das Optimierungsproblem

$$\begin{aligned} &\text{Minimiere} && F(x) \\ &\text{u. d. N.} && \Phi(x) = 0, \\ & && \Psi(x) \in K, \\ & && x \in U. \end{aligned} \tag{4.9}$$

Durch Setzen von $\Psi_j(x) := \langle \zeta_j, \Psi(x) \rangle$, $j \in \{1, \dots, m\}$ können das gleichbedeutend mit

$$\begin{aligned} &\text{Minimiere } F(x) \\ &\text{u. d. N. } \Phi(x) = 0, \\ &\Psi_j(x) \leq 0, \quad j \in \{1, \dots, m\} \\ &x \in U. \end{aligned} \tag{4.10}$$

Das K polyhedrisch ist bedeutet also, dass nur endliche viele Ungleichheitsbedingungen im Optimierungsproblem auftauchen. Wegen [Lemma 4.33](#) sehen die Karush-Kuhn-Tucker-Bedingungen (4.20) bis (4.24) in diesem Fall wie folgt aus: Es gibt ein $\lambda_* \in Y$ und $\mu_{*j} \in \mathbb{R}$, $j \in \{1, \dots, m\}$, so dass

$$DF(x_*) + \lambda_* D\Phi(x_*) + \sum_{j=1}^m \mu_{*j} D\Psi_{*j}(x_*) = 0, \tag{4.11}$$

$$\Phi(x_*) = 0, \tag{4.12}$$

$$\Psi_j(x_*) \leq 0 \quad \text{für } j \in \{1, \dots, m\}, \tag{4.13}$$

$$\mu_{*j} \geq 0 \quad \text{für } j \in \{1, \dots, m\}, \tag{4.14}$$

$$\sum_{j=1}^m \mu_{*j} \Psi_j(x_*) = 0. \tag{4.15}$$

Wir wollen sehen, wann die KKT-Bedingungen notwendig für einen lokalen Minimierer sind. Dafür müssen wir polyhedrische Kegel etwas genauer studieren. Dies wird uns zu einer einfacheren Version der Farkas-Alternative bringen (siehe [Lemma 4.35](#)), welche uns danach sagt, dass die Guinard-Bedingung allein im polyhedrischen Fall schon ausreicht (siehe [Satz 4.36](#)).

Da $L_0^-(\zeta_j)$ abgeschlossene Halbräume sind, ist an der Definition auch einfach abzulesen, dass polyhedrische Kegel immer konvex und abgeschlossen sind. Für endlich erzeugte Kegel gilt das auch; es ist aber etwas schwieriger zu beweisen:

Lemma 4.31 *Sei A eine endliche Menge im lokalkonvexen Raum Z . Dann ist der endlich erzeugte Kegel $\text{cone}(A)$ abgeschlossen. Insbesondere gilt dann $\text{cone}(A) = \overline{\text{cone}(A)}$.*

BEWEIS. Schritt 1.: Wir zeigen, dass es für jedes $x \in \text{cone}(A) \setminus \{0\}$ eine linear unabhängige Teilmenge $B \subset A$ gibt, so dass $x \in \text{cone}(B)$. Dazu betrachten wir die Menge

$$\{ \#(B) \mid B \subset A \exists t: B \rightarrow]0, \infty[: x = \sum_{b \in B} t(b) b \}.$$

Diese hat gewiss ein Minimum $m = m(x)$. Wir können x schreiben als

$$x = \sum_{i=1}^m t_i b_i \quad \text{mit } t_1, \dots, t_m > 0 \text{ und } b_1, \dots, b_m \in B.$$

Angenommen, $b_1, \dots, b_m$ sind linear abhängig. Dann gibt es $(s_1, \dots, s_m)^T \in \mathbb{R}^m \setminus \{0\}$ mit $\sum_{i=1}^m s_i b_i = 0$. Durch Umnummerierung können wir erreichen, dass $0 < \frac{t_m}{s_m} \leq \frac{t_i}{s_i}$ gilt für alle i mit $\frac{t_i}{s_i} > 0$. Dann gilt

$$x = \sum_{i=1}^m t_i b_i - \frac{t_m}{s_m} \sum_{i=1}^m s_i b_i = \sum_{i=1}^{m-1} \left(t_i - \frac{t_m}{s_m} s_i \right) b_i$$

und $t_i - \frac{s_m}{t_m} s_i > 0$ für $i \in \{1, \dots, m-1\}$. Das ist aber ein *Widerspruch* zur Minimalität von m .

Schritt 2.: Nach Schritt 1 gilt

$$\text{cone}(A) = \bigcup_{B \subset A \text{ ist linear unabhängig}} \text{cone}(B).$$

Ist aber $B \subset A$ linear unabhängig, so ist $\text{cone}(B)$ als homöomorphes Bild eines Koordinatenthantens abgeschlossen. Da A endlich ist, ist $\text{cone}(A)$ also eine endliche Vereinigung abgeschlossener Mengen und als solche wieder abgeschlossen. $\square$

Lemma 4.32 Sei Z ein normierter Raum und $K = \text{cone}(z_1, \dots, z_m) \subset Z$ sei endlich erzeugt. Dann gilt

$$\text{pol}(K) = \bigcap_{j=1}^m L_0^-(\psi_Z(z_j))$$

und $\text{pol}(K)$ ist somit polyhedrisch. Hierbei ist $\psi_Z: Z \rightarrow Z''$ die kanonische Abbildung in den Bidualraum.

BEWEIS. Wir kürzen ab $L := \bigcap_{j=1}^m L_0^-(\psi_Z(z_j))$. Ist $\mu \in \text{pol}(K)$, so gilt zwangsläufig

$$\langle \psi_Z(z_j), \mu \rangle = \langle \mu, z_j \rangle \leq 0 \quad \text{für alle } j \in \{1, \dots, m\}.$$

Folglich gilt $\text{pol}(K) \subset L$. Ist aber $\mu \in L$ und $z \in K$ ein beliebiger Punkt, so kann man schreiben $z = \sum_{j=1}^m t_j z_j \in K$ mit $t_j \geq 0$. Dann hat man

$$\langle \mu, z \rangle = \sum_{j=1}^m t_j \langle \mu, z_j \rangle = \sum_{j=1}^m t_j \langle \psi_Z(z_j), \mu \rangle \leq 0$$

und dies zeigt schon $L \subset \text{pol}(K)$. $\square$

Lemma 4.33 Sei Z ein normierter Raum und $K = \bigcap_{j=1}^m L_0^-(\zeta_j) \subset Z$ ein polyhedrischer Kegel. Dann gilt

$$\text{pol}(K) = \text{cone}(\zeta_1, \dots, \zeta_m).$$

Der Kegel $\text{pol}(K)$ ist also endlich erzeugt. Insbesondere bedeutet das, dass sich jedes $\mu \in \text{pol}(K)$ darstellen lässt als

$$\mu = \mu_1 \zeta_1 + \dots + \mu_m \zeta_m \quad \text{mit} \quad \mu_1, \dots, \mu_m \geq 0.$$

BEWEIS. Wir kürzen ab $L := \text{cone}(\zeta_1, \dots, \zeta_m)$. Es gilt $\zeta_j \in \text{pol}(K)$ und weil $\text{pol}(K)$ konvex ist, ist auch $L \subset \text{pol}(K)$. Mit [Lemma 4.31](#) und dem Bipolarensatz (siehe [Satz 4.23](#)) können wir wie folgt rechnen:

$$K = \overline{\text{cone}}(K) = \text{prepol}(\text{pol}(K)) \subset \text{prepol}(L) = \bigcap_{\zeta \in L} L_0^-(\zeta) \subset \bigcap_{j=1}^m L_0^-(\zeta_j) = K.$$

Wenden wir noch einmal die Polare an, so erhalten wir aus dem zweiten Teil des Bipolarensatzes [Satz 4.23](#), dass

$$\text{pol}(K) = \text{pol}(\text{prepol}(L)) = \overline{\text{cone}}(L).$$

Da aber der Kegel L endlich erzeugt ist, ist er nach [Lemma 4.31](#) schon schwach-*-abgeschlossen und es folgt $\text{pol}(K) = \overline{\text{cone}}(L) = L$. $\square$

Folgerung 4.34 *Seien X und Z normierte Räume. Ist $K \subset Z$ polyhedrisch und ist $A \in L(X; Z)$, so gilt $\text{pol}(A^{-1}(K)) = A'(\text{pol}(K))$.*

BEWEIS. Sei $K = \bigcap_{j=1, \dots, m} L_0^-(\zeta_j)$. Aus [Lemma 4.33](#) folgt $\text{pol}(K) = \text{cone}(\zeta_1, \dots, \zeta_m)$. Wir haben daher

$$A^{-1}(K) = \bigcap_{j=1, \dots, m} L_0^-(A' \zeta_j) \quad \text{und} \quad A'(\text{pol}(K)) = \text{cone}(A' \zeta_1, \dots, A' \zeta_m).$$

Andererseits schließen wir abermals mit [Lemma 4.33](#), dass

$$\text{pol}(A^{-1}(K)) = \text{cone}(A' \zeta_1, \dots, A' \zeta_m)$$

gelten muss. Wir haben also $\text{pol}(A^{-1}(K)) = A'(\text{pol}(K))$. $\square$

Die Bedingungen in der Farkas-Alternative [Lemma 4.26](#), dass der Kegel $A'(\text{pol}(K))$ schwach-*-abgeschlossen sein muss, können wir für polyhedrische Kegel also fallen lassen:

Lemma 4.35 (Farkas-Alternative für polyhedrische Kegel)

Seien X und Z normierte Räume. Sei $K \subset Z$ ein polyhedrischer Kegel und $A \in L(X; Z)$ eine stetige, lineare Abbildung. Für jedes $\xi \in X'$ ist dann genau eine der beiden folgenden Aussagen wahr:

1. *Es gibt ein $u \in X$ mit $Au \in K$ und $\langle \xi, u \rangle < 0$.*
2. *Es gibt ein $\mu \in \text{pol}(K)$ mit $\xi + A' \mu = 0$.*

Als Lohn unserer Mühen erhalten wir deutlich leichter überprüfbare Bedingungen für die Notwendigkeit der KKT-Bedingungen:

Satz 4.36 (Guignard CQ + polyhedrisch $\Rightarrow$ KKT)

Sei $x_* \in U$ ein lokales Minimum des Optimierungsproblems (4.9), bei dem die Guignard-Bedingung erfüllt ist.

Dann ist x_* notwendigerweise ein Karush-Kuhn-Tucker-Punkt im Sinne von (4.11)–(4.15).

4.1.7 Übungsaufgaben

Aufgabe 4.37

Sei Z ein normierter Raum und $A, A_1, A_2 \subset Z$ beliebige Mengen. Dann ist $\text{pol}(A)$ ein schwach*-abgeschlossener, konvexer Kegel. Gilt $A_1 \subset A_2$, so folgt $\text{pol}(A_1) \supset \text{pol}(A_2)$.

Aufgabe 4.38

Sei Z ein normierter Raum und $B, B_1, B_2 \subset Z'$ beliebige Mengen. Dann ist $\text{prepol}(B)$ ein abgeschlossener, konvexer Kegel. Gilt $B_1 \subset B_2$, so folgt $\text{prepol}(B_1) \supset \text{prepol}(B_2)$.

Aufgabe 4.39

Sei X ein normierter Raum und $C \subset X$ eine konvexe Menge. Dann ist auch der Abschluss $\overline{C}$ konvex.

Aufgabe 4.40

Sei Z ein normierter Raum und $K \subset Z$ ein abgeschlossener, konvexer Kegel. Zeigen sie, dass $\text{pol}(T_z K) = \{\mu \in \text{pol}(K) \mid \langle \mu, z \rangle = 0\}$ für beliebiges $z \in K$ gilt.

Aufgabe 4.41 (affin + polyhedrisch $\Rightarrow$ Abadie)

Seien X und Z normierte Räume, $K \subset Z$ ein polyedrischer Kegel und $\Psi: X \rightarrow Z$ eine stetige, affine Abbildung. Wir betrachten den Zulässigkeitsbereich $M := \Psi^{-1}(K)$. Dann ist für jeden zulässigen Punkt $a \in M$ die Abadie-Bedingung erfüllt.

Aufgabe 4.42

Ein Landwirt verfügt über 50 ha Ackerland, das er zum Anbau von Braugerste oder Zuckerrüben nutzen kann. Für die Frühjahrsarbeit sind bei Braugerste 10 Stunden pro ha und bei Zuckerrüben 40 Stunden pro ha notwendig. Es stehen für diese Zeit 800 Arbeitsstunden zur Verfügung. Für die Ernte sind bei Braugerste 8 Stunden pro ha und bei Zuckerrüben 20 Stunden pro ha erforderlich. Im Erntezeitraum stehen aber nur 460 Arbeitsstunden zur Verfügung. Wegen des notwendigen Fruchtwechsels dürfen nicht mehr als 18 ha Zuckerrüben angebaut werden. Der Gewinn je ha Braugerste beträgt 200 € und je ha Zuckerrüben 600 €. Wie hoch ist der in einem Jahr maximal erzielbare Gewinn? Wie viel ist von den zwei Sorten anzubauen, damit dieser erreicht wird? Formulieren Sie dieses Problem als restringiertes Optimierungsproblem und lösen Sie es grafisch.

Aufgabe 4.43 (Verallgemeinerte Ungleichungen)

Sei $K \subset Z$ ein abgeschlossener, konvexer und spitzer Kegel im reellen, lokalkonvexen Raum Z . Für $z_1, z_2 \in Z$ definieren wir die Relation

$$z_1 \succcurlyeq_K z_2 \quad \Leftrightarrow \quad z_1 \in z_2 + K.$$

Insbesondere hat man $z \succcurlyeq_K 0$ genau dann, wenn $z \in K$. Damit kann man also das Optimierungsproblem (4.2) auch so umformulieren:

$$\begin{array}{ll} \text{Minimiere} & F(x) \\ \text{u. d. N.} & \Psi(x) \succcurlyeq_K 0, \\ & x \in U. \end{array}$$

Die KKT-Bedingungen aus [Definition 4.10](#) kann man dann wie folgt umformulieren:

$$\begin{aligned} DF(x_*) + \mu_* D\Psi(x_*) &= 0, \\ \Psi(x_*) &\succcurlyeq_K 0 \\ \mu_* &\succcurlyeq_{\text{pol}(K)} 0, \\ \langle \mu_*, \Psi(x_*) \rangle &= 0. \end{aligned}$$

Zeigen Sie für die Relation $\succcurlyeq_K$: Durch $\succcurlyeq_K$ wird eine *Halbordnung* auf Z definiert, das heißt, es gelten für alle z, z_1, z_2 und $z_3 \in Z$:

1. (Reflexivität) $z \succcurlyeq_K z$.
2. (Antisymmetrie) Gilt $z_1 \succcurlyeq_K z_2$ und $z_2 \succcurlyeq_K z_1$, dann gilt schon $z_1 = z_2$.
3. (Transitivität) Gilt $z_1 \succcurlyeq_K z_2$ und $z_2 \succcurlyeq_K z_3$, dann gilt auch $z_1 \succcurlyeq_K z_3$.

Außerdem ist $\succcurlyeq_K$ verträglich mit der Vektorraumstruktur, das heißt, für $t_1 \geq 0$ und $t_2 \geq 0$ hat man:

4. Gilt $t \geq 0, z_1 \succcurlyeq_K z_2$, dann gilt auch $t z_1 \succcurlyeq_K t z_2$.
5. Gilt $z_1 \succcurlyeq_K z_2$ und $z_3 \succcurlyeq_K z_4$, dann gilt auch $z_1 + z_3 \succcurlyeq_K z_2 + z_4$.

4.2 Regularitätsbedingungen

4.2.1 LICQ bei Gleichungsnebenbedingungen

Natürlich stellt sich nun die Frage, wie man denn praktisch die Guinard- oder gleich die Abadie-Bedingung nachprüfen kann. Zumindest für Gleichheitsnebenbedingungen gibt es dafür eine einfache Antwort. Wir betrachten weiterhin Optimierungsprobleme der Form (4.2), allerdings nur für Kegel, die aus einem einzigen Punkt bestehen. Konkreter seien X und Y zwei normierte Räume, $U \subset X$ eine offene Menge, $F: U \rightarrow \mathbb{R}$ eine differenzierbare Funktion und $\Phi: U \rightarrow Y$ eine differenzierbare Abbildung. Wir betrachten das Optimierungsproblem

$$\begin{aligned} &\text{Minimiere} && F(x) \\ &\text{u. d. N.} && \Phi(x) = 0, \\ &&& x \in U. \end{aligned} \tag{4.16}$$

Die zulässige Menge ist also

$$M = \{x \in U \mid \Phi(x) = 0\}.$$

Definition 4.44 (LICQ)

Wir sagen, am Punkt $a \in M$ ist die *Bedingung der linearen Unabhängigkeit* (englisch: *linear independence constraint qualification*, kurz *LICQ*) für das Problem (4.16) erfüllt, genau dann wenn Φ stetig differenzierbar bei a ist und wenn $D\Phi(a): X \rightarrow Y$ surjektiv ist.

Der folgende Satz ist das klassische Resultat von Lagrange über die Multiplikatorregel:

Satz 4.45 (LICQ + $K = \{0\} \Rightarrow$ Abadie)

Seien X und Y beides Banachräume und sei bei $a \in M$ die LICQ erfüllt. Dann ist auch die Abadie CQ bei a erfüllt.

BEWEIS. Wir verwenden dazu den Satz über implizite Funktionen (siehe [15, Kapitel 4]) und daher ist es wesentlich, dass sowohl X und Y Banachräume sind. Denn dann folgt aus dem Satz über die offene Abbildung, dass $D\Phi(a): X \rightarrow Y$ eine offene Abbildung ist, dass also $D\Phi(a)$ offene Menge auf offene Mengen abbildet. Mehr Detail dazu findet man wie gesagt in [15], oder – in deutlich allgemeinerer, aber vielleicht nicht ganz so anschaulicher Darstellung in Abschnitt 4.3.2 und Abschnitt 4.3.3.

Für $x \in X$ sei $d(x, M) := \inf \{\|x - y\| \mid y \in M\}$ der minimale Abstand zur zulässigen Menge. Der Satz über implizite Funktionen besagt im Wesentlichen, dass es eine offene Umgebung $W \subset X$ von a und ein $\gamma \geq 0$, so dass gilt (vgl. Satz 4.69):

$$d(x, M) \leq \gamma \|\Phi(x)\|.$$

Sei $u \in X$. Für $t > 0$ ergibt die Definition der Fréchet-Ableitung, dass

$$\|\Phi(x)\| \leq \|\Phi(a) + D\Phi(a)(x - a)\| + \|x - a\| \varrho(x)$$

mit einer Funktion $\varrho: X \rightarrow [0, \infty]$, die stetig ist bei a und $\varrho(a) = 0$ erfüllt. Sei nun $u \in T_a^{\text{lin}} M = \ker(D\Phi(a))$. Wegen $\Psi(a) = 0$ gilt dann für $t > 0$ und $x = a + tu$, dass

$$\|\Phi(x)\| = \|\Phi(a + tu)\| \leq \|tu\| \varrho(a + tu).$$

Der Satz über implizite Funktionen sagt uns nun, dass

$$d(a + tu, M) \leq \gamma \|\Phi(a + tu)\| \leq \gamma \|tu\| \varrho(a + tu)$$

gelten muss. Es gibt also ein $w = w(t)$ mit $\|w(t)\| \leq \gamma t \|u\| \varrho(a + tu)$, so dass $a + tu + w(t) \in M$. Nun können wir

$$t_k := \frac{1}{k}, \quad x_k := a + t_k u + w(t_k) \in M$$

wählen und beobachten

$$\lim_{k \rightarrow \infty} \frac{x_k - a}{t_k} = \lim_{k \rightarrow \infty} \left(u + \frac{w(t_k)}{t_k} \right) = u.$$

Also ist $u \in T_a M$. Da wir dies für alle $u \in T_a^{\text{lin}} M$ tun können, ist also die Abadie-Bedingung erfüllt. $\square$

Folgerung 4.46 (LICQ + $K = \{0\} \Rightarrow \text{KKT}$)

Seien X und Y beides Banachräume. Sei $x_* \in M$ ein lokales Minimum von (4.16), bei dem die LICQ erfüllt ist. Dann ist x_* auch ein KKT-Punkt. Insbesondere gibt es einen Lagrange-Multiplikator $\lambda_* \in Y'$ so dass folgenden KKT-Bedingungen erfüllt sind:

$$\begin{aligned} DF(x_*) + \lambda_* D\Phi(x_*) &= 0, \\ \Phi(x_*) &= 0. \end{aligned}$$

BEWEIS. Aus Satz 4.45 wissen wir, dass die Abadie-Bedingung und damit auch die Guinard-Bedingung erfüllt sind. Da $K = \{0\}$ nur aus einem Punkt besteht, ist $T_0 K = \{0\}$ und $D\Phi(x_*)'(K) = \{0\} \subset X'$. Das ist aber mit Sicherheit eine schwach-*-abgeschlossene Menge (Beweis?). Also folgt aus Satz 4.28, dass x_* ein KKT-Punkt sein muss.

Die KKT-Bedingungen erhält man nun aus (4.3)–(4.6): Für die Polare des Restriktionskegels $K = \{0\}$ gilt $\text{pol}(K) = Y$. Also ist $\lambda_* \in \text{pol}(K)$ (vgl. (4.5)) gar keine Einschränkung. Und (4.6) ist trivialerweise erfüllt wegen $\Phi(x_*) \in K = \{0\}$. $\square$

4.2.2 Mangasarian-Fromovitz Regularitätsbedingung (MFCQ)

Seien X, Y und Z Banachräume, $U \subset X$ eine offene Menge und $K \subset Z$ ein abgeschlossener, konvexer Kegel. Ferner seien mit $\Phi: U \rightarrow Y$ und $\Psi: U \rightarrow Z$ zwei differenzierbare Abbildungen gegeben. Dazu betrachten wir den Zulässigkeitsbereich

$$M := \{ x \in U \mid \Phi(x) = 0 \text{ und } \Psi(x) \in K \}.$$

Das heißt, wir betrachten das Optimierungsproblem

$$\begin{aligned} &\text{Minimiere} && F(x) \\ &\text{u. d. N.} && \Phi(x) = 0, \\ &&& \Psi(x) \in K, \\ &&& x \in U. \end{aligned} \tag{4.17}$$

Um dies in das Setting von (4.2) zu bringen, definieren wir $\tilde{\Psi} := (\Phi, \Psi): U \rightarrow \tilde{Z} := Y \oplus Z$, $\tilde{C} := X$ und $\tilde{K} := \{0\} \times K$. Dann ist der Zulässigkeitsbereich gegeben durch

$$M = \{ x \in U \mid \tilde{\Psi}(x) \in \tilde{K} \}.$$

Definition 4.47 (MFCQ)

Wir sagen, für das Problem (4.17) ist am Punkt $a \in M$ die *Mangasarian-Fromovitz-Regularitätsbedingung* (MFCQ) erfüllt, genau dann wenn gilt:

1. Die Abbildungen Φ und Ψ sind bei a stetig differenzierbar.
2. Das Differential $D\Phi(a): X \rightarrow Y$ ist surjektiv.
3. Es gibt ein $v \in X$ mit $D\Phi(a)v = 0$ und $D\Psi(a)v \in (T_{\Psi(a)}K)^\circ$, d. h., $D\Psi(a)v$ liegt im topologischen Inneren des Tangentialkegels $T_{\Psi(a)}K$.

Bemerkung 4.48 Man beachte, dass die letzte Bedingung aus Definition 4.47 impliziert, dass der Kegel K ein nichtleeres Inneres besitzen muss. Das ist bei manchen Problemen eine echte Einschränkung.

Aufgabe 4.49 (MFCQ + polyhedrisch)

Nehmen wir an, wir haben einen polyedrischen Kegel vorliegen, also dass

$$K := \bigcap_{j=1}^m L_{\zeta_j}^-(0), \quad \text{mit } \zeta_1, \dots, \zeta_m \in Z'.$$

Zeigen Sie, dass für das Problem (4.10) die MFCQ am zulässigen Punkt $a \in M$ äquivalent ist zu:

1. Das Differential $D\Phi(a): X \rightarrow Y$ ist surjektiv.

2. Die Abbildungen Φ und Ψ sind bei a stetig differenzierbar.

3'. Es gibt einen Vektor $v \in X$ so dass $D\Phi(a)v = 0$ und so dass für alle $j \in \{1, \dots, m\}$ mit $\Psi_j(a) = 0$ auch $D\Psi_j(a)v < 0$ gilt.

Zeigen Sie dazu, für den Kegel

$$T_z K = \{ w \in Z \mid \forall j \in \{1, \dots, m\} \text{ mit } \langle \zeta_j, z \rangle = 0: \langle \zeta_j, w \rangle \leq 0 \} \quad \text{und} \\ (T_z K)^\circ = \{ w \in Z \mid \forall j \in \{1, \dots, m\} \text{ mit } \langle \zeta_j, z \rangle = 0: \langle \zeta_j, w \rangle < 0 \}.$$

Satz 4.50 (MFCQ + polyhedrisch $\Rightarrow$ Abadie)

Seien zusätzlich X und Y Banachräume und sei $K \subset Z$ polyhedrisch. Dann impliziert die MFCQ bei a auch die Abadie-Bedingung.

BEWEIS. Sei $u \in T_a^{\text{lin}} M$ und $v \in X$ so wie in Definition 4.47. Für $s > 0$ definieren wir $u(s) := u + s v$. Als nächstes zeigen wir, dass $u(s) \in T_a M$. Weil $T_a M$ abgeschlossen ist und $u(s)$ für $s \searrow 0$ gegen u konvergiert, folgt daraus die Behauptung $T_a^{\text{lin}} M \subset T_a M$.

Zunächst beobachten wir, dass $D\Phi(a)u(s) = 0$ gilt. Analog zu Satz 4.45 haben wir daher

$$\|\Phi(a + t u(s))\| \leq \|t u(s)\| o(a + t u(s)).$$

Aufgrund des Satzes über implizite Funktionen gibt es nun ein $w(t) \in X$ mit

$$\Phi(a + t u(s) + w(t)) = 0 \quad \text{und} \quad \|w(t)\| \in o(\|t u(s)\|) \quad \text{for } t \searrow 0.$$

Mit

$$t_k := \frac{1}{k} \quad \text{und} \quad x_k := a + t_k u(s) + w(t_k)$$

gilt dann wie gehabt

$$\Phi(x_k) = 0, \quad \lim_{k \rightarrow \infty} x_k = a \quad \text{und} \quad \lim_{k \rightarrow \infty} \frac{x_k - a}{t_k} = u(s).$$

Es bleibt zu zeigen, dass für hinreichend große $k \in \mathbb{N}$ auch $\Psi(x_k) \in K$ gilt, denn dann ist garantiert, dass $x_k \in M$ ein zulässiger Punkt ist und dass $u(s) \in T_a M$. Der Kegel K sei gegeben durch

$$K := \bigcap_{j=1}^m L_{\zeta_j}^-(0), \quad \text{mit } \zeta_1, \dots, \zeta_m \in Z'.$$

Wir zeigen nun, dass es für jedes $j \in \{1, \dots, m\}$ ein $k(j) \in \mathbb{N}$ gibt, so dass $\langle \zeta_j, \Psi(x_k) \rangle < 0$ für alle $k > k(j)$ gilt. Wir unterscheiden dazu zwei Fälle:

Fall 1: $\langle \zeta_j, \Psi(a) \rangle < 0$. Dann gibt es so ein $k(j)$ aufgrund der Stetigkeit von ζ_j und Ψ und weil x_k gegen a konvergiert.

Fall 2: $\langle \zeta_j, \Psi(a) \rangle = 0$. Weil Ψ Fréchet-differenzierbar ist, hat man

$$\begin{aligned} \langle \zeta_j, \Psi(x_k) \rangle &= \langle \zeta_j, \Psi(a) \rangle + s t_k \langle \zeta_j, D\Psi(a) u \rangle + t_k \langle \zeta_j, D\Psi(a) v \rangle + \langle \zeta_j, D\Psi(a) w(t_k) \rangle + o(t_k) \\ &= s t_k \langle \zeta_j, D\Psi(a) u \rangle + t_k \langle \zeta_j, D\Psi(a) v \rangle + o(t_k). \end{aligned}$$

Aus [Aufgabe 4.49](#) und $D\Psi(a) u \in T_{\Psi(a)}K$ folgt nun $\langle \zeta_j, D\Psi(a) u \rangle \leq 0$. Analog folgt aus [Aufgabe 4.49](#) und $D\Psi(a) v \in (T_{\Psi(a)}K)^\circ$, dass $\langle \zeta_j, D\Psi(a) v \rangle < 0$ gelten muss. Wir erhalten

$$\langle \zeta_j, \Psi(x_k) \rangle \leq t_k \langle \zeta_j, D\Psi(a) v \rangle + o(t_k).$$

Und für hinreichend große $k > k(j)$ wird dies dominiert durch $t_k \langle \zeta_j, D\Psi(a) v \rangle < 0$.

Da m endlich ist, sind alle diese Ungleichungen für $k > \max \{k(1), \dots, k(m)\}$ gleichzeitig erfüllt. Damit ist also $\Psi(x_k)$ für späte k sogar ein innerer Punkt von K . Dies schließt den Beweis ab. $\square$

Folgerung 4.51 (MFCQ + polyhedrisch $\Rightarrow$ KKT)

Ist der Kegel K polyhedrisch und ist a ein lokaler Minimierer von Problem (4.17), dann gibt es Lagrange-Multiplikatoren $\lambda_* \in Y'$ und $\mu_{*j} \in \mathbb{R}$, $j \in \{1, \dots, m\}$, so dass die KKT-Bedingungen (4.11)–(4.15) erfüllt sind.

BEWEIS. Wegen [Satz 4.50](#) ist also Abadie- und damit die Guinard-Bedingung erfüllt. Also folgt die Notwendigkeit der KKT-Bedingungen aus [Satz 4.36](#). $\square$

4.2.3 LICQ bei Ungleichungsnebenbedingungen

Wir betrachten das Problem (4.10).

Definition 4.52 Für einen zulässigen Punkt $x \in U \subset X$ des Optimierungsproblems (4.10) und $j \in \{1, \dots, m\}$ nennen wir die Nebenbedingung Ψ_j *aktiv*, falls $\Psi_j(x) \geq 0$; andernfalls nennen wir sie *inaktiv*. Die *geordnete* Menge der Indizes der aktiven Nebenbedingungen bezeichnen wir als *aktive Menge*. Die *inaktive Menge* der inaktiven Nebenbedingungen mit $\mathcal{I}(x)$.

Definition 4.53 (LICQ)

Für einen Punkt $a \in M$ definieren wir den Raum $Z_{\mathcal{A}(a)} := \mathbb{R}^{\#(\mathcal{A}(a))}$, wobei $\#(\mathcal{A}(a))$ die Kardinalität von $\mathcal{A}(a)$ bedeute, und die Abbildung

$$\Psi_{\mathcal{A}(a)}: U \rightarrow Z_{\mathcal{A}(a)}, \quad \Psi_{\mathcal{A}(a)}(x) := \left(\langle \zeta_j, \Psi(x) \rangle \right)_{j \in \mathcal{A}(a)},$$

Wir sagen, für das Problem (4.17) ist am Punkt $a \in M$ die *Regularitätsbedingung der linearen Unabhängigkeit* (englisch: *linear independence constraint qualification*, kurz *LICQ*) erfüllt, genau dann wenn gilt:

1. Die Abbildungen Φ und Ψ sind bei a stetig differenzierbar.

2. Die lineare Abbildung

$$\begin{pmatrix} D\Phi(a) \\ D\Psi_{\mathcal{A}(a)}(a) \end{pmatrix}: X \rightarrow Y \oplus Z_{\mathcal{A}(a)}$$

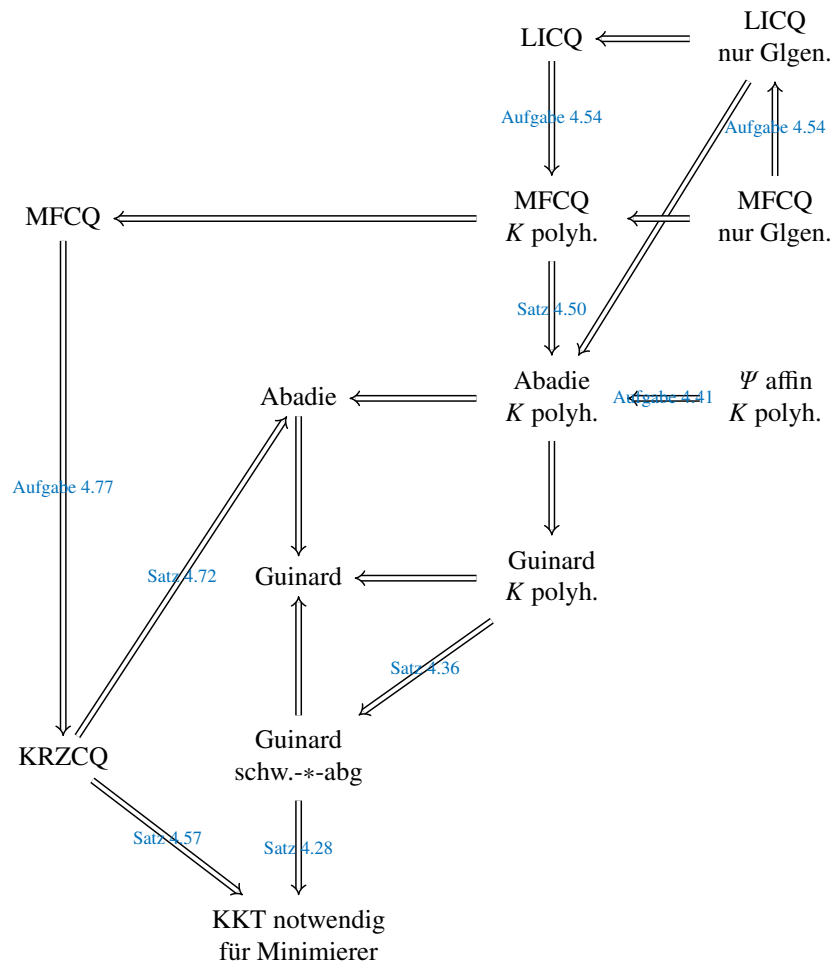
ist surjektiv.

Aufgabe 4.54 (LICQ $\Rightarrow$ MFCQ + polyhedrisch)

Nutzen Sie [Aufgabe 4.49](#), um zu zeigen, dass für das Problem (4.10) die LICQ die MFCQ impliziert. Im Falle, dass $K = Z$ (wenn also gar keine Ungleichungsnebenbedingungen vorhanden sind), sind LICQ und MFCQ äquivalent.

4.2.4 Zusammenfassung

Wie wir gesehen haben, gibt es vielfältige Implikationen zwischen den unterschiedlichen Regularitätsbedingungen. Da es schwer ist, dabei den Überblick zu behalten, hier ein Implikationsgraph für die Resultate, die wir bisher gezeigt haben. Zur Einfachheit nehmen wir an, dass X , Y und Z alle Banachräume sind und dass die Abbildungen $\Phi: U \subset X \rightarrow Y$ und $\Psi: U \subset X \rightarrow Z$ stetig differenzierbar sind bei den zu untersuchenden Punkten. Außerdem nehmen wir hier einen Vorgriff auf [Abschnitt 4.3](#) über die Kurcyusz-Robinson-Zowe- Regularitätsbedingung (KRZCQ). Sie ist eine Verallgemeinerung der MFCQ. Ihre Behandlung ist aber sehr technisch, daher empfehlen wir diesen Abschnitt nur wirklich für sehr interessierte Leser. Trotzdem wollen wir sie hier Einordnen, insbesondere, da wir mit ihrer Hilfe einsehen können, dass LICQ und MFCQ auch ohne polyhedrische Kegel hinreichend dafür sind, dass die KKT-Bedingungen notwendig sind (KKT).



4.3 Kurcyusz-Robinson-Zowe-Regularitätsbedingung (KRZCQ)

Hier beschäftigen wir uns mit etwas allgemeineren Optimierungsproblem:

Seien X und Z Banachräume, sei $U \subset X$ eine offene Menge, $C \subset X$ eine abgeschlossene, konvexe Menge und $K \subset Z$ ein abgeschlossener, konvexer Kegel. Zu den differenzierbaren Abbildungen $F: U \rightarrow \mathbb{R}$ und $\Psi: U \rightarrow Z$ betrachten wir das Optimierungsproblem

$$\begin{aligned} &\text{Minimiere} && F(x) \\ &\text{u. d. N.} && \Psi(x) \in K, \\ &&& x \in C, \\ &&& x \in U. \end{aligned} \tag{4.18}$$

Zusätzlich zu (4.2) führen wir also noch eine konvexe Nebenbedingung hinzu. Die zulässige Menge ist also

$$M := \{x \in U \cap C \mid \Psi(x) \in K\} = C \cap \Psi^{-1}(K).$$

Der *linearisierte Tangentialkegel* an M am Punkt $x \in X$ hat nun die Form

$$T_x^{\text{lin}} M := \{u \in T_x C \mid D\Psi(x)u \in T_{\Psi(x)} K\} = T_x C \cap D\Psi(x)^{-1}(T_{\Psi(x)} K).$$

Immernoch gilt stets $T_x M \subset T_x^{\text{lin}} M$ und bei Gleichheit sagen wir wieder, die *Abadie-Bedingung* sei erfüllt. Wir erhalten einen nur wenig von Definition 4.10 abweichenden Begriff von Karush-Kuhn-Tucker-Punkten:

Definition 4.55 Wir nennen einen Punkt $x_* \in X$ einen *Karush-Kuhn-Tucker-Punkt* oder kurz *KKK-Punkt* genau dann, wenn es sogenannte *Lagrange-Multiplikatoren* $\kappa_* \in X'$ und $\mu_* \in Z'$ gibt, so dass die *Karush-Kuhn-Tucker-Bedingungen* erfüllt sind:

$$DF(x_*) + \kappa_* + \mu_* D\Psi(x_*) = 0, \tag{4.19}$$

$$x_* \in C, \tag{4.20}$$

$$\Psi(x_*) \in K, \tag{4.21}$$

$$\kappa_* \in \text{pol}(C - x_*), \tag{4.22}$$

$$\mu_* \in \text{pol}(K), \tag{4.23}$$

$$\langle \mu_*, \Psi(x_*) \rangle = 0. \tag{4.24}$$

Definition 4.56 Sei $a \in M$ ein zulässiger Punkt des Optimierungsproblems (4.2). Wir sagen, bei a ist die *Kurcyusz-Robinson-Zowe-Regularitätsbedingung* erfüllt genau dann, wenn Ψ bei a stetig differenzierbar ist und wenn gilt:

$$Z = D\Psi(a)(\text{cone}(C - a)) - \text{cone}(K - \Psi(a)). \tag{4.25}$$

Das heißt: Für jedes $z \in Z$ kann man ein $u \in \text{cone}(C - a)$ und ein $w \in \text{cone}(K - \Psi(a))$ finden, so dass gilt:

$$z = D\Psi(a)u - w.$$

Es ist unser Ziel in diesem Abschnitt, das folgende Theorem zu beweisen:

Satz 4.57 (KRZCQ $\Rightarrow$ KKT) *Sei $x_* \in M$ ein lokales Minimum von (4.2), bei dem die Kurcyusz-Robinson-Zowe-Regularitätsbedingung erfüllt ist. Dann ist x_* auch ein Karush-Kuhn-Tucker-Punkt im Sinne von Definition 4.55.*

Die Kurcyusz-Robinson-Zowe-Regularitätsbedingung ist bemerkenswert, weil sie nicht explizit verlangt, dass $D\Psi(x)'(\text{pol}(T_{\Psi(x)}K))$ abgeschlossen ist, andererseits aber schon ausreicht, um die KKT-Bedingungen zu notwendigen Optimalitätsbedingungen zu machen. Andere Regularitätsbedingungen wie die Mangasarian-Fromovitz-Regularitätsbedingungen (siehe Definition 4.47) benötigen für ihre Formulierung, dass K ein nichtleeres Inneres hat. Auch das wird hier nicht benötigt.

Bemerkung 4.58 Es ist sicher hilfreich, einen einfachen Fall ohne Ungleichungsbedingungen durchzuspielen. Dazu nehmen wir an, dass $C = X$ und dass $K = \{0\} \subset Z$ gelten. Die zulässige Menge ist also

$$M = \{x \in U \mid \Psi(x) = 0\}.$$

Für $a \in M$ gilt $\text{cone}(C - a) = X$ und $\text{cone}(K - \Psi(a)) = \text{cone}(\{0\}) = \{0\}$. Folglich schmilzt die reichlich komplizierte Kurcyusz-Robinson-Zowe-Regularitätsbedingung zusammen auf:

$$D\Psi(a)(X) = Z,$$

das heißt, $D\Psi(a)$ ist surjektiv. Das ist aber nichts weiter als die LICQ-Bedingung für gleichungsrestringierte Problem, siehe Abschnitt 4.2.1. Nach dem Satz von der offenen Abbildung aus der Funktionalanalysis³ bedeutet dies insbesondere, dass $D\Psi(a)$ eine offene Abbildung ist. Es folgt aus dem Satz über die implizite Abbildung⁴, dass die Abadie-Bedingung für jedes $a \in U$ erfüllt ist, d. h., dass $T_a M = T_a^{\text{lin}} M$ gilt. Den ausführlichen Beweis hierfür haben wir in Satz 4.45 geführt.

Für den allgemeinen Fall werden wir nun diese Argumentation im Wesentlichen nachturnen.

4.3.1 Existenz von Lagrange-Multiplikatoren

Beim Beweis von Satz 4.57 spielen Kegel mit nichtleerem Inneren aber auch eine wichtige Rolle. Bevor wir die wesentliche Beweisidee angreifen, benötigen wir noch ein Hilfsresultat aus der Funktionalanalysis:

Satz 4.59 (Trennungssatz)

Sei X ein reeller, normierter Raum, $V \subset X$ eine offene und konvexe Menge mit $0 \notin V$. Dann gibt es ein $\xi \in X'$ mit $\langle \xi, x \rangle > 0$ für alle $x \in V$.

³Wir erhalten ihn später als Folgerung einer verallgemeinerten Version, siehe Satz 4.62 und Folgerung 4.63.

⁴Auch diesen beweisen wir "neu" als Folgerung aus einer Verallgemeinerung, siehe Satz 4.69 und Aufgabe 4.70.

Dies ist eine Folgerung aus dem *Satz von Hahn-Banach*; diesen verwendet man üblicherweise auch, um [Satz 2.34](#) zu beweisen. Der Beweis des Satzes von Hahn-Banach ist nicht besonders schwer nachzuvollziehen, ist aber etwas technisch und benötigt einen tiefen Griff in die Eingeweide der modernen Mathematik (für endlichdimensionale Räume genügt eine Induktion; im Allgemeinen zieht man das zum Auswahlaxiom äquivalente Lemma von Zorn heran). Deshalb soll er hier nicht vorgeführt werden. Interessierte seien auf die Lehrbuchliteratur verwiesen, z. B. auf [\[13\]](#).

Satz 4.60 (Kurcyusz-Zowe, [\[16\]](#)) Sei x_* ein lokales Minimum, an dem die Abadie-Bedingung erfüllt ist und der Kegel

$$D\Psi(x_*)(\text{cone}(C - x_*) - \text{cone}(K - \Psi(x_*)))$$

besitze 0 als inneren Punkt. Dann ist x_* ein KKT-Punkt.

BEWEIS. Sei $\varepsilon > 0$, so dass $B(0; \varepsilon) \subset D\Psi(x_*)(\text{cone}(C - x_*) - \text{cone}(K - z_*))$. Wir definieren $z_* := \Psi(x_*)$ und den konvexen Kegel $L \subset \mathbb{R} \times Z$ durch

$$L := \left\{ \left(DF(x_*)u + s, D\Psi(x_*)u - w \right) \mid u \in \text{cone}(C - x_*), w \in \text{cone}(K - z_*), s \geq 0 \right\}.$$

Versehen mit der Norm $\|(t, z)\| := |t| + \|z\|$ ist $\mathbb{R} \times Z$ ein normierter Raum. Mit $c > \|DF(x_*)\|$ liegt nach Voraussetzung die offene Menge $]cs, \infty[\times B(0; \varepsilon) \subset \mathbb{R} \times Z$ noch komplett in L . Das heißt, L° ist ein nichtleerer, konvexer Kegel (siehe [Aufgabe 4.74](#)).

Der Punkt $(0, 0)$ liegt *nicht* in L° ; andernfalls wäre $(0, -t) \in L^\circ$ für hinreichend kleines $t > 0$. Dann gäbe es ein $u \in \text{cone}(C - x_*) \subset T_{x_*}C$, ein $w \in \text{cone}(K - z_*) \subset T_{z_*}K$ und ein $s \geq 0$ mit $D\Psi(x_*)u = w$ und $DF(x_*)u + s < 0$. Daraus folgt aber wegen der Abadie-Bedingung $u \in T_{x_*}^{\text{lin}}M \subset T_{x_*}M$ und $DF(x_*)u < 0$. Das wäre aber ein Widerspruch zur notwendigen Optimalitätsbedingung erster Ordnung aus [Satz 4.3](#).

Aus dem Trennungssatz [Satz 4.59](#) folgt, dass es ein $(\alpha, \mu) \in \mathbb{R} \times Z' = (\mathbb{R} \times Z)'$ gibt, so dass

$$\langle (\alpha, \mu), (t, z) \rangle > 0 = \langle (\alpha, \mu), (0, 0) \rangle \quad \text{für alle } (t, z) \in L^\circ.$$

Insbesondere ist dann aber auch $(\alpha, \mu) \neq 0$. Weil L° dicht liegt in L (siehe [Aufgabe 4.74](#)), gilt insbesondere

$$\langle (\alpha, \mu), (t, z) \rangle \geq 0 = \langle (\alpha, \mu), (0, 0) \rangle \quad \text{für alle } (t, z) \in L.$$

Das impliziert aber

$$\alpha(DF(x_*)u + s) + \langle \mu, D\Psi(x_*)u - w \rangle \geq 0$$

für alle $u \in \text{cone}(C - x_*)$, $w \in \text{cone}(K - z_*)$, und $s \geq 0$.

Angenommen, $\alpha = 0$. Dann verschwindet μ aber auf dem gesamten Ball $B(0; \varepsilon)$; demzufolge würde auch $\mu = 0$ gelten. Das ist aber ein Widerspruch zu $(\alpha, \mu) \neq 0$.

Wir setzen nun $\mu_* := \alpha^{-1} \mu$ und erhalten

$$DF(x_*)u + s + \langle \mu_*, D\Psi(x_*)u - w \rangle \geq 0$$

für alle $u \in \text{cone}(C - x_*)$, $w \in \text{cone}(K - z_*)$ und $s \geq 0$.

Für $s = 0$ und $w = 0$ erhalten wir dann $DF(x_*)u + \langle \mu_*, D\Psi(x_*)u \rangle \geq 0$ für alle $u \in \text{cone}(C - x_*)$. Damit ist schon (4.19) erfüllt:

$$DF(x_*) + \kappa_* + \mu_* D\Psi(x_*) = 0 \quad \text{mit einem } \kappa_* \in \text{pol}(\text{cone}(C - z_*)).$$

Für $u = 0$ und $s = 0$ erhalten wir $\langle \mu_*, w \rangle \leq 0$ für alle $w \in \text{cone}(K - z_*)$, also $\mu_* \in \text{pol}(\text{cone}(K - z_*))$. Das bedeutet aber gerade $\mu_* \in \text{pol}(K)$ und $\langle \mu_*, \Psi(x_*) \rangle = 0$ (siehe Aufgabe 4.40). Das sind gerade (4.23) und (4.24). Wegen $x_* \in M$ gilt natürlich auch $\Psi(x_*) \in K$, also (4.21). $\square$

BEWEIS (von Satz 4.57). Die Offenheitsvoraussetzung im obigen Satz ist genau dann erfüllt, wenn die Kurcyusz-Robinson-Zowe-Regularitätsbedingung erfüllt ist. Dazu werden wir im nächsten Abschnitt eine verallgemeinerte Version des Satz über offene Abbildung (Satz 4.62 bzw. Lemma 4.65) herleiten. Für Satz 4.57 müssen wir also “lediglich” zeigen, dass die Kurcyusz-Robinson-Zowe-Regularitätsbedingung auch die Abadie-Bedingung impliziert. Genau dafür benötigen wir eine Verallgemeinerung des Satzes über implizite Funktionen (Satz 4.69): Dieser hilft uns, zu einer gegebenen Richtung $u \in T_a^{\text{lin}} M$ Folgen $x_n \in M$ und $t_n \searrow 0$ mit

$$\frac{x_n - a}{t_n} \xrightarrow{n \rightarrow \infty} u$$

zu konstruieren. Im Wesentlichen wird dies dadurch passieren, dass wir beobachten, dass $a + t_n u$ schon “fast” in M liegt und wir dies nur durch eine Störung der Größenordnung $o(t_n)$ beheben können (siehe Satz 4.72). $\square$

4.3.2 Satz von der offenen Abbildung

Ohne uns Gedanken darüber zu machen, warum er so heißt (bei Interesse findet man die Details dazu z. B. in [13]), behandeln wir den Baireschen Kategoriensatz.

Satz 4.61 (Bairescher Kategoriensatz)

Sei (X, d) ein vollständiger metrischer Raum und $(C_n)_{n \in \mathbb{N}}$ eine Folge abgeschlossener Mengen mit $X \subset \bigcup_{n \in \mathbb{N}} C_n$. Dann hat mindestens eine dieser Mengen ein nichtleeres Inneres, das heißt, es gibt ein $\varepsilon > 0$, ein $x \in X$ und ein $n \in \mathbb{N}$, so dass $B(x; \varepsilon) \subset C_n$.

BEWEIS. Wir definieren die offenen Mengen $U_n := X \setminus C_n$ und beobachten $\bigcap_{n \in \mathbb{N}} U_n = \emptyset$. Angenommen, keines der C_n hat einen inneren Punkt. Das bedeutet: Für alle $\varepsilon > 0$, für alle $x \in X$ und für alle $n \in \mathbb{N}$ ist $B(x; \varepsilon) \cap U_n \neq \emptyset$. Beginnend mit einem beliebigem Startpunkt $x_1 \in U_1$ und einem beliebigen $0 < \varepsilon_1 < 1$ wählen wir nun rekursiv eine Folge $(x_n)_{n \in \mathbb{N}}$ mit

$$\varepsilon_{n+1} < \frac{1}{n+1} \quad \text{und} \quad B(x_{n+1}; \varepsilon_{n+1}) \subset \bar{B}(x_{n+1}; \varepsilon_{n+1}) \subset B(x_n; \varepsilon_n) \cap U_n \subset U_n.$$

Für jedes $n \geq m$ gilt nun $x_n \in B(x_m; \frac{1}{m})$, also $d(x_n, x_m) < \frac{1}{m}$. Damit ist $(x_n)_{n \in \mathbb{N}}$ eine Cauchyfolge und konvergiert gegen einen Punkt $x \in X$. Das führt aber zu folgendem *Widerspruch*:

$$x \in \bigcap_{n \in \mathbb{N}} \bar{B}(x_{n+1}; \varepsilon_{n+1}) \subset \bigcap_{n \in \mathbb{N}} U_n \subset \emptyset. \quad \square$$

Satz 4.62 (Verallgemeinerter Satz von der offenen Abbildung, [16])

Seien X und Z Banachräume, $C \subset X$ eine abgeschlossene und konvexe Menge, $K \subset Z$ ein abgeschlossener, konvexer Kegel und $A \in L(X; K)$ eine stetige, lineare Abbildung. Dann sind für $x \in C$ und $z \in K$ äquivalent:

1. Es gilt $Z = A(\text{cone}(C - x)) - \text{cone}(K - z)$.
2. Es gibt ein $\varepsilon > 0$ mit $\bar{B}(0; \varepsilon) \subset A((C - x) \cap \bar{B}(0; 1)) - ((K - z) \cap \bar{B}(0; 1))$

BEWEIS. Wir nutzen die Abkürzungen

$$\begin{aligned} L &:= A(\text{cone}(C - x)) - \text{cone}(K - z) \\ Q(r) &:= A((C - x) \cap \bar{B}(0; r)) - ((K - z) \cap \bar{B}(0; r)) \quad \text{für } r \geq 0. \end{aligned}$$

“Punkt 2 $\Rightarrow$ Punkt 1”: Sei $\varepsilon > 0$ mit $B(0; \varepsilon) \subset Q(1)$. Weil $Q(1) \subset L$ gilt und weil L ein Kegel ist, gilt auch

$$Z = \bigcup_{r>0} r B(0; \varepsilon) \subset \bigcup_{r>0} r L \subset L \subset Z.$$

“Punkt 1 $\Rightarrow$ Punkt 2” Dies ist der eigentliche Knackpunkt. Die Abschlüsse $\bar{Q}(r) := \overline{Q(r)}$ sind konvex (siehe Aufgabe 4.39). Außerdem erfüllen sie $\bar{Q}(r) = r \bar{Q}(1)$ und es gilt wegen Aufgabe 4.66

$$Z \subset L = \bigcup_{k \in \mathbb{N}} k \bar{Q}(1) = \bigcup_{k \in \mathbb{N}} \bar{Q}(k).$$

Nach dem Baireschen Kategoriensatz Satz 4.61 muss es daher ein $k \in \mathbb{N}$ geben, so dass $\bar{Q}(k)$ ein nichtleeres Inneres besitzt. Sei a ein innerer Punkt von $\bar{Q}(k)$. Weil die Mengen $(\bar{Q}(k))_{k \in \mathbb{N}}$ ganz Z überdecken, muss es ein $m \in \mathbb{N}$ geben mit $-a \in \bar{Q}(m)$. Folglich liegt $b := -\frac{k}{m}a$ in $\bar{Q}(k)$ und wegen $0 = a(t - 1) + t b$ mit $t := \frac{m}{k+m} \in]0, 1[$ ist 0 ein innerer Punkt von $\bar{Q}(k)$. Dann ist 0 aber auch ein innerer Punkt von $\bar{Q}(1)$ und es gibt ein $\varepsilon > 0$ mit $\bar{B}(0; 2\varepsilon) \subset \bar{Q}(1)$. Das ist aber noch *nicht ganz* das, was wir zeigen wollten: Wir müssen noch den Abschluss von $\bar{Q}(1)$ entfernen und das tun wir auf Kosten eines ε . Sei $z_0 \in \bar{B}(0; \varepsilon)$. Wir zeigen jetzt durch Induktion, dass es zu jedem $n \in \mathbb{N}$ ein $x_n \in (C - x) \cap \bar{B}(0; 1)$, ein $y_n \in (K - z) \cap \bar{B}(0; 1)$, und ein $r_n \in \bar{B}(0; 2^{-n} \varepsilon) \subset Z$ gibt mit

$$z_0 = A\left(\sum_{k=1}^n 2^{-k} x_k\right) - \left(\sum_{k=1}^n 2^{-k} y_k\right) + r_n \quad \text{für alle } n \in \mathbb{N}. \quad (4.26)$$

Induktionsanfang: Für $n = 1$ ist das gerade wegen

$$\bar{B}(0; \varepsilon) = \frac{1}{2} \bar{B}(0; 2\varepsilon) \subset \frac{1}{2} \bar{Q}(1) = \bar{Q}(\frac{1}{2}) \subset Q(\frac{1}{2}) + \bar{B}(0; \frac{1}{2}\varepsilon). \quad (4.27)$$

möglich.

Induktionsschritt: Ist $r_n \in \bar{B}(0; 2^{-n}\varepsilon)$, so folgt die Darstellung

$$r_n = A(2^{-(n+1)}x_{n+1}) + 2^{-(n+1)}y_{n+1} + r_{n+1}$$

mit $x_{n+1} \in (C - x) \cap \bar{B}(0; 1)$, $y_{n+1} \in (K - z) \cap \bar{B}(0; 1)$ und $r_{n+1} \in \bar{B}(0; 2^{-(n+1)}\varepsilon)$ aus

$$\bar{B}(0; 2^{-n}\varepsilon) = 2^{-n} \bar{B}(0; \varepsilon) \subset 2^{-n} \left(Q(\frac{1}{2}) + \bar{B}(0; \frac{1}{2}\varepsilon) \right) = Q(2^{-(n+1)}) + \bar{B}(0; 2^{-(n+1)}\varepsilon). \quad (4.28)$$

Mit der geometrischen Reihe rechnet man leicht nach, dass die beiden Summen in (4.26) Cauchyfolgen definieren. Weil $C - x$ konvex ist und den Nullpunkt enthält, gilt

$$\sum_{k=1}^n 2^{-k} x_k = \sum_{k=1}^n 2^{-k} x_k + t \cdot 0 \in C - x \quad \text{mit } t := 1 - \sum_{k=1}^n 2^{-k} \in]0, 1[.$$

Weil $(C - x) \cap \bar{B}(0; 1)$ und $(K - z) \cap \bar{B}(0; 1)$ abgeschlossen sind, konvergieren die Reihen gegen Elemente $x \in (C - x) \cap \bar{B}(0; 1)$ und $y \in (K - z) \cap \bar{B}(0; 1)$. Außerdem konvergiert r_n gegen 0 für $n \rightarrow \infty$ und wir erhalten im Limes

$$z_0 = Ax - y \in A((C - x) \cap \bar{B}(0; 1)) - ((K - z) \cap \bar{B}(0; 1)). \quad \square$$

Als Nebenprodukt erhalten wir zwei der wichtigsten Resultate aus der Funktionalanalysis.

Folgerung 4.63 (Satz von der offenen Abbildung) Seien X und Z Banachräume und $A \in L(X; Z)$ ein stetiger, surjektiver Operator. Dann ist A auch offen, das heißt, A bildet offene Mengen auf offene Mengen ab.

BEWEIS. Man wählt hierzu $C = X$ und $K = \{0\}$. $\square$

Folgerung 4.64 Seien X und Z Banachräume und $A \in L(X; Z)$. Ist A bijektiv, dann ist A schon stetig invertierbar.

Das für uns hier wichtigste Resultat ist dieses hier:

Lemma 4.65 (Zowe-Kurcyusz, [16]) Unter den Bedingungen von Satz 4.62 sind äquivalent:

1. $0 \in (A(C - x) - (K - z))^\circ$.
2. $0 \in (A(\text{cone}(C - x)) - \text{cone}(K - z))^\circ$.
3. $Z = A(\text{cone}(C - x)) - \text{cone}(K - z)$.

BEWEIS. “Punkt 1 $\Rightarrow$ Punkt 2”: Dies folgt aus $C - x \subset \text{cone}(C - x)$ und $K - z \subset \text{cone}(K - z)$.
 “Punkt 2 $\Rightarrow$ Punkt 3”: Dies erhält man mit geeignetem $\varepsilon > 0$ durch

$$Z = \bigcup_{r>0} B(0; \varepsilon) \subset \bigcup_{r>0} r(A(\text{cone}(C - x)) - \text{cone}(K - z)) = A(\text{cone}(C - x)) - \text{cone}(K - z)$$

“Punkt 3 $\Rightarrow$ Punkt 1”: Man wendet Satz 4.62. □

Aufgabe 4.66

Sei $C \subset X$ eine konvexe Menge im normierten Raum X und sei $x \in C$. Dann gilt

$$\text{cone}(C - x) = \bigcup_{n \in \mathbb{N}} n((C - x) \cap \bar{B}(0; 1)).$$

Aufgabe 4.67

Seien X und Z Banachräume und sei $A \in L(X; Z)$. Dann sind die folgenden beiden Aussagen äquivalent:

1. A ist injektiv und das Bild $\text{im}(A)$ ist abgeschlossen.
2. Es gibt ein $c > 0$ mit $\|A u\| \geq c \|u\|$ für alle $u \in X$.

Hinweis: Für “1 $\Rightarrow$ 2” überlege man sich, dass ein abgeschlossener Untervektorraum in einem Banachraum wieder ein Banachraum ist (mit der Einschränkung der Norm).

Für “2 $\Rightarrow$ 1” nutze man, dass jede Folge in $\text{im}(A)$, die konvergent ist, auch eine Cauchyfolge sein muss.

4.3.3 Satz über implizite Funktionen von Robinson

Lemma 4.68

Seien X und Z Banachräume, Y ein topologischer Raum, $\Omega \subset X$ offen und seien $L \subset X$ und $K \subset Z$ abgeschlossene, konvexe Kegel. Sei $\Psi_y: X \rightarrow Z$, $y \in Y$ eine Familie von differenzierbaren Abbildungen. Für die Punkte $a \in L \cap \Omega$ und $b \in Y$ sei erfüllt:

1. $\Psi_b(a) \in K$.
2. Regularitätsbedingung: $D\Psi_b(a)(L) - K = Z$.
3. Die Abbildungen $(x, y) \mapsto \Psi_y(x)$ und $(x, y) \mapsto D\Psi_y(x)$ sind stetig bei (a, b) .

Wir definieren die Mengen-wertige Abbildung

$$M: Y \rightarrow \mathfrak{P}(X), \quad M(y) := \{x \in C \cap \Omega \mid \Psi_y(x) \in K\}.$$

Dann gibt es Umgebungen U von a in Ω und V von b in Y und ein $\gamma \geq 0$ so dass gilt

$$M(y) \neq \emptyset \quad \text{und} \quad d(x, M(y)) \leq \gamma d(\Psi_y(x), K) \quad \text{für alle } x \in C \cap U \text{ und alle } y \in V.$$

BEWEIS. Wir nehmen an $K \neq Z$, denn sonst kann man $U = \Omega$, $V = Y$ und $\Lambda = 1$ wählen.

Wir möchten den Robinsonschen Fixpunktsatz ?? anwenden, um die Residualabschätzung in (??) auszunutzen. Daher müssen wir unser Problem $x \in M(y)$ zunächst in ein Fixpunktproblem umwandeln. Wir kürzen ab $A := D\Psi_b(a)$ und führen die folgende Mengenwertige Abbildung ein:

$$\Phi_y: X \rightarrow \mathfrak{P}(Z), \quad \Phi_y(x) = \begin{cases} \Psi_y(x) - K, & x \in L \cap \Omega, \\ \emptyset, & x \notin L \cap \Omega. \end{cases}$$

Zusätzlich definieren wir noch die “Ableitung”

$$D\Phi_y: X \rightarrow \mathfrak{P}(Z), \quad D\Phi_y(x)u = \begin{cases} D\Psi_y(x)u - K, & u \in L, \\ \emptyset, & u \notin L. \end{cases}$$

Suggestiv definieren wir noch die “Inverse“ der Ableitung $D\Phi_y(x)^{-1}: Z \rightarrow \mathfrak{P}(X)$ durch

$$D\Phi_y(x)^{-1}w := \{u \in X \mid w \in D\Phi_y(x)u\} = \{u \in L \mid w \in D\Psi_y(x)u - K\}.$$

Die Idee des Beweises besteht nun darin, die Iterationsabbildung des modifizierten Newton-Verfahrens wie folgt für Φ_y zu verallgemeinern:⁵

$$\Theta_y(x) := D\Phi_b(a)^{-1}(D\Psi_b(a)x - \Psi_y(x)).$$

Ausgeschrieben erhält man

$$\Theta_y(x) = \{u \in L \mid Ax - \Psi_y(x) \in Au - K\}. \quad (4.29)$$

Ein Punkt $x \in U$ ist also genau dann ein Fixpunkt von Θ_y (d. h. $x \in \Theta_y(x)$), wenn $x \in M(y)$ gilt. Deswegen können wir Θ_y auch als Abbildung $\Theta_y: L \rightarrow \mathfrak{P}(L)$ auf dem metrischen Raum L auffassen.

Wir möchten zeigen, dass Θ_y auf $\bar{B}(a; R)$ für hinreichend kleines $R > 0$ eine Kontraktion ist. Dazu untersuchen wir die Abbildung Φ_y zunächst etwas genauer und nutzen die Abkürzungen $c := \Psi_b(a)$ und

$$G(w) := D\Phi_b(a)^{-1}w = \{u \in L \mid w \in Au - K\}.$$

Behauptung 1: Es gilt $0 < \Lambda := \sup_{w \neq 0} \frac{d(0, G(w))}{\|w\|} < \infty$.

⁵Dass dies wirklich eine Verallgemeinerung ist, sieht man am besten, wenn man für den Moment $\Omega = L = X$ und $K = \{0\}$ einsetzt. Dann hat man nämlich $\Phi_y(x) = \{\Psi_y(x)\}$ und $D\Psi_y(x)u = \{D\Phi_y(x)u\}$ und

$$\Theta_y(x) = \{D\Psi_b(a)^{-1}(D\Psi_b(a)x - \Psi_y(x))\} = \{x - D\Psi_b(a)^{-1}\Psi_y(x)\}.$$

Aus der Regularitätsbedingung und dem Satz von der offenen Abbildung [Satz 4.62](#) folgt, dass es ein $\delta > 0$ geben muss mit

$$\bar{B}(0; \delta) \subset A(L \cap \bar{B}(0; 1)) - (K \cap \bar{B}(0; 1)).$$

Für jedes $w \in Z$ gilt daher

$$d(0, G(w)) = \inf \{ \|u\| \mid u \in L \text{ und } w \in A u - K \} \leq \frac{1}{\delta} \|w\|.$$

Es folgt $\Lambda < \infty$. Weil $G(w)$ stets abgeschlossen ist, würde aus $\Lambda = 0$ folgen, dass 0 in $G(w)$ liegen muss für *jedes* $w \in Z$. Das bedeutet aber $w \in A 0 - K$ für alle $w \in Z$ und daher $K = Z$. Das hatten wir aber eingangs ausgeschlossen. Damit ist Behauptung 1 bewiesen.

Behauptung 2: Für alle $w_1, w_2 \in Z$ gilt

$$\text{dist}(G(w_1), G(w_2)) \leq \Lambda \|w_1 - w_2\|. \quad (4.30)$$

Wir geben uns ein beliebiges $\varepsilon > 0$ vor. Sei $u_1 \in G(w_1)$. Mit Behauptung 1 findet man ein $u_2 \in G(w_2 - w_1)$ mit $\|u_2\| < \Lambda \|w_2 - w_1\| + \varepsilon$. Wie man leicht nachrechnet gilt

$$u_2 = u_2 + u_1 - u_1 \in G(w_2 - w_1) + G(w_1) - u_1 \subset G(w_2) - u_1.$$

Wir erhalten

$$d(u_1, \Phi(w_2)) = \inf \{ \|u - u_1\| \mid u \in G(w_2) \} = \inf \{ \|u_2\| \mid u_2 \in G(w_2) - u_1 \} < \Lambda \|w_1 - w_2\| + \varepsilon.$$

Analog sieht man auch $d(u_2, G(w_1)) < \Lambda \|w_1 - w_2\| + \varepsilon$ für beliebiges $u_2 \in G(w_2)$ ein. Da dies für beliebige $\varepsilon > 0$ gilt, zeigt dies schließlich auch Behauptung 2.

Behauptung 3: Für alle $x \in L \cap \mathcal{Q}$ und $y \in Y$ gilt $d(x, \Theta_y(x)) \leq \Lambda d(\Psi_y(x), K)$.

Wir beobachten zunächst für beliebes $w \in Z$:

$$G(w) = \{ u \in L \mid \exists \tilde{w} \in K: w + \tilde{w} \in A u - K \} = \bigcup_{\tilde{w} \in K} G(w + \tilde{w}) = G(w + K). \quad (4.31)$$

Für beliebiges $x \in L \cap \mathcal{Q}$ gilt $x \in G(A x)$. Dies beides zusammen mit Behauptung 2 ergibt

$$\begin{aligned} d(x, \Theta_y(x)) &= d(x, G(A x - \Psi_y(x))) \\ &\leq \text{dist}(G(A x), G(A x - \Psi_y(x))) \\ &= \text{dist}(G(A x), G(A x - \Psi_y(x) + K)) \\ &\leq \Lambda \text{dist}(\{A x\}, (A x - \Psi_y(x) + K)) \\ &= \Lambda \text{dist}(\{0\}, (K - \Psi_y(x))) = \Lambda d(\Psi_y(x), K). \end{aligned}$$

Behauptung 4: Zu jedem $q \in]0, 1[$ gibt es $R > 0$ mit $\bar{B}(a; R) \subset \Omega$ und $S > 0$, so dass

$$\text{dist}(\Theta_y(x_1), \Theta_y(x_2)) \leq q \|x_1 - x_2\| \quad \text{für alle } x_1, x_2 \in L \cap \bar{B}(a; R) \text{ und } y \in \bar{B}(b; S).$$

Wegen Behauptung 2 können wir nun für beliebige $x_1, x_2 \in L \cap \Omega$ nachrechnen:

$$\begin{aligned} \text{dist}(\Theta_y(x_1), \Theta_y(x_2)) &= \text{dist}(G(Ax_1 - \Psi_y(x_1)), G(Ax_2 - \Psi_y(x_2))) \\ &\leq \Lambda \|(Ax_1 - \Psi_y(x_1)) - (Ax_2 - \Psi_y(x_2))\| \\ &= \Lambda \|\Psi_y(x_2) - \Psi_y(x_1) - D\Psi_b(a)(x_2 - x_1)\|. \end{aligned}$$

Der Hauptsatz der Differential- und Integralrechnung liefert uns die Abschätzung

$$\begin{aligned} \|\Psi_y(x_2) - \Psi_y(x_1) - D\Psi_b(a)(x_2 - x_1)\| &\leq \int_0^1 \|D\Psi_y(x_1(1-t) + tx_2) - D\Psi_b(a)\| \|x_2 - x_1\| dt \\ &\leq \sup_{t \in [0,1]} \omega_y(x_1(1-t) + tx_2) \|x_1 - x_2\|. \end{aligned}$$

wobei $\omega_y(x) := \|D\Psi_y(x) - D\Psi_b(a)\| \rightarrow 0$ sei. Wegen der Stetigkeit von $(x, y) \mapsto D\Psi_y(x)$ gilt dabei $\omega_y(x) \rightarrow 0$ für $(x, y) \rightarrow (a, b)$. Zu einem gegebenen $q \in]0, 1[$ können wir also $R > 0$, $S > 0$ wählen mit $\bar{B}(a; R) \subset \Omega$ und so dass $\Lambda \omega_y(x) \leq q$ gilt für alle $x \in L \cap \bar{B}(a; R)$ und alle $y \in \bar{B}(b; S)$. Weil $L \cap \bar{B}(a; R)$ konvex ist, beweist dies Behauptung 3.

Anwendung des Fixpunktsatzes

Sei $\varepsilon > 0$ und $q := \frac{\varepsilon}{\Lambda + \varepsilon} \in]0, 1[$. Wir wählen $R > 0$ und $S > 0$ wie in Behauptung 4. Wegen $\Psi_y(x) \in K$ und der Stetigkeit von Ψ schaffen wir es, $0 < R_1 < R/2$ und $0 < S_1 < S$ zu finden mit

$$d(x, \Theta_y(x)) \leq \Lambda d(\Psi_y(x), K) < \frac{1}{2}(1-q)R \quad \text{für alle } x \in L \cap \bar{B}(a; R_1) \text{ und alle } y \in \bar{B}(b; S_1).$$

Fixiere $x \in L \cap \bar{B}(a; R_1)$ und $y \in \bar{B}(b; S_1)$. Mit dem Radius $R_2 := \frac{R}{2}$ sind nun wegen $L \cap \bar{B}(x; R_2) \subset L \cap \bar{B}(a; R)$ und Behauptung 4 die Voraussetzungen des Robinsonschen Fixpunktsatzes ?? erfüllt. Zusammen mit Behauptung 3 erhalten wir

$$d(x, M(y)) \leq \frac{1}{1-q} d(x, \Theta_y(x)) \leq \frac{\Lambda}{1-q} d(\Psi_y(x), K) = (\Lambda + \varepsilon) d(\Psi_y(x), K). \quad \square$$

Satz 4.69 (Satz über implizite Funktionen nach Robinson, [11])

Seien X und Z Banachräume, Y ein topologischer Raum, $\Omega \subset X$ offen und seien $C \subset X$ eine abgeschlossene, konvexe Menge und $K \subset Z$ ein abgeschlossener, konvexer Kegel. Sei $\Psi_y: X \rightarrow Z$, $y \in Y$ eine Familie von differenzierbaren Abbildungen. Für die Punkte $a \in C \cap \Omega$ und $b \in Y$ sei erfüllt:

1. $\Psi_b(a) \in K$.
2. Regularitätsbedingung: $D\Psi_b(a)(\text{cone}(C - a)) - \text{cone}(K - \Psi_b(a)) = Z$.

3. Die Abbildungen $(x, y) \mapsto \Psi_y(x)$ und $(x, y) \mapsto D\Psi_y(x)$ sind stetig bei (a, b) .

Wir definieren die Mengen-wertige Abbildung

$$M: Y \rightarrow \mathfrak{P}(X), \quad M(y) := \{x \in C \cap \Omega \mid \Psi_y(x) \in K\}.$$

Dann gibt es eine Umgebungen U von a in Ω und V von b in Y und ein $\gamma \geq 0$ so dass gilt

$$M(y) \neq \emptyset \quad \text{und} \quad d(x, M(y)) \leq \gamma d(\Psi_y(x), K) \quad \text{für alle } x \in L \cap U \text{ und alle } y \in V.$$

BEWEIS. Wir möchten [Lemma 4.68](#) anwenden und müssen dafür die Menge C irgendwie durch einen geeigneten, abgeschlossenen und konvexen Kegel ersetzen. Diesen Kegel definieren wir wie folgt

$$\widetilde{L} := \overline{\text{cone}}((C - a) \times \{1\}) = \{0\} \cup \{(x, s) \in X \times]0, \infty[\mid a + x/s \in C\}.$$

Außerdem definieren wir die offene Menge

$$\widetilde{\Omega} := \text{cone}((\Omega - a) \times \{1\}) \setminus \{0\} = \{(x, s) \in X \times]0, \infty[\mid a + x/s \in \Omega\},$$

die Abbildung

$$\widetilde{\Psi}_y: \widetilde{\Omega} \rightarrow Z, \quad \widetilde{\Psi}_y(x, s) := s \Psi_y(a + x/s)$$

und den Zulässigkeitsbereich

$$\begin{aligned} \widetilde{M}(y) &:= \{(x, s) \in \widetilde{\Omega} \cap \widetilde{L} \mid \widetilde{\Psi}_y(x, s) \in K\} \\ &= \{(x, s) \in \widetilde{\Omega} \cap \widetilde{L} \mid s \Psi_y(a + x/s) \in K\} \\ &= \{0\} \cup \{(x, s) \in X \times \mathbb{R} \mid a + x/s \in \Omega \cap C \text{ und } \Psi_y(a + x/s) \in K\} \\ &= \overline{\text{cone}}((M(y) - a) \times \{1\}). \end{aligned}$$

Mit der Kettenregel am Punkt $\tilde{a} := (0, 1)$ rechnen wir einfach folgendes nach:

$$\begin{aligned} \widetilde{\Psi}_y(\tilde{a}) &= \Psi(a) \in K, \\ D\widetilde{\Psi}_y(x, s)(u, r) &= D\Psi_y(a + x/s)u + r \Psi_y(a + x/s) - D\Psi_y(a + x/s)(x/s)r, \\ D\widetilde{\Psi}_y(\tilde{a})(u, r) &= D\widetilde{\Psi}_y(0, 1)(u, r) = D\Psi_y(a)u + r \Psi_y(a). \end{aligned}$$

Nun gilt

$$D\widetilde{\Psi}_b(\tilde{a})(\widetilde{L}) - K \supset D\widetilde{\Psi}_b(\tilde{a})((C - a) \times \{1\}) - K = D\Psi_b(a)(C - a) - (K - \Psi_b(a)).$$

Wegen [Lemma 4.65](#) enthält letztere Menge 0 als inneren Punkt; also ist 0 auch innerer Punkt des Kegels $D\widetilde{\Psi}_b(\tilde{a})(\widetilde{L}) - K$ und folglich gilt $D\widetilde{\Psi}_b(\tilde{a})(\widetilde{L}) - K = Z$. Mit [Satz 4.69](#) finden wir nun geeignete Umgebungen $\widetilde{U}$ um $\tilde{a}$ in $\widetilde{\Omega}$ und V um b in Y und $\widetilde{\gamma} \geq 0$, so dass

$$\widetilde{M}(y) \neq \emptyset \quad \text{und} \quad d(\tilde{x}, \widetilde{M}(y)) \leq \widetilde{\gamma} d(\widetilde{\Psi}_y(\tilde{x}), K) \quad \text{für alle } \tilde{x} \in \widetilde{L} \cap \widetilde{U} \text{ und alle } y \in V.$$

Insbesondere gilt das für Punkte der Form $\tilde{x} = (x - a, 1)$ mit $x \in C$ und $(x, 1) \in \tilde{U}$:

$$\tilde{M}(y) \neq \emptyset \quad \text{und} \quad d((x - a, 1), \tilde{M}(y)) \leq \tilde{\gamma} d(\Psi_y(x), K).$$

Die Menge $\tilde{U}$ kann so gewählt werden, dass überdies stets $\tilde{\gamma} d(\Psi_y(x), K) < 1/2$ gelten muss, wenn $(x, 1) \in \tilde{U}$ ist. Zu $\eta > 0$ und $(x, 1) \in \tilde{U}$ findet man nun ein $x_1 \in X$ und ein $s_1 \in \mathbb{R}$ mit $(x_1, s_1) \in \tilde{M}(y)$ und $\|(x - a) - x_1\| + \|1 - s_1\| \leq (\tilde{\gamma} + \eta) d(\Psi_y(x), K)$. Dann gilt aber $|1 - s_1| < \frac{1}{2}$ und $a + x_1/s_1 \in M(y)$ und man findet eine Konstante $\gamma \geq 0$ mit

$$\|x - x_1/s_1\| = \|(x - a) - (a + x_1/s_1)\| \leq \gamma d(\Psi_y(x), K).$$

Mit $U := \{x \in X \mid (x - a, 1) \in \tilde{U}\}$ gilt schließlich

$$d(x, M(y)) \leq \gamma d(\Psi_y(x), K) \quad \text{für alle } x \in C \cap U \text{ und alle } y \in V. \quad \square$$

Aufgabe 4.70

Leiten Sie aus [Satz 4.69](#) den klassischen Satz über implizite Funktionen her:

Seien X, Y und Z Banachräume, $U \subset X$ und $V \subset Y$ offen, $\Psi: U \times V \rightarrow Z$ eine im ersten Argument differenzierbare Abbildung und seien $a \in U$ und $b \in V$ zwei Punkte.

1. $\Psi(a, b) = 0$.
2. $\partial_x \Psi(a, b): X \rightarrow Z$ ist invertierbar.
3. Die Abbildungen Ψ und $\partial_x \Psi$ sind stetig bei (a, b) .

Dann gibt es eine Umgebungen U_1 von a in U und V_1 von b in V und eine bei b stetige Abbildung $m: V_1 \rightarrow U_1$ und ein $\gamma \geq 0$ mit

$$\Psi(m(y), y) = 0 \quad \text{und} \quad \|x - m(y)\| \leq \gamma \|\Psi(x, y)\| \quad \text{für alle } x \in U_1 \text{ und alle } y \in V_1.$$

4.3.4 KRZCQ und Abadie-Bedingung

Lemma 4.71 *Ist die KRZCQ (4.25) bei $a \in M$ erfüllt, so gilt*

$$T_a^{\text{lin}} M = \overline{\{u \in \text{cone}(C - a) \mid D\Psi(a)u \in \text{cone}(K - \Psi(a))\}}$$

BEWEIS. Wir definieren $Q := \{u \in \text{cone}(C - a) \mid D\Psi(a)u \in \text{cone}(K - \Psi(a))\}$. Wegen [Aufgabe 4.40](#) gilt

$$T_a^{\text{lin}} M = \{u \in \overline{\text{cone}(C - a)} \mid D\Psi(a)u \in \overline{\text{cone}(K - \Psi(a))}\}.$$

Letztere Menge ist abgeschlossen und enthält Q . Hieraus folgt $T_a^{\text{lin}} M \supset \overline{Q}$.

Sei nun $u \in T_a^{\text{lin}} M$. Dann gibt es eine Folge $u_n \in \text{cone}(C - a)$ mit $u_n \rightarrow u$. Wegen $d(D\Psi(a)u_n, K - \Psi(a)) \rightarrow 0$ können wir durch Auswählen einer Teilfolge erreichen, dass $d(D\Psi(a)u_n, K - \Psi(a)) \leq \frac{1}{2n}$ ist. Dann gibt es also ein $w_n \in K - \Psi(a)$ und ein $z_n \in Z$ mit

$$D\Psi(a)u_n = w_n + z_n, \quad \|z_n\| \leq \frac{1}{n}.$$

Wegen der KRZCQ und [Satz 4.62](#) finden wir ein $0 < R < \infty$ mit

$$\bar{B}(0; 1/R) \subset D\Psi(a) \left(\text{cone}(C - a) \cap \bar{B}(0; 1) \right) - \left(\text{cone}(K - \Psi(a)) \cap \bar{B}(0; 1) \right).$$

Folglich finden wir $v_n \in \text{cone}(C - a) \cap \bar{B}(0; \frac{R}{n})$ und $\tilde{w}_n \in \text{cone}(K - \Psi(a)) \cap \bar{B}(0; \frac{R}{n})$ mit $-z_n = D\Psi(a)v_n - \tilde{w}_n$. Wir erhalten

$$u_n + v_n \in \text{cone}(C - a), \quad D\Psi(a)(u_n + v_n) = w_n + \tilde{w}_n \in \text{cone}(K - \Psi(a)) \quad \text{und} \quad u_n + v_n \rightarrow u.$$

Also ist $u_n + v_n \in Q$ und $u \in \bar{Q}$. Dies zeigt schließlich $T_a^{\text{lin}} M \subset \bar{Q}$. $\square$

Satz 4.72 (KRZCQ $\Rightarrow$ Abadie)

Sei $a \in M$ ein Punkt, an dem die KRZCQ ([4.25](#)) erfüllt ist. Dann ist auch die Abadie-Bedingung erfüllt.

BEWEIS. Da $T_a M \subset T_a^{\text{lin}} M$ immer gilt (vgl. [Proposition 4.7](#)), haben wir lediglich $T_a^{\text{lin}} M \subset T_a M$ zu zeigen. Weil $T_a M$ abgeschlossen ist und weil die Menge

$$Q := \{u \in \text{cone}(C - a) \mid D\Psi(a)u \in \text{cone}(K - \Psi(a))\}$$

nach [Lemma 4.71](#) dicht liegt in $T_a^{\text{lin}} M$, reicht es, die Inklusion $Q \subset T_a M$ nachzuweisen.

Sei also $u \in Q$. Dann hat man $d(\Psi(a) + t D\Psi(a)u, K) = o(t)$ für $t > 0$. Seien $x_n \in C$ und $t_n > 0$ mit

$$x_n \rightarrow a, \quad t_n \rightarrow 0 \quad \text{und} \quad \frac{x_n - a}{t_n} \rightarrow u.$$

Mit $\|x_n - a\| = O(t_n)$ und $\|x_n - a - t_n u\| = o(t_n)$ erhalten wir

$$\begin{aligned} \Psi(x_n) - \Psi(a) - t_n D\Psi(a)u &= \Psi(x_n) - \Psi(a) - D\Psi(a)(x_n - a) + D\Psi(a)(x_n - a - t_n u) \\ &= o(\|x_n - a\|) + o(t_n) = o(t_n). \end{aligned}$$

Dies ergibt

$$d(\Psi(x_n), K) \leq \|\Psi(x_n) - \Psi(a) - t_n D\Psi(a)u\| + d(\Psi(a) + t_n D\Psi(a)u, K) = o(t_n).$$

Der Satz über implizite Funktionen (siehe [Satz 4.69](#)) sagt uns nun, dass $d(x_n, M) = o(t_n)$ gilt. Zu jedem x_n finden wir also ein $y_n \in M$ mit $\|y_n - x_n\| = o(t_n)$. Wegen

$$u = \lim_{n \rightarrow \infty} \frac{x_n - a}{t_n} = \lim_{n \rightarrow \infty} \left(\frac{y_n - a}{t_n} + \frac{x_n - y_n}{t_n} \right) = \lim_{n \rightarrow \infty} \frac{y_n - a}{t_n}$$

sehen wir nun endlich ein, dass $u \in T_a M$ ein Tangentialvektor ist. Folglich gilt $Q \subset T_a M$. $\square$

Hiermit ist endlich auch der Beweis von [Satz 4.57](#) abgeschlossen.

4.3.5 Übungsaufgaben

Aufgabe 4.73

Sei C eine abgeschlossene und konvexe Menge im normierten Raum X und sei $a \in C$.

Dann gilt $\overline{\text{cone}}(C - a) = T_a C$.

Aufgabe 4.74

Sei $C \subset X$ eine konvexe Menge.

1. Dann ist auch C° konvex.
2. Ist C° nichtleer, dann gilt $\overline{(C^\circ)} = \overline{C}$. Insbesondere ist das Innere von C dicht in C .
Hinweis: Man kann jeden Punkt von C mit einem inneren Punkt durch eine Linie verbinden.

Aufgabe 4.75

Sei $C \subset \mathbb{R}^n$ eine nichtleere, konvexe Menge. Dann gilt auch $(\overline{C})^\circ = C^\circ$.

Aufgabe 4.76

Selbst für nichtleere, konvexe Mengen kann in unendlichdimensionalen Räumen $(\overline{C})^\circ$ ungleich C° sein: Wir definieren den Banachraum

$$X := \{x: \mathbb{N} \rightarrow \mathbb{R} \mid \sum_{n=1}^{\infty} |x_n| < \infty\} \quad \text{mit} \quad \|x\| := \sum_{n=1}^{\infty} |x_n|$$

und die Menge $C := \{x: \mathbb{N} \rightarrow \mathbb{R} \mid \exists N \in \mathbb{N} \forall n \geq N: x_n = 0\}$. Dann ist C nichtleer, konvex und dicht in X , also $(\overline{C})^\circ = X$ aber $C^\circ = \emptyset$.

Aufgabe 4.77 (MFCQ $\Rightarrow$ KRZCQ)

Seien X , Y und Z Banachräume, $U \subset X$ eine offene Menge und $K \subset Z$ ein abgeschlossener, konvexer Kegel mit nichtleerem Inneren. Ferner seien mit $\Phi: U \rightarrow Y$ und $\Psi: U \rightarrow Z$ zwei differenzierbare Abbildungen gegeben. Wir definieren $\tilde{\Psi} := (\Phi, \Psi): U \rightarrow \tilde{Z} := Y \oplus Z$, $\tilde{C} := X$ und $\tilde{K} := \{0\} \times K$. Dazu betrachten wir den Zulässigkeitsbereich

$$M := \{x \in U \mid \Phi(x) = 0 \text{ und } \Psi(x) \in K\} = \{x \in U \mid \tilde{\Psi}(x) \in \tilde{K}\}.$$

Am Punkt $a \in M$ sei die *Mangasarian-Fromovitz-Regularitätsbedingung* (MFCQ) erfüllt (siehe [Definition 4.47](#)) Dann ist bei a auch die Kurcyusz-Robinson-Zowe-Regularitätsbedingung für $\tilde{\Psi}$ erfüllt:

$$\tilde{Z} = \text{im}(D\tilde{\Psi}(a)) - \text{cone}(\tilde{K} - \tilde{\Psi}(a)).$$

4.3.6 Optimalitätsbedingungen zweiter Ordnung

In diesem Abschnitt betrachten wir Optimalitätsbedingungen zweiter Ordnung für das Optimierungsproblem (4.18) im Fall $C = X$, also für

$$\begin{aligned} &\text{Minimiere} && F(x) \\ &\text{u. d. N.} && \Psi(x) \in K, \\ &&& x \in U. \end{aligned} \quad (4.32)$$

Außerdem nehmen wir an, dass die Abbildung F und Ψ zweimal stetig differenzierbar sind.

Satz 4.78 (Notwendige Optimalitätsbedingungen zweiter Ordnung)

Beim Optimierungsproblem (4.32) seien die Abbildung F und Ψ zweimal stetig differenzierbar. Sei $x_* \in M$ ein lokales Minimum, an dem die x_* ist die Kurcyusz-Robinson-Zowe-Regularitätsbedingung erfüllt ist. Dann gilt:

1. x_* ist ein KKT-Punkt mit Lagrange-Multiplikator μ_* .
2. $D^2F(x_*)(u, u) + \mu_* D^2\Psi(x_*)(u, u) \geq 0$ für alle $u \in T_{x_*}^{\text{lin}} M \cap \ker(DF(x_*))$.

BEWEIS. Sei $u \in T_{x_*}^{\text{lin}} M \cap \ker(DF(x_*))$. Aus der ersten KKT-Bedingung $0 = DF(x_*)u + \mu_* D\Psi(x_*)u = 0$ und $DF(x_*)u = 0$ folgt $\mu_* D\Psi(x_*)u = 0$. Die complementary-slackness-Bedingung liefert $\mu_* \Psi(x_*) = 0$.

Schritt 1: Unser erstes Ziel ist es, eine Folge $x_k \in U$ und $t_k > 0$ zu finden mit

$$x_k \rightarrow x_*, \quad t_k \rightarrow 0, \quad u = \lim_{k \rightarrow \infty} \frac{x_k - x_*}{t_k}, \quad \Psi(x_k) \in K \quad \text{und} \quad \langle \mu_*, \Psi(x_k) \rangle = 0. \quad (4.33)$$

Bis auf die letzte Bedingung wird uns dies durch Satz 4.72 garantiert. Wir konstruieren daher ein Hilfsproblem mit $\langle \mu_*, \Psi(x_k) \rangle = 0$ als zusätzlicher Nebenbedingung:

Als abgeschlossener Untervektorraum im Banachraum Z ist $Z_0 := \ker(\mu_*)$ auch selbst ein Banachraum. Der Kegel $K_0 := Z_0 \cap K$ ist abgeschlossen und konvex. Ist $\mu_* \neq 0$, so finden wir $z_0 \in Z$ mit $\langle \mu_*, z_0 \rangle \neq 0$. Dann definieren wir den Projektor $P: Z \rightarrow Z_0$ durch $Pw := w - z_0 \frac{\langle \mu_*, w \rangle}{\langle \mu_*, z_0 \rangle}$. Ist $\mu_* = 0$, so gilt $Z_0 = Z$ und wir verwenden einfach $P := \text{id}_Z$.⁶ Wir definieren

$$\Psi_0 := P \circ \Psi \quad \text{und} \quad M_0 := \{x \in U \mid \Psi_0(x) \in K_0\} = \{x \in M \mid \langle \mu_*, \Psi(x) \rangle = 0\}.$$

Wegen $\Psi(x_*) \in K$ und $\langle \mu_*, \Psi(x_*) \rangle = 0$ haben wir auch $\Psi_0(x_*) \in K_0$ und $x_* \in M_0$. Aus der KRZCQ für Ψ , M und K folgt die KRZCQ für Ψ_0 , M_0 und K_0 :

$$Z_0 = \text{im}(D\Psi_0(x_*)) - \text{cone}(K_0 - \Psi_0(x_*)).$$

Nach Satz 4.72 gilt also

$$T_{x_*} M_0 = T_{x_*}^{\text{lin}} M_0 = T_{x_*}^{\text{lin}} M \cap \ker(\mu_* D\Psi(x_*)) \supset T_{x_*}^{\text{lin}} M \cap \ker(DF(x_*)) \ni u.$$

⁶Im Grunde müssten wir im Fall von $\mu_* = 0$ diesen ganzen Zirkus ja gar nicht veranstalten.

Das beweist aber schon die Existenz von $(x_k)_{k \in \mathbb{N}}$ und $(t_k)_{k \in \mathbb{N}}$ wie in (4.33)

Schritt 2: Nach Konstruktion haben wir nun

$$F(x_k) + \mu_* \Psi(x_k) = F(x_*).$$

Der Satz von Taylor zusammen mit $x_k - x_* = t_k u + o(t_k)$ ergibt die Identitäten

$$\begin{aligned} F(x_k) &= F(x_*) + t_k \overline{DF(x_*)} u + \frac{1}{2} t_k^2 D^2 F(x_*) (u, u) + o(t_k^2), \\ \mu_* \Psi(x_k) &= \mu_* \overline{\Psi(x_*)} + t_k \mu_* \overline{D\Psi(x_*)} u + \frac{1}{2} t_k^2 \mu_* D^2 \Psi(x_*) (u, u) + o(t_k^2). \end{aligned}$$

Zusammengesetzt erhalten wir also

$$F(x_k) = F(x_k) + \mu_* \Psi(x_k) = F(x_*) + \frac{1}{2} t_k^2 \left(D^2 F(x_*) (u, u) + \mu_* D^2 \Psi(x_*) (u, u) \right) + o(t_k^2).$$

Wäre aber $D^2 F(x_*) (u, u) + \mu_* D^2 \Psi(x_*) (u, u) < 0$, so ergäbe sich hieraus ein Widerspruch dazu, dass x_* ein lokales Minimum ist. $\square$

Satz 4.79 (Hinreichende Optimalitätsbedingungen zweiter Ordnung [6])

Beim Optimierungsproblem (4.32) seien die Abbildung F und Ψ zweimal stetig differenzierbar. Der Punkt $x_* \in M$ gelte:

1. Die KRZCQ ist bei x_* erfüllt.
2. x_* ist ein KKT-Punkt mit Lagrange-Multiplikator μ_* .
3. Es gibt $\beta > 0$ und $\gamma > 0$ sodass für alle $u \in \{u \in T_{x_*}^{\text{lin}} M \mid DF(x_*) u \leq \beta \|u\|\}$ gilt:

$$D^2 F(x_*) (u, u) + \mu_* D^2 \Psi(x_*) (u, u) \geq \gamma \|u\|^2.$$

Dann ist x_* ein striktes lokales Minimum.

BEWEIS. Der Satz über implizite Funktionen Satz 4.69 angewendet auf die Abbildung $y \mapsto D\Psi(x_*) (y - x_*)$ und den Kegel $\overline{\text{cone}}(K - \Psi(x_*))$ zeigt für $x \in M$:

$$\begin{aligned} d(x - x_*, T_{x_*}^{\text{lin}} M) &\leq \gamma_0 d\left(D\Psi(x_*) (x - x_*), \overline{\text{cone}}(K - \Psi(x_*))\right) \\ &\leq \gamma_0 d\left(D\Psi(x_*) (x - x_*), K - \Psi(x_*)\right) \\ &= \gamma_0 d(\Psi(x_*) + D\Psi(x_*) (x - x_*), K) \\ &\leq \gamma_0 \|\Psi(x_*) + D\Psi(x_*) (x - x_*) - \Psi(x)\| + \gamma_0 \overline{d(\Psi(x), K)} = o(\|x - x_*\|). \end{aligned}$$

Es gibt daher eine Abbildung $u: M \rightarrow T_{x_*}^{\text{lin}} M$ mit $\|(x - x_*) - u(x)\| = o(\|x - x_*\|)$. Mit $w: M \rightarrow X$ bezeichnen wir die Abbildung $w(x) := (x - x_*) - u(x)$. Für jedes $\eta > 0$ gibt es ein $\varepsilon = \varepsilon(\eta) > 0$

so dass $\|w(x)\| \leq \eta \|u(x)\|$ für alle $x \in \bar{B}(x_*; \varepsilon)$ gilt. Wir werden η und ε im Folgenden sukzessive verkleinern.

Fall 1: Wir zeigen: Für hinreichend kleines $\varepsilon > 0$ gilt

$$F(x) \geq F(x_*) + \frac{\beta}{2} \|x - x_*\| \quad \text{für alle } x \in M \cap B(x_*; \varepsilon) \text{ mit } DF(x_*) u(x) > \beta \|u(x)\|.$$

Sei $\varrho(r) := \sup_{x \in B(x_*, r)} \|DF(x_*) - DF(x)\|$.

$$\begin{aligned} F(x) &\geq F(x_*) + DF(x_*) (x - x_*) - \varrho(r) \|x - x_*\| \\ &\geq F(x_*) + DF(x_*) u(x) + DF(x_*) w(x) - \varrho(\varepsilon) \|x - x_*\| \\ &\geq F(x_*) + (\beta - \|DF(x_*)\| \eta) \|u(x)\| - \varrho(\varepsilon) \|x - x_*\| \end{aligned}$$

Für hinreichend kleines $\eta > 0$ haben wir $\|u(x)\| \leq \frac{3}{4} \|x - x_*\|$ und $\|DF(x_*)\| \eta \leq \frac{1}{6} \beta$. Überdies können wir $\varepsilon > 0$ so wählen, sodass $\varrho(\varepsilon) \leq \frac{1}{8} \beta$ gilt. Wir erhalten

$$F(x) \geq F(x_*) + (\frac{3}{4} \beta - \frac{3}{4} \cdot \frac{1}{6} \beta) \|x - x_*\| - \frac{1}{8} \beta \|x - x_*\| = F(x_*) + \frac{\beta}{2} \|x - x_*\|.$$

Fall 2: Wir zeigen: Für hinreichend kleines $\varepsilon > 0$ gilt

$$F(x) \geq F(x_*) + \frac{\gamma}{4} \|x - x_*\|^2 \quad \text{für alle } x \in M \cap B(x_*; \varepsilon) \text{ mit } DF(x_*) u(x) \leq \beta \|u(x)\|.$$

Wir verwenden folgende Abkürzungen:

$$L_\beta := \{u \in T_{x_*}^{\text{lin}} M \mid DF(x_*) u \leq \beta \|u\|\} \quad \text{und} \quad b := D^2 F(x_*) + \mu_* D^2 \Psi(x_*).$$

Sei $\eta > 0$. Für $u \in L_\beta$ und $w \in X$ mit $\|w\| \leq \eta \|u\|$ gilt dann einerseits

$$\begin{aligned} b(u + w, u + w) &= b(u, u) + 2b(u, w) + b(w, w) \\ &\geq \gamma \|u\|^2 - 2\|b\| \eta \|u\|^2 - \|b\| \eta^2 \|u\|^2 = (\gamma - 2\|b\| \eta - \|b\| \eta^2) \|u\|^2. \end{aligned}$$

Andererseits gilt auch

$$\|u + w\|^2 \leq \|u\|^2 + 2\|u\| \|w\| + \|w\|^2 \leq (1 + 2\eta + \eta^2) \|u\|^2.$$

Mit $\delta(\eta) := \frac{\gamma - 2\|b\| \eta - \|b\| \eta^2}{(1 + 2\eta + \eta^2)}$ erhalten wir

$$b(u + w, u + w) \geq \delta(\eta) \|u + w\|^2 \quad \text{für alle } u \in L_\beta \text{ und } w \in X \text{ mit } \|w\| \leq \eta \|u\|. \quad (4.34)$$

Wir definieren $H := F + \mu_* \Psi$ und $\omega(r) := \sup_{x \in \bar{B}(x_*, r)} \|D^2 H(x_*) - D^2 H(x)\|$. Für $x \in M \cap B(x_*; \varepsilon)$ schreiben wir $x - x_* = u(x) + w(x)$ und beobachten $\|w(x)\| \leq \eta \|u(x)\|$. Wir rechnen mit einer Taylorentwicklung für H und mit $H(x_*) = F(x_*)$:

$$\begin{aligned} F(x) &\geq F(x) + \mu_* \Psi(x) = H(x) \\ &\geq H(x_*) + \overline{DH(x_*)(x - x_*)} + \frac{1}{2} D^2 H(x_*)(x - x_*, x - x_*) - \frac{1}{2} \omega(\varepsilon) \|x - x_*\|^2 \\ &\geq F(x_*) + \frac{1}{2} b(x - x_*, x - x_*) - \frac{1}{2} \omega(\varepsilon) \|x - x_*\|^2. \end{aligned}$$

Nun wenden wir (4.34) an auf $x - x_* = u(x) + w(x)$ und erhalten

$$F(x) \geq F(x_*) + \frac{1}{2}(\delta(\eta) - \omega(\varepsilon)) \|x - x_*\|^2.$$

Wegen $\delta(\eta) \xrightarrow{\eta \rightarrow 0} \gamma > 0$ kann man zunächst η so klein wählen, dass $\delta(\eta) \geq \frac{3}{4}\gamma$ gilt. Anschließend verkleinern wir $\varepsilon > 0$ noch so weit, dass $\|u(x)\| \leq \eta \|w(x)\|$ und $\omega(\varepsilon) \leq \frac{1}{4}\gamma$ gelten. Fall 1 und Fall 2 zusammen ergeben die Behauptung. $\square$

Folgerung 4.80 (Endlichdimensionaler Fall)

Beim Optimierungsproblem (4.32) seien die Abbildung F und Ψ zweimal stetig differenzierbar und X sei endlichdimensional. Der Punkt $x_* \in M$ gelte:

1. Die KRZCQ ist bei x_* erfüllt.
2. x_* ist ein KKT-Punkt mit Lagrange-Multiplikator μ_* .
3. $D^2F(x_*)(u, u) + \mu_* D^2\Psi(x_*)(u, u) > 0$ für alle $u \in T_{x_*}^{\text{lin}} M \cap \ker(DF(x_*))$ mit $u \neq 0$.

Dann ist x_* ein striktes lokales Minimum.

BEWEIS. Angenommen, solche Konstanten $\beta > 0$ und $\gamma > 0$ wie in Bedingung 3 von Satz 4.79 gibt es nicht. Dann findet man eine Folge $(u_k)_{k \in \mathbb{N}}$ in $T_{x_*}^{\text{lin}} M$ mit

$$\|u_n\| = 1, \quad DF(x_*) u_n \leq \frac{1}{n} \quad \text{und} \quad D^2F(x_*)(u_n, u_n) + \mu_* D^2\Psi(x_*)(u_n, u_n) \leq \frac{1}{n}.$$

Weil der Einheitsball in X (folgen-)kompakt ist, muss die Folge $(u_k)_{k \in \mathbb{N}}$ eine konvergente Teilfolge besitzen. Sei u deren Grenzwert. Weil $T_{x_*}^{\text{lin}} M$ abgeschlossen ist gilt:

$$u \in T_{x_*}^{\text{lin}} M, \quad \|u\| = 1, \quad DF(x_*) u = 0 \quad \text{und} \quad D^2F(x_*)(u, u) + \mu_* D^2\Psi(x_*)(u, u) = 0.$$

Das ist aber ein Widerspruch zu den Annahmen des Satzes. $\square$

Beispiel 4.81 ([6]) Die Bedingungen 1–3 aus Folgerung 4.80 sind im unendlichdimensionalen nicht hinreichend. Um das einzusehen betrachten wir den Hilbertraum

$$X = Z := \left\{ (x_n)_{n \in \mathbb{N}} \mid x_n \in \mathbb{R}, \sum_{n \in \mathbb{N}} x_n^2 < \infty \right\}$$

mit dem Skalarprodukt $g(x, y) := \sum_{n \in \mathbb{N}} x_n y_n$. Dazu betrachten wir den abgeschlossenen, konvexen Kegel $K := \{z \in Z \mid \forall n \in \mathbb{N}: z_n \leq 0\}$, das lineare Funktional $\xi \in \text{pol}(K)$, $\langle \xi, x \rangle := \sum_{n \in \mathbb{N}} \frac{1}{n} x_n$, die Zielfunktion $F: X \rightarrow \mathbb{R}$, $F(x) = -\langle \xi, x \rangle - g(x, x)$ und die Abbildung $\Psi: X \rightarrow \mathbb{R}$, $\Psi(x) = x$. Am Punkt $x_* = 0 \in X$ haben wir $D\Psi(x_*) = \text{id}_X$; die KRZCQ ist also an diesem Punkt erfüllt. Weiters ist $x_* = 0 \in X$ auch ein KKT-Punkt mit Lagrange-Multiplikator $\mu_* = \xi$:

$$DF(x_*) + \mu_* D\Psi(x_*) = -\xi + \xi \text{id}_X = 0, \quad \Psi(x_*) \in K, \quad \mu \in \text{pol}(K) \quad \text{and} \quad \mu_* \Psi(x_*) = 0.$$

Wir haben $T_{x_*}^{\text{lin}} M = K$ und $\ker(DF(x_*)) = \ker(\xi)$. Ist $u \in \ker(\xi) \setminus \{0\}$, so muss es ein $n \in \mathbb{N}$ geben mit $u_n < 0$. Folglich ist $T_{x_*}^{\text{lin}} M \cap \ker(DF(x_*)) = \{0\}$ und

$$D^2F(x_*)(u, u) + \mu_* D^2\Psi(x_*)(u, u) \geq \|u\|^2 \quad \text{für alle } u \in T_{x_*}^{\text{lin}} M \cap \ker(DF(x_*)).$$

gilt trivialerweise. Trotzdem ist $x_* = 0$ *kein* lokales Minimum. Dies sieht man ein, indem man die durch $x_n^{(k)} := -\frac{2}{k} \delta_{kn}$ definiert Folge $(x^{(k)})_{k \in \mathbb{N}}$ in X betrachtet. Wegen $\|x^{(k)}\| = \frac{2}{k}$ konvergiert diese gegen $0 \in X$. Allerdings gilt

$$F(x^{(k)}) = \frac{2}{k^2} - \frac{4}{k^2} = -\frac{2}{k^2} < 0 = F(x_*).$$

Aufgabe 4.82

Zu $r > 0$ betrachten wir das Optimierungsproblem

$$\begin{aligned} &\text{Minimiere} && F(x, y) := -(x+1)^2 - y^2 \\ &\text{u. d. N.} && (x-r)^2 + y^2 = r^2, \\ &&& (x, y) \in \mathbb{R}^2. \end{aligned}$$

Sei $a := (0, 0)$. Bestimmen Sie zu jeder der folgenden Aussagen alle Werte von $r > 0$ für die sie erfüllt sind:

1. Die KKT-Bedingungen sind bei a erfüllt.
2. Die notwendigen Bedingungen zweiter Ordnung sind bei a erfüllt.
3. Die hinreichenden Bedingungen zweiter Ordnung sind bei a erfüllt.
4. Es gibt ein $c > 0$ mit $DF(a)(u, u) \geq c \|u\|^2$ für alle $u \in T_a M$.

4.4 Newton-Verfahren bei Gleichungsnebenbedingungen

Seien X und Y Banachräume, sei $F: X \rightarrow \mathbb{R}$ eine zweimal stetig differenzierbare Funktion und $\Phi: X \rightarrow Y$ eine zweimal stetig differenzierbare Abbildung. Wir betrachten im Wesentlichen das Optimierungsproblem aus (4.16) auf dem Ganzraum X :

$$\begin{aligned} &\text{Minimiere} && F(x) \\ &\text{u. d. N.} && \Phi(x) = 0. \end{aligned} \tag{4.35}$$

Wie wir in [Folgerung 4.46](#) gesehen haben, vereinfachen sich die Karush-Kuhn-Tucker-Bedingungen (4.3)–(4.6) zu

$$DF(x_*) + \lambda_* D\Phi(x_*) = 0, \tag{4.36}$$

$$\Phi(x_*) = 0. \tag{4.37}$$

Die Lösungen dieses Gleichungssystems stimmen also mit den Nullstellen der folgenden Abbildung überein:

$$\Theta: U \times Y' \rightarrow X' \times Y, \quad \Theta(x, \lambda) := \begin{pmatrix} DF(x) + \lambda D\Phi(x) \\ \Phi(x) \end{pmatrix}.$$

Sind X und Y endlichdimensional, so gilt wegen des Dimensionssatzes $X \times Y' \cong X' \times Y$. Falls X und Y beides Hilberträume sind, folgt $X \times Y' \cong X' \times Y$ aus dem Darstellungssatz von Fréchet-Riesz (siehe [Satz 2.35](#)). In diesen Fällen sollte es zumindest *prinzipiell* möglich sein, Lösungen des KKT-Systems durch das Newton-Verfahren für Θ aufzufinden.

Die Ableitung von Θ lautet

$$D\Theta(x, \lambda)(u, \eta) = \begin{pmatrix} DDF(x)u + \lambda D\Phi(x)u + \eta D\Phi(x) \\ D\Phi(x)u \end{pmatrix} \quad \text{für } u \in X \text{ und } \eta \in Y'.$$

In Blockmatrixform sieht das wie folgt aus:

$$D\Theta(x, \lambda) = \begin{pmatrix} DDF(x) + \lambda D\Phi(x) & D\Phi(x)' \\ D\Phi(x) & 0 \end{pmatrix}.$$

Dies ist eine Blockmatrix von der Form

$$\begin{pmatrix} A & B \\ C & 0 \end{pmatrix}.$$

Solche Matrizen werden *Sattelpunktmatrizen* genannt. Dem Leser sei der Überblicksartikel [1] wärmstens ans Herz gelegt. Dort findet etliche Anwendungsbeispiele, Lösungstheorie und numerische Methoden zum Lösen von linearen Sattelpunktsystemen. Wir konzentrieren uns hier nur auf die Lösbarkeit des KKT-Systems in folgendem besonderen Fall:

Satz 4.83 (Newtonverfahren)

Im Optimierungsproblem (4.35) seien F und Φ jeweils zweimal stetig differenzierbar mit lokal Lipschitz-stetigen zweiten Ableitungen. Seien (X, g) und (Y, h) endlichdimensionale Hilberträume und der Punkt $x_* \in M$ erfülle:

1. $D\Phi(x_*)$ surjektiv.
2. x_* ist ein KKT-Punkt mit Lagrange-Multiplikator λ_* .
3. Es gebe ein $c > 0$ mit $D^2F(x_*)(u, u) + \lambda_* D^2\Phi(x_*)(u, u) \geq c \|u\|^2$ für alle $u \in T_{x_*}^{\text{lin}} M \cap \ker(DF(x_*))$.

Dann ist $D\Theta(x_*, \lambda_*)$ stetig invertierbar. Für Startwerte (x_0, λ_0) hinreichend nahe bei (x_*, λ_*) konvergiert das Newton-Verfahren

$$(x_{k+1}, \lambda_{k+1}) := (x_k, \lambda_k) - D\Theta(x_k, \lambda_k)^{-1} \Theta(x_k, \lambda_k)$$

also quadratisch gegen (x_*, λ_*) .

BEWEIS. Wir setzen $a := D^2F(x_*) + \lambda_* D^2\Phi(x_*)$, $A := J_g^{-1} J_a$ und $B := D\Phi(x_*, \lambda_*)$. Damit hat $D\Theta(x_*, \lambda_*)$ die Blockmatrixform

$$D\Theta(x_*, \lambda_*) = \begin{pmatrix} J_a & B' \\ B & 0 \end{pmatrix}: X \oplus Y' \rightarrow X' \oplus Y.$$

Wir definieren die Abbildung $\mathcal{A}: X \oplus Y \rightarrow X \oplus Y$ durch

$$\mathcal{A} = \begin{pmatrix} J_g^{-1} & 0 \\ 0 & \text{id}_Y \end{pmatrix} \begin{pmatrix} J_a & A' \\ A & 0 \end{pmatrix} \begin{pmatrix} \text{id}_X & 0 \\ 0 & J_h \end{pmatrix} = \begin{pmatrix} A & B^* \\ B & 0 \end{pmatrix}.$$

Es ist also $D\Theta(x_*, \lambda_*)$ genau dann stetig invertierbar, wenn $\mathcal{A}$ es ist. Wegen der hinreichenden Bedingung zweiter Ordnung gibt es ein $c > 0$ so dass

$$a(u, u) \geq c \|u\|^2 \quad \text{für alle } u \in T_{x_*}^{\text{lin}} M \cap \ker(DF(x_*))$$

gelte. Weil $D\Phi(x_*)$ surjektiv ist, ist also LICQ und damit die Abadie-Bedingung erfüllt (siehe Satz 4.45). Daher gilt $T_{x_*} M = T_{x_*}^{\text{lin}} M = \ker(B)$. Weil $T_{x_*} M$ also ein Vektorraum ist, folgt aus der ersten KKT-Bedingung, dass $T_{x_*} M \subset \ker(DF(x_*))$ gelten muss. Wir erhalten also

$$g(A u, u) = a(u, u) \geq c \|u\|^2 \quad \text{für alle } u \in \ker(B).$$

Aus Lemma 4.85 folgt nun, dass $\mathcal{A}$ und damit auch $D\Theta(x_*, \lambda_*)$ stetig invertierbar ist. Damit ist aber $D\Theta(x, \lambda)$ auch auf einer geeigneten Umgebung von (x_*, λ_*) stetig invertierbar (vgl. Aufgabe 3.33) und der Satz von Newton-Kantorovich Satz 3.27 ist anwendbar; dieser zeigt schließlich die quadratische Konvergenz. $\square$

Bemerkung 4.84 Die Bedingung 2 des vorangehenden Satzes sind insbesondere dann erfüllt, wenn die hinreichenden Optimalitätsbedingungen zweiter Ordnung aus [Satz 4.79](#) erfüllt sind.

Lemma 4.85 ([\[4\]](#),[\[1\]](#)) Seien (X, g) und (Y, h) endlichdimensionale Hilberträume. Sei $A \in L(X; X)$ selbstadjungiert (d. h. $A^* = A$) und $B \in L(X; Y)$ surjektiv. Es gebe ein $c > 0$ sodass

$$g(Au, u) \geq c \|u\|^2 \quad \text{für alle } u \in \ker(B). \quad (4.38)$$

Dann ist die Sattelpunktmatrix $\mathcal{A} := \begin{pmatrix} A & B^* \\ B & 0 \end{pmatrix}$ stetig invertierbar und es gilt $\mathcal{A}^{-1} = \mathcal{B}$ mit

$$\mathcal{B} := \begin{pmatrix} L & (\text{id}_X - LA) B^\dagger \\ (B^\dagger)^* (\text{id}_X - AL) & -(B^\dagger)^* (A - ALA) B^\dagger \end{pmatrix}. \quad (4.39)$$

Hierbei ist $C: \ker(B) \hookrightarrow X$ die kanonische Einbettung, $L := C(C^*AC)^{-1}C^*$ und $B^\dagger = B^*(BB^*)^{-1}$ die Moore-Penrose-Pseudoinverse von B .

BEWEIS. Auf $\ker(B)$ verwenden wir die Einschränkung g_0 des Skalarproduktes g . Die Abschätzung (4.38) besagt dann gerade, dass $C^*AC: \ker(B) \rightarrow \ker(B)$ eine stetige und stetig invertierbare, lineare Abbildung ist.⁷ Wie wir schon in [Aufgabe 3.57](#) nachgerechnet haben, ist $B^\dagger B$ der Orthoprojektor auf $\text{im}(B^*) = \ker(B)^\perp$. Gleichzeitig ist CC^* der Orthoprojektor auf $\ker(B)$. Daher gilt $B^\dagger B = \text{id}_X - CC^*$. Außerdem gilt $BC = 0$ und daher auch $C^*B^* = 0$ und $LB^* = 0$. Wir rechnen nun

$$\mathcal{B}\mathcal{A} = \begin{pmatrix} LA + (\text{id}_X - LA)B^\dagger B & LB^* \\ (B^\dagger)^*(\text{id}_X - AL)A - (B^\dagger)^*(A - ALA)B^\dagger B & (B^\dagger)^*(\text{id}_X - AL)B^* \end{pmatrix}.$$

Für den linken oberen Block rechnen wir

$$\begin{aligned} LA + (\text{id}_X - LA)B^\dagger B &= LA + (\text{id}_X - LA)(\text{id}_X - CC^*) \\ &= LA + \text{id}_X - LA - CC^* + LACC^* \\ &= \text{id}_X - CC^* + C \underbrace{(C^*AC)^{-1}}_{=0} C^* ACC^* = \text{id}_X. \end{aligned}$$

Für den linken unteren Block gilt

$$\begin{aligned} (B^\dagger)^*(\text{id}_X - AL)A - (B^\dagger)^*(A - ALA)B^\dagger B &= (B^\dagger)^*(A - ALA - (A - ALA)(\text{id}_X - CC^*)) \\ &= (B^\dagger)^*(A - ALA - A + ALA + ACC^* - ALACC^*) \\ &= (B^\dagger)^*(ACC^* - AC(C^*AC)^{-1}C^*ACC^*) = 0. \end{aligned}$$

⁷Genau gilt $J_{g_0} C^*AC = J_a$ für die stetige Bilinearform $a: \ker(B) \times \ker(B) \rightarrow \mathbb{R}$, $a(u, u) := g(ACu, Cu) = g_0(C^*ACu, u)$. Wegen $a(u, v) \leq \|A\| \|u\|_{g_0} \|v\|_{g_0}$ und $a(u, v) \geq c g_0(u, u)$ für alle $u, v \in \ker(B)$ und wegen [Satz 2.35](#) ist J_a stetig invertierbar und somit auch C^*AC .

Schließlich gilt für den rechten unteren Block

$$(B^\dagger)^*(\text{id}_X - A L) B^* = (B^\dagger)^* B^* - A L B^* = (B B^\dagger)^* = \text{id}_Y^* = \text{id}_Y.$$

Wir haben also $\mathcal{B} \mathcal{A} = \text{id}_{X \oplus Y}$. Weil sowohl $\mathcal{A}$ als auch $\mathcal{B}$ selbstadjungiert sind, gilt auch $\mathcal{A} \mathcal{B} = (\mathcal{B}^* \mathcal{A}^*)^* = (\mathcal{B} \mathcal{A})^* = \text{id}_{X \oplus Y}^* = \text{id}_{X \oplus Y}$. $\square$

Bemerkung 4.86 Satz 4.83 und Lemma 4.85 gelten wörtlich so auch für unendlichdimensionale Hilberträume. Wesentlich dabei ist lediglich, dass $B B^*$ invertierbar und damit $B^\dagger$ existent ist. Das kann man z. B. nachweisen, indem man Aufgabe 4.89, Aufgabe 4.90 und Aufgabe 4.91 bearbeitet.

Bemerkung 4.87 Ein wichtiger Spezialfall sind quadratische Optimierungsprobleme (ohne Ungleichungsbedingungen) *quadratisches Optimierungsproblem*, das sind solche, bei denen die Zielfunktion quadratisch und Φ affin-linear ist. Die KKT-Bedingungen sind dann ein lineares Gleichungssystem und das Newton-Verfahren konvergiert bereits nach einem Schritt.

Das Newton-Verfahren für allgemeinere F und Φ ist ein Spezialfall des sogenannten *sequential quadratic programming* (SQP).

Bemerkung 4.88 Die Formeln (4.39) und (4.40) sind üblicherweise sehr schlecht dazu geeignet, die auftretenden Gleichungssysteme auch numerisch zu lösen. Je nach vorliegendem Matrixtyp sollte man andere Löser verwenden, zum Beispiel:

- direkte Löser für dicht besetzte Matrizen (nur geeignet für Systeme mit bis zu einigen tausend Freiheitsgraden)
- direkte Löser für dünn besetzte Matrizen (geeignet für Systeme mit bis zu einigen Millionen Freiheitsgraden bei geringer Bandbreite oder bei besonders geeigneter Besetzungsstruktur)
- iterative Löser für dünn besetzte Matrizen oder für matrixfreie Operationen (für große Systeme sind geeignete Präkonditionierer zwingend erforderlich!)

In der Regel wird man verschiedene Löser ausprobieren müssen, bevor man den geeigneten gefunden hat. Für Details verweisen wir abermals auf [1].

4.4.1 Übungsaufgaben

Aufgabe 4.89

Seien X und Y normierte Räume und sei $B \in L(X; Y)$. Dann gilt:

1. $\ker(B') = \text{pol}(\text{im}(B))$ und $\ker(B) = \text{prepol}(\text{im}(B'))$.
2. $\text{prepol}(\ker(B')) = \overline{\text{im}(B)}$.

3. $\text{pol}(\ker(B)) \supset \overline{\text{im}(B')}$.
4. Ist X reflexiv, so gilt sogar $\text{pol}(\ker(B)) = \overline{\text{im}(B')}$.

Hinweis: Nutzen Sie den Bipolarensatz [Satz 4.23](#).

Aufgabe 4.90

Seien (X, g) und (Y, h) Hilberträume und sei $B \in L(X; Y)$. Dann gilt:

1. $\ker(B^*) = \text{im}(B)^\perp$ und $\ker(B) = \text{im}(B^*)^\perp$.

Hinweis: Verwenden Sie $B^* = J_g^{-1} B' J_h$ und [Aufgabe 4.89](#).

2. Ist $\text{im}(B)$ abgeschlossen, so gilt auch $\ker(B^*)^\perp = \text{im}(B)$ und $\ker(B)^\perp = \text{im}(B^*)$.

Hinweis: Verwenden Sie ohne Beweis, dass $\text{im}(B^*)$ abgeschlossen ist genau dann, wenn $\text{im}(B)$ abgeschlossen ist. Das ist Gegenstand des *Satzes vom abgeschlossenen Bild*. Einen Beweis dafür findet man z. B. in [\[13\]](#), S. 159–161.

Aufgabe 4.91

Seien (X, g) und (Y, h) Hilberträume und sei $B \in L(X; Y)$ surjektiv. Dann ist BB^* stetig invertierbar.

Hinweis: Folgern Sie aus [Aufgabe 4.90](#), dass BB^* bijektiv sein muss und wenden Sie [Folgerung 4.64](#) an.

Aufgabe 4.92

Seien X_1, X_2 und Y_1, Y_2 Banachräume und $A \in L(X_1; Y_1)$, $B \in L(X_2; Y_1)$, $C \in L(X_1; Y_2)$ und $D \in L(X_2; Y_2)$ stetige, lineare Abbildungen und sei A stetig invertierbar. Die Blockmatrix $\mathcal{A} := \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ ist genau dann stetig invertierbar, wenn das *Schur-Komplement* $S := D - C A^{-1} B$ stetig invertierbar ist und es gilt

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1} B S^{-1} C A^{-1} & -A^{-1} B S^{-1} \\ -S^{-1} C A^{-1} & S^{-1} \end{pmatrix}. \quad (4.40)$$

Aufgabe 4.93

Im Setting von [Aufgabe 3.75](#) suchen wir nun die Ruhelage einer *starren Kette*, bei dem jedes Kettenglied die feste Länge $\|x_{i+1} - x_i\| = \ell$ hat. Lösen Sie dieses Problem numerisch mit dem Newtonverfahren.

4.5 Newton-Verfahren bei Ungleichungsnebenbedingungen

Wir beschäftigen uns weiterhin mit dem Optimierungsproblem (4.10) und wollen versuchen, die KKT-Gleichungen (4.11)–(4.15) algorithmisch durch iterative Verfahren zu lösen.

FiXme Note:
Zitate
ergänzen

4.5.1 Aktive Mengen

Wir erinnern noch einmal an die Definition der aktiven Menge Definition 4.52.

Für zulässige Punkte gilt stets $\mathcal{A}(x) \cup \mathcal{I}(x) = \{1, \dots, m\}$. Ist (x_*, λ_*, μ_*) ein KKT-Tripel, so gilt wegen (4.15)

$$j \in \mathcal{I}(x_*) \Rightarrow \mu_{*j} = 0.$$

Sei $\mathcal{J} = (j_1, \dots, j_k) \subset (1, \dots, m)$ eine geordnete Indexmenge, $z \in Z$ und $\mu \in Z'$. Wir schreiben

$$Z_{\mathcal{J}} := \mathbb{R}^k, \quad z_{\mathcal{J}} := (z_{j_1}, \dots, z_{j_k})^T \quad \text{und} \quad \mu_{\mathcal{J}} := (\mu_{j_1}, \dots, \mu_{j_k}).$$

Für eine geordnete Menge $\mathcal{S} \subset (1, \dots, m)$ und ihr geordnetes Komplement $\mathcal{R}$ definieren wir die Abbildung

$$\Theta_{\mathcal{S}}: U \times Y' \times Z'_{\mathcal{S}} \times Z'_{\mathcal{R}} \rightarrow X' \times Y \times Z_{\mathcal{S}} \times Z_{\mathcal{R}}, \quad (4.41)$$

durch

$$\Theta_{\mathcal{S}}(x, \lambda, \mu_{\mathcal{S}}, \mu_{\mathcal{R}}) := \begin{pmatrix} DF(x) + \lambda D\Phi(x) + \mu_{\mathcal{S}} D\Psi_{\mathcal{S}}(x) + \mu_{\mathcal{R}} D\Psi_{\mathcal{R}}(x) \\ \Phi(x) \\ \Psi_{\mathcal{S}}(x) \\ -\mu_{\mathcal{R}}(x) \end{pmatrix}. \quad (4.42)$$

Proposition 4.94 Ein Punkt $x_* \in M$ ist genau dann ein KKT-Punkt von (4.10), wenn es eine geordnete Teilmenge $\mathcal{A}_*$ von $\{1, \dots, m\}$ und Lagrange-Multiplikatoren $\lambda_* \in Y'$ und $(\mu_*)_{\mathcal{A}_*} \in (Z_{\mathcal{A}_*})'$ gibt, so dass

$$\Theta_{\mathcal{A}_*}(x_*, \lambda_*, \mu_*) = 0, \quad \mu_{* \mathcal{A}_*} \geq 0 \quad \text{und} \quad \Psi_{\mathcal{I}_*}(x_*) < 0 \quad (4.43)$$

erfüllt sind. Hierbei sei $\mathcal{I}_* := (1, \dots, m) \setminus \mathcal{A}_*$ das geordnete Komplement von $\mathcal{A}_*$.

BEWEIS. “ $\Rightarrow$ ”: Sei (x_*, λ_*, μ_*) ein KKT-Tripel. Mit $\mathcal{A}_* := \mathcal{A}(x_*)$ und $\mathcal{I}_* := \mathcal{I}(x_*)$ extrahieren wir aus den KKT-Bedingungen die folgende notwendige Bedingung:

$$\begin{aligned} DF(x_*) + \lambda_* D\Phi(x_*) + \mu_* D\Psi(x_*) &= 0, \\ \Phi(x_*) &= 0, \\ \Psi_{\mathcal{A}_*}(x_*) &= 0, \\ \Psi_{\mathcal{I}_*}(x_*) &< 0, \\ \mu_{\mathcal{A}_*} &\geq 0, \\ \mu_{\mathcal{I}_*} &= 0. \end{aligned}$$

Das ist aber gleichbedeutend mit (4.43); man muss lediglich

$$\mu_* D\Psi(x_*) = \mu_{*\mathcal{A}_*} D\Psi_{\mathcal{A}_*}(x_*) + \mu_{*I_*} D\Psi_{I_*}(x_*)$$

beachten. □

Die Ableitung der in (4.41) und (4.42) definierten Abbildung ist

$$D\Theta_S(x, \lambda, \mu_S, \mu_R) = \begin{pmatrix} \mathcal{H}(x, \lambda, \mu) & D\Phi(x)' & D\Psi_S(x)' & D\Psi_R(x)' \\ D\Phi(x) & 0 & 0 & 0 \\ D\Psi_S(x) & 0 & 0 & 0 \\ 0 & 0 & 0 & -\text{id}_{Z_R} \end{pmatrix}, \quad (4.44)$$

wobei wir hier und im Folgenden die Abkürzung

$$\mathcal{H}(x, \lambda, \mu) := DDF(x) + \lambda DD\Phi(x) + \mu DD\Psi(x) \quad (4.45)$$

verwenden wollen.

Satz 4.95 (Newtonverfahren mit Orakel unter LICQ)

Seien (X, g) und (Y, h) endlichdimensionale Hilberträume, $Z = \mathbb{R}^m$ und der Punkt $x_* \in M$ erfülle die LICQ (vgl. Definition 4.53) und die hinreichenden Bedingungen zweiter Ordnung aus Satz 4.79 mit Lagrange-Multiplikatoren λ_* und μ_* für das Problem (4.10). Sei $\mathcal{A}_* := \mathcal{A}(x_*)$ die aktive Menge bei x_*

Dann ist $D\Theta_{\mathcal{A}_*}(x_*, \lambda_*, \mu_{*\mathcal{A}_*}, \mu_{*I_*})$ stetig invertierbar. Für Startwerte $(x_0, \lambda_0, \mu_{0\mathcal{A}_*}, \mu_{0I_*})$ hinreichend nahe bei $(x_*, \lambda_*, \mu_{*\mathcal{A}_*}, \mu_{*I_*})$ konvergiert das Newton-Verfahren für $\Theta_{\mathcal{A}_*}$ quadratisch gegen $(x_*, \lambda_*, \mu_{*\mathcal{A}_*}, \mu_{*I_*})$.

BEWEIS. Nach (4.43) hat die Ableitung von $\Theta_{\mathcal{A}_*}$ bei $(x_*, \lambda_*, \mu_{*\mathcal{A}_*}, \mu_{*I_*})$ folgende Blockmatrixform:

$$D\Theta_{\mathcal{A}_*}(x_*, \lambda_*, \mu_{*\mathcal{A}_*}, \mu_{*I_*}) = \begin{pmatrix} \mathcal{H}(x_*, \lambda_*, \mu_*) & D\Phi(x_*)' & D\Psi_{\mathcal{A}_*}(x_*)' & D\Psi_{I_*}(x_*)' \\ D\Phi(x_*) & 0 & 0 & 0 \\ D\Psi_{\mathcal{A}_*}(x_*) & 0 & 0 & 0 \\ 0 & 0 & 0 & -\text{id}_{Z_{I_*}} \end{pmatrix}.$$

Analog zu Satz 4.83 folgt mit Lemma 4.85 die stetige Invertierbarkeit der linken, oberen 3×3 -Blocks aus der hinreichenden Bedingung zweiter Ordnung und der LICQ. Hieraus überlegt man sich leicht mit Aufgabe 4.92, dass die Gesamtmatrix auch stetig invertierbar sein muss.

Für $(x, \lambda, \mu_{\mathcal{A}_*}, \mu_{I_*})$ hinreichend nahe bei $(x_*, \lambda_*, \mu_{*\mathcal{A}_*}, \mu_{*I_*})$ ist dann $D\Theta_{\mathcal{A}_*}(x, \lambda, \mu_{\mathcal{A}_*}, \mu_{I_*})$ immernoch stetig invertierbar und der Satz von Newton-Kantorovich Satz 3.27 impliziert quadratische Konvergenz. □

Bemerkung 4.96 Für das Newton-Verfahren muss man lineare Gleichungssysteme der Form

$$\begin{pmatrix} \mathcal{H}(x, \lambda, \mu) & D\Phi(x)' & D\Psi_S(x)' & D\Psi_R(x)' \\ D\Phi(x) & 0 & 0 & 0 \\ D\Psi_S(x) & 0 & 0 & 0 \\ 0 & 0 & 0 & -\text{id}_{Z_R} \end{pmatrix} \begin{pmatrix} \delta x \\ \delta \lambda \\ \delta \mu_S \\ \delta \mu_R \end{pmatrix} = - \begin{pmatrix} DF(x) + \lambda D\Phi(x) + \mu D\Psi(x) \\ \Phi(x) \\ \Psi_S(x) \\ -\mu_R \end{pmatrix} \quad (4.46)$$

lösen. Hierbei sind S und R geordnete Mengen mit $S \amalg R = (1, \dots, m)$.

Aus der letzten Zeile erhält man $\delta \mu_R = -\mu_R$. Addiert man $D\Psi_R(x)' \mu_R = \mu_R D\Psi_R(x)$ zur ersten Zeile und beachtet $\mu D\Psi(x) = \mu_S D\Psi_S(x) + \mu_R D\Psi_R(x)$, so kann man das System verschlanken auf

$$\begin{pmatrix} \mathcal{H}(x, \lambda, \mu) & D\Phi(x)' & D\Psi_S(x)' \\ D\Phi(x) & 0 & 0 \\ D\Psi_S(x) & 0 & 0 \end{pmatrix} \begin{pmatrix} \delta x \\ \delta \lambda \\ \delta \mu_S \end{pmatrix} = - \begin{pmatrix} DF(x) + \lambda D\Phi(x) + \mu_S D\Psi_S(x) \\ \Phi(x) \\ \Psi_S(x) \end{pmatrix}, \quad \delta \mu_R = -\mu_R.$$

Dieses System hat die begrüßenswerte Eigenschaft, dass die auftretende Matrix symmetrisch ist, so dass Lösungsalgorithmen für symmetrische Matrizen benutzt werden können. Da die Matrix bei nichttrivialen Nebenbedingungen aber stets indefinit sein muss, lässt es sich weder mit dem iterativen CG-Verfahren, noch mit einer Cholesky-Zerlegung behandeln. Eine $\mathbf{LDL}^T$ -Zerlegung mit einer unteren Dreiecksmatrix $\mathbf{L}$ und einer Diagonalmatrix $\mathbf{D}$ ist aber möglich. Verwendet der Lösungsalgorithmus eine geeignete Pivotierung, so ist er auch dann noch relativ robust, wenn D Nulleinträge hat, das heißt, wenn die Systemmatrix nicht invertierbar ist.

4.5.2 Aktive-Mengen-Strategien

In der Praxis kennt man die aktive Menge $\mathcal{A}_* = \mathcal{A}(x_*)$ des Optimums aber nicht im Voraus. Sogenannte *Aktive-Mengen-Strategien* beginnen daher mit einem “Schätzer” $\mathcal{A}_0$ von $\mathcal{A}_*$ und versuchen ihn durch eine Folge $\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2, \dots$ zu verbessern. Für eine gegebene aktive Menge $\mathcal{A}_k$ wendet man das Newton-Verfahren für $\Theta_{\mathcal{A}_k}$ an, um ein Quadrupel $(x_k, \lambda_k, \mu_{k\mathcal{A}_k}, \mu_{kI_k})$ mit $\Theta_{\mathcal{A}_k}(x_k, \lambda_k, \mu_{k\mathcal{A}_k}, \mu_{kI_k}) = 0$ zu erhalten. Anschließend überprüft man, ob $\Psi(x_k) \leq 0$, $\mu_k \geq 0$ und $\langle \mu, \Psi(x_k) \rangle = 0$ gelten. Ist dem so, so ist x_k ein KKT-Punkt mit Lagrange-Multiplikatoren λ_k und μ_k . Andernfalls ist $\mathcal{A}_k$ wohl doch nicht die richtige aktive Menge und man modifiziert $\mathcal{A}_k$ anhand dieser Informationen zu einem $\mathcal{A}_{k+1}$ und fährt fort, bis ein KKT-Punkt gefunden ist.

Dieses Vorgehen ist besonders attraktiv für quadratische Optimierungsprobleme (englisch: *quadratic programs*, also solche Optimierungsprobleme, bei denen ein F Polynom zweiten Grades Φ und Ψ beide affin-lineare Abbildungen sind.⁸ Denn bei diesen terminiert das Newton-Verfahren für $\Theta_{\mathcal{A}_k}$ bereits nach einem Schritt. Ein solches Verfahren haben wir in [Algorithmus 8](#) umrissen. Da man in jeder Iteration die “aktive Menge” $\mathcal{A}_k$ höchstens um ein Element ändert, kann es bei sehr vielen Ungleichungsbedingungen entsprechend lange dauern, bis $\mathcal{A}_*$ gefunden ist. Außerdem kann a priori nicht garantiert werden, dass die Folge

⁸Um Lösbarkeit zu garantieren kann man sich auf *stark konvexe*, quadratische Funktionen F beschränken, d. h. $c > 0$ mit $DDF(x)(u, u) \geq c \|u\|^2$ für alle $x \in X$ und alle $u \in X$.

$\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2, \dots$ keine Zykel enthält. In diesem Fall würde Algorithmus 8 nicht terminieren. Dies kann aber durch geeignete Buchführung vermieden werden.

Algorithmus 8 erwähnen wir hier nur, weil er beim *sequential quadratic programming* (SQP), einem klassischen Verfahren für allgemeinere nichtlineare Optimierungsprobleme Anwendung findet. Bei diesem löst man – z. B. mit Algorithmus 8 – eine Folge quadratischer Optimierungsprobleme, die man durch Taylorapproximation aus (4.10) erhält. Wegen der oben beschriebenen Problematik bei vielen Nebenbedingungen gehen wir nicht genauer hierauf ein.

Algorithm 8 Aktive-Mengen-Strategie für quadratische Optimierungsprobleme (QP)**Require:** F strongly convex, quadratic, Φ, Ψ affine, $\Phi(x_0) = 0, \Psi(x_0) \leq 0$,

```

1: function ACTIVESETSTRATEGYQP( $x_0, F, \Phi, \Psi$ )
2:    $k \leftarrow 0$ 
3:    $\text{stop} \leftarrow \text{false}$ 
4:    $\mathcal{A}_0 \leftarrow \{ \}$ 
5:   while  $\neg \text{stop}$  do
6:      $\mathcal{I}_0 \leftarrow \{ 1, \dots, m \} \setminus \mathcal{A}_0$ 
7:      $(\delta x_k, \lambda_k, \mu_{k\mathcal{A}_k}, \mu_{k\mathcal{I}_k}) \leftarrow D\Theta_{\mathcal{A}_k}(x_k, 0, 0, 0)^{-1} \Theta_{\mathcal{A}_k}(x_k, 0, 0, 0)$ 
8:     if  $\delta x_k = 0$  then
9:       if  $(\mu_k)_{\mathcal{A}_k} \geq 0$  then
10:         $\text{stop} \leftarrow \text{true}$  ▷ KKT-Bedingungen erfüllt
11:       else
12:        choose  $j \in \arg \min((\mu_k)_{\mathcal{A}_k})$ 
13:         $\mathcal{A}_{k+1} \leftarrow \mathcal{A}_k \setminus \{ j \}$  ▷ deaktiviere  $j$  mit größter Abweichung von  $\Psi_j \geq 0$ 
14:         $x_{k+1} \leftarrow x_k$ 
15:       end if
16:     else
17:        $\mathcal{W}_k \leftarrow \{ j \in \{ 1, \dots, m \} \mid D\Psi_j(x) \delta x > 0 \}$ 
18:       for  $j \in \mathcal{W}_k$  do
19:         $t_{k,j} \leftarrow -\frac{\Psi_j(x_k)}{D\Psi_j(x) \delta x_k}$  ▷  $\Psi_j(x_k + t_{k,j} \delta x_k) = 0$ 
20:       end for ▷ bestimme Schrittweiten bis zum “Anschlag”
21:       if  $\min \{ t_{k,j} \mid j \in \mathcal{W}_k \} \geq 1$  then
22:         $x_{k+1} \leftarrow x_k + \delta x_k$  ▷ nutze vollen Newton-Schritt, wenn zulässig
23:       else
24:        choose  $j \in \arg \min(\{ t_{k,j} \mid j \in \mathcal{W}_k \})$ 
25:         $x_{k+1} \leftarrow x_k + t_{k,j} \delta x_k$  ▷ laufe so weit es geht in Richtung  $\delta x_k$ 
26:         $\mathcal{A}_{k+1} \leftarrow \mathcal{A}_k \cup \{ j \}$  ▷ aktiviere  $j$ 
27:       end if
28:     end if
29:      $k \leftarrow k + 1$ 
30:   end while
31:   return  $(x_k, \lambda_k, \mu_k)$  ▷  $(x_k, \lambda_k, \mu_k)$  ist KKT-Punkt
32: end function

```

4.5.3 Semi-glattes Newton-Verfahren bei Ungleichungsnebenbedingungen

Wie wir gesehen haben, machen uns die Ungleichungen in den KKT-Bedingungen einige Probleme. [Satz 4.95](#) sagt uns, dass die KKT-Bedingungen mit dem Newton-Verfahren gelöst werden können, wenn die aktive Menge $\mathcal{A}_* = \mathcal{A}(x_*)$ eines Optimums x_* bekannt ist. Bei m gibt es aber 2^m potentiell mögliche aktive Menge; sie alle durchzuprobieren ist für große m nicht effizient.

Nehmen wir uns noch einmal diejenigen KKT-Bedingungen [\(4.13\)](#), [\(4.14\)](#) und [\(4.15\)](#) vor, die uns das Leben so schwer machen:

$$\Psi_j(x_*) \leq 0, \quad \mu_{*j} \geq 0 \quad \text{und} \quad \mu_{*j} \Psi_j(x_*) = 0.$$

Überraschenderweise kann man diese drei Bedingungen zu einer Gleichheitsbedingung zusammenfassen (siehe [Aufgabe 4.101](#)):

$$\chi(\Psi_j(x_*), \mu_{*j}) = 0 \tag{4.47}$$

mit der Funktion

$$\chi: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}, \quad \chi(s, t) := \max(s + t, 0) - t = \begin{cases} s, & s + t \geq 0, \\ -t, & s + t < 0. \end{cases}$$

Die KKT-Bedingungen werden hierdurch zu den folgenden *Gleichungen* vereinfacht:

$$DF(x_*) + \lambda_* D\Phi(x_*) + \mu_* D\Psi(x_*) = 0, \tag{4.48}$$

$$\Phi(x_*) = 0, \tag{4.49}$$

$$\chi(\Psi_j(x_*), \mu_{*j}) = 0 \quad \text{für } j \in \{1, \dots, m\}. \tag{4.50}$$

Sind auch X und Y endlichdimensional, so sind das $\dim(X) + \dim(Y) + \dim(Z)$ Gleichungen in $\dim(X) + \dim(Y) + \dim(Z)$ Variablen. Das ist schon einmal vielversprechend. Allerdings scheint es so, als könnte man hier das Newton-Verfahren nicht anwenden, denn χ ist zwar auf der Menge $\{(s, t) \in \mathbb{R}^2 \mid s + t \neq 0\}$ differenzierbar; auf der Menge $\{(s, t) \in \mathbb{R}^2 \mid s + t = 0\}$ allerdings nicht. Trotzdem definieren wir

$$\Theta: X \times Y' \times Z' \rightarrow X' \times Y \times Z \quad \Theta(x, \lambda, \mu) := \begin{pmatrix} DF(x) + \lambda D\Phi(x) + \mu D\Psi(x) \\ \Phi(x) \\ \chi(\Psi(x), \mu) \end{pmatrix}, \tag{4.51}$$

wobei wir abkürzend folgende Schreibweise verwenden

$$\chi(\Psi(x), \mu) := \begin{pmatrix} \chi(\Psi_1(x), \mu_1) \\ \vdots \\ \chi(\Psi_m(x), \mu_m) \end{pmatrix}.$$

Für das Newton-Verfahren benötigen wir eine Ableitung – oder etwas ähnliches. Wir mogeln⁹ jetzt ein ganz kleines Bisschen und *definieren*

$$D\chi(s, t) := \begin{pmatrix} \partial_s \chi(s, t) & \partial_t \chi(s, t) \end{pmatrix} := \begin{cases} (1, 0), & s + t \geq 0, \\ (0, -1), & s + t < 0. \end{cases}$$

Mit der Kettenregel, der Abkürzung $\mathcal{H}(x, \lambda, \mu) := DDF(x) + \lambda DD\Phi(x) + \mu DD\Psi(x)$ und mit einem feste zugekniffenen Auge rechnet man nun die Ableitung nach:

$$D\Theta(x, \lambda, \mu) := \begin{pmatrix} \mathcal{H}(x, \lambda, \mu) & D\Phi(x)' & D\Psi(x)' \\ D\Phi(x) & 0 & 0 \\ \text{diag}(\sigma) D\Psi(x) & 0 & \text{diag}(\sigma - 1) \end{pmatrix},$$

wobei

$$\sigma_j = \sigma_j(\Psi(x), \mu) := \begin{cases} 1, & \Psi_j(x) + \mu_j \geq 0, \\ 0, & \Psi_j(x) + \mu_j < 0. \end{cases}$$

Ein (gedämpfter) Newton-Schritt besteht demnach im Lösen des Gleichungssystems

$$\begin{pmatrix} \mathcal{H}(x, \lambda, \mu) & D\Phi(x)' & D\Psi(x)' \\ D\Phi(x) & 0 & 0 \\ \text{diag}(\sigma) D\Psi(x) & 0 & \text{diag}(\sigma - 1) \end{pmatrix} \begin{pmatrix} \delta x \\ \delta \lambda \\ \delta \mu \end{pmatrix} = - \begin{pmatrix} D\Phi(x) + \lambda D\Psi(x) \\ \Phi(x) \\ \chi(\Psi(x), \mu) \end{pmatrix} \quad (4.52)$$

und der Aufdatierung

$$x + \tau \delta x, \quad \lambda + \tau \delta \lambda \quad \text{und} \quad \mu + \tau \delta \mu$$

mit einer geeigneten Schrittweite $\tau > 0$.

Das zu lösende System können wir noch etwas vereinfachen. Dafür benötigen wir die folgende Definition.

Definition 4.97 Wir bezeichnen die geordnete Menge aller $j \in \{1, \dots, m\}$, für die $\Psi_j(x) + \mu_j \geq 0$ gilt, mit $\mathcal{S} = \mathcal{S}(x, \mu)$. In der englischsprachigen Literatur wird sie als *primal-dual active set* bezeichnet; sie sollte aber nicht mit der aktiven Menge $\mathcal{A}(x)$ verwechselt werden. Das Komplement bezeichnen wir mit $\mathcal{R} = \mathcal{R}(x, \mu)$.

Manche Autoren verlangen auch $\Psi_j(x) + \mu_j > 0$ für aktive Indizes und $\Psi_j(x) + \mu_j \leq 0$ für inaktive. Wir haben die Wahlen gerade so getroffen, dass folgendes Lemma gilt:

⁹Was wir hier als Mogeln bezeichnen wird durch eine rigorose Theorie der semi-glaten Newton-Verfahren getragen. Als weitere Lektüre sei z. B. das sehr lesenswerte Skript [5] und die darin referenzierte Literatur empfohlen.

Lemma 4.98 Ist x_* ein KKT-Punkt mit Lagrange-Multiplikatoren λ_* und μ_* , so gilt

$$\mathcal{S}(x_*, \mu_*) = \mathcal{A}(x_*) \quad \text{und} \quad \mathcal{R}(x_*, \mu_*) = \mathcal{I}(x_*).$$

Zu jedem Punkt $(x_0, \mu_0) \in U \times Z$ gibt es eine Umgebung V von (x_0, μ_0) in $U \times Z'$, so dass

$$\mathcal{S}(x_0, \mu_0) \supset \mathcal{S}(x, \mu) \quad \text{und} \quad \mathcal{R}(x_0, \mu_0) \subset \mathcal{R}(x, \mu) \quad \text{für alle } (x, \mu) \in V.$$

BEWEIS. Ist $j \in \mathcal{A}(x_*)$, so gilt $\Psi_j(x_*) \geq 0$ und $\mu_{*j} \geq 0$, also $\Psi_j(x_*) + \mu_{*j} \geq 0$ und daher $j \in \mathcal{S}(x_*, \mu_*)$. Ist andererseits aber $j \in \mathcal{I}(x_*)$, so gilt $\Psi_j(x_*) < 0$ und $\mu_{*j} = 0$, also $\Psi_j(x_*) + \mu_{*j} < 0$ und $j \in \mathcal{R}(x_*, \mu_*)$. Dies zeigt die erste Aussage. Die zweite folgt direkt aus der Stetigkeit der Abbildungen $(x, \mu) \mapsto \Psi_j(x) + \mu_j$ und der Tatsache, dass es nur endlich viele skalare Ungleichungsbedingungen gibt. $\square$

Wenden wir auf das Gleichungssystem die Permutation $(\mathcal{S}, \mathcal{R})$ in der Z -Komponente an, so erhalten wir

$$\begin{pmatrix} \mathcal{H}(x, \lambda, \mu) & D\Phi(x)' & D\Psi_{\mathcal{S}}(x)' & D\Psi_{\mathcal{R}}(x)' \\ D\Phi(x) & 0 & 0 & 0 \\ D\Psi_{\mathcal{S}}(x) & 0 & 0 & 0 \\ 0 & 0 & 0 & -\text{id}_{Z_{\mathcal{R}}} \end{pmatrix} \begin{pmatrix} \delta x \\ \delta \lambda \\ \delta \mu_{\mathcal{S}} \\ \delta \mu_{\mathcal{R}} \end{pmatrix} = - \begin{pmatrix} DF(x) + \lambda D\Phi(x) + \mu D\Psi(x) \\ \Phi(x) \\ \chi(\Psi_{\mathcal{S}}(x), \mu_{\mathcal{S}}) \\ -\mu_{\mathcal{R}} \end{pmatrix} \quad (4.53)$$

Dieses System kommt uns sehr bekannt vor (vgl. [Bemerkung 4.96](#)). Man beachte die kleine Modifikation der dritten Komponente der rechten Seite. Zusammen mit Armijo-Backtracking für das Residuum von Θ zur Bestimmung der Schrittweite $\tau > 0$ ergibt sich [Algorithmus 9](#).

Wenn seine Bedingungen auch nicht gerade einfach in der Praxis zu überprüfen sind, so hat man doch folgenden bemerkenswerten Konvergenzsatz:

Satz 4.99 (Lokale Konvergenz des semi-glatte Newton-Verfahrens)

Im Optimierungsproblem (4.10) seien F , Φ und Ψ jeweils zweimal stetig differenzierbar. Sei (x_*, λ_*, μ_*) ein KKT-Tripel. Es gebe eine Umgebung W von $p_* := (x_*, \lambda_*, \mu_*)$ in $U \times Y' \times Z'$, Konstanten $C > 0$, $L \geq 0$ und $\alpha \in]0, 1]$ mit:

1. $D\Theta(p)$ ist stetig invertierbar mit $\|D\Theta(p)^{-1}\| \leq C$ für alle $p \in W$.
2. Für alle $p \in W$ gilt

$$\|D^2F(x) - D^2F(y)\|, \|D^2\Phi(x) - D^2\Phi(y)\|, \|D^2\Psi(x) - D^2\Psi(y)\| \leq L \|x - y\|^\alpha.$$

Dann gibt es ein $r > 0$, so dass für jeden Startwert $p_0 \in \bar{B}(p_*; r)$ das semi-glatte Newton-Verfahren

$$p_{k+1} = p_k - D\Theta(p_k)^{-1} \Theta(p_k)$$

superlinear gegen p_* konvergiert:

$$p_k \xrightarrow{k \rightarrow \infty} p \quad \text{und} \quad \|p_{k+1} - p_*\| \leq \Lambda C \|p_k - p_*\|^{(1+\alpha)}.$$

Algorithm 9 Primal-Dual-Active-Set-Algorithmus/semi-glattes Newtonverfahren

Require: $\sigma \in \mathbb{R}$, $\gamma \in]0, 1[$, $\text{TOL} > 0$

```

1: function PRIMALDUALACTIVEMETHOD( $x_0, \lambda_0, \mu_0, F, \Phi, \Psi$ )
2:    $k \leftarrow 0$ 
3:   while  $\neg((x_k, \lambda_k, \mu_k)$  is KKT-Tripel up to TOL) do
4:      $S_k \leftarrow \mathcal{S}(x_k, \mu_k)$ 
5:      $\mathcal{R}_k \leftarrow \mathcal{R}(x_k, \mu_k)$ 
6:      $(\delta\mu_k)_{\mathcal{R}_k} \leftarrow -(\mu_k)_{\mathcal{R}_k}$ 
7:      $\xi \leftarrow DF(x_k) + \lambda_k D\Phi(x_k) + (\mu_k)_{S_k} D\Psi_{S_k}(x_k)$ 
8:     Solve
       
$$\begin{pmatrix} \mathcal{H}(x_k, \lambda_k, \mu_k) & D\Phi(x_k)' & D\Psi_{S_k}(x_k)' \\ D\Phi(x_k) & 0 & 0 \\ D\Psi_{S_k}(x_k) & 0 & 0 \end{pmatrix} \begin{pmatrix} \delta x_k \\ \delta \lambda_k \\ (\delta\mu_k)_{S_k} \end{pmatrix} = - \begin{pmatrix} \xi_k \\ \Phi(x_k) \\ \chi(\Psi_{S_k}(x_k), (\mu_k)_{S_k}) \end{pmatrix}$$

9:      $t \leftarrow 1$ 
10:    while  $\|\Theta((x_{k+1}, \lambda_{k+1}, \mu_{k+1}) + t(\delta x_k, \delta \lambda_k, \delta \mu_k))\|^2 > (1 - \sigma t) \|\Theta(x_k, \lambda_k, \mu_k)\|^2$  do
11:       $t \leftarrow \gamma t$ 
12:    end while ▷ backtracking line search
13:     $(x_{k+1}, \lambda_{k+1}, \mu_{k+1}) \leftarrow (x_{k+1}, \lambda_{k+1}, \mu_{k+1}) + t(\delta x_k, \delta \lambda_k, \delta \mu_k)$ 
14:     $k \leftarrow k + 1$ 
15:  end while
16:  return  $(x_k, \lambda_k, \mu_k)$  ▷  $(x_k, \lambda_k, \mu_k)$  ist KKT-Punkt bis auf Fehler  $< \text{TOL}$ 
17: end function

```

BEWEIS. Wir wählen ein beliebiges $q \in]0, 1[$. Nach [Lemma 4.100](#) gibt es ein $\Lambda \geq 0$ und ein $r > 0$ so dass gilt

$$\|\Theta(p_*) - \Theta(p) - D\Theta(p)(p_* - p)\| \leq \Lambda \|p_* - p\|^{(1+\alpha)} \quad \text{für alle } p \in \bar{B}(p_*; r). \quad (4.54)$$

Wir können nun $r > 0$ so weit verkleinern, dass $\bar{B}(p_*; r)$ in W enthalten ist und so dass

$$\|\Theta(p_*) - \Theta(p) - D\Theta(p)(p_* - p)\| \leq \frac{q}{C} \|p_* - p\| \quad \text{für alle } p \in \bar{B}(p_*; r) \text{ gilt.}$$

Analog wie im Beweis des Satzes von Newton-Kantorovich (siehe [Satz 3.27](#)) rechnen wir nun für beliebiges $p \in \bar{B}(p_*; r)$:

$$\begin{aligned} \|(p - D\Theta(p)^{-1} \Theta(p)) - p_*\| &= \|(p - D\Theta(p)^{-1} \Theta(p)) + D\Theta(p)^{-1} \Theta(p_*) - p_*\| \\ &\leq \|D\Theta(p)^{-1}\| \|\Theta(p_*) - \Theta(p) - D\Theta(p)(p_* - p)\| \\ &\leq C \|\Theta(p_*) - \Theta(p) - D\Theta(p)(p_* - p)\| \\ &\leq C \Lambda \|p_* - p\|^{1+\alpha}. \end{aligned} \quad (4.55)$$

Insbesondere gilt $\|(p - D\Theta(p)^{-1} \Theta(p)) - p_*\| \leq q \|p_* - p\| < r$. Die Newton-Iterierten sind also wohldefiniert und es gilt

$$\|p_{k+1} - p_*\| \leq q \|p_k - p_*\| \leq \dots \leq q^{k+1} \|p_0 - p_*\|.$$

Wegen $0 < q < 1$ folgt Konvergenz $p_k \rightarrow p_*$. Aus (4.55) folgt somit die super-lineare Konvergenz. $\square$

Lemma 4.100 *Unter den Bedingungen von Satz 4.99 gilt für die Abbildung $\Theta: U \times Y' \times Z' \rightarrow X' \times Y \times Z$ aus (4.51), gilt für festes $p_* := (x_*, \lambda_*, \mu_*)$ und beliebiges $p := (x, \lambda, \mu)$ die folgende Asymptotik:*

$$\|\Theta(p_*) - \Theta(p) - D\Theta(p)(p_* - p)\| = O(\|p - p_*\|^{(1+\alpha)}) \quad \text{für } p \rightarrow p_*.$$

BEWEIS. Die X' - und die Y -Komponente werden durch den Satz von Taylor abgedeckt. Nur die Z -Komponente von Θ ist daher strittig. Mit $q = (\Psi(x), \mu)$ und $q_* = (\Psi(x_*), \mu_*)$

$$\begin{aligned} \chi(q_*) - \chi(q) - D\chi(q) \begin{pmatrix} D\Psi(x)(x_* - x) \\ \mu_* - \mu \end{pmatrix} \\ = \chi(q_*) - \chi(q) - D\chi(q) \begin{pmatrix} \Psi(x) - \Psi(x_*) \\ \mu_* - \mu \end{pmatrix} + D\chi(q) \begin{pmatrix} \Psi(x_*) - \Psi(x) - D\Psi(x)(x_* - x) \\ 0 \end{pmatrix} \\ = \chi(q_*) - \chi(q) - D\chi(q)(q_* - q) + D\chi(q) \begin{pmatrix} \Psi(x_*) - \Psi(x) - D\Psi(x)(x_* - x) \\ 0 \end{pmatrix}. \end{aligned}$$

Aus Aufgabe 4.102 und $\|q_* - q\| = O(\|p_* - p\|)$ folgt

$$\chi(q_*) - \chi(q) - D\chi(q)(q_* - q) = O(\|q_* - q\|^2) = O(\|p_* - p\|^2).$$

Falls Ψ von der Hölder-Klasse $C^{2,\alpha}$, $\alpha \in [0, 1]$ ist, so impliziert der Satz von Taylor $\Psi(x_*) - \Psi(x) - D\Psi(x)(x_* - x) = O(\|x_* - x\|^{(1+\alpha)})$. Wegen $\|D\chi(q)\| \leq 1$ hat man dann

$$D\chi(q) \begin{pmatrix} \Psi(x_*) - \Psi(x) - D\Psi(x)(x_* - x) \\ 0 \end{pmatrix} = O(\|x_* - x\|^{(1+\alpha)}) \leq O(\|p_* - p\|^{(1+\alpha)}).$$

Folglich gilt für die Z -Komponente von Θ folgende Asymptotik:

$$\chi(\Psi(x_*), \mu_*) - \chi(\Psi(x), \mu) - D\chi(\Psi(x), \mu) \begin{pmatrix} D\Psi(x)(x_* - x) \\ \mu_* - \mu \end{pmatrix} = O(\|p_* - p\|^{(1+\alpha)}). \quad \square$$

4.5.4 Übungsaufgaben

Aufgabe 4.101

Wir betrachten die Abbildung $\chi: \mathbb{R}^2 \rightarrow \mathbb{R}$

$$\chi(s, t) := \max(s + t, 0) - t = \begin{cases} s, & s + t \geq 0, \\ -t, & s + t < 0. \end{cases}$$

Zeigen Sie, dass folgende Aussagen für $(s, t) \in \mathbb{R}^2$ äquivalent sind:

1. $s \leq 0$ und $t \geq 0$ und $st = 0$.
2. $\chi(s, t) = 0$.

Aufgabe 4.102

Zur Funktion χ aus [Aufgabe 4.101](#) definieren wir eine “verallgemeinerte Ableitung” $D\chi: \mathbb{R}^2 \rightarrow L(\mathbb{R}^2; \mathbb{R})$ wie folgt:

$$D\chi(s, t) := \begin{pmatrix} \partial_s \chi(s, t) & \partial_t \chi(s, t) \end{pmatrix} := \begin{cases} (1, 0), & s + t \geq 0, \\ (0, -1), & s + t < 0. \end{cases}$$

Zeigen Sie für festes $q_0 = (s_0, t_0) \in \mathbb{R}^2$ und für $q = (s, t)$ die Identität

$$\chi(q) - \chi(q_0) - D\chi(q)(q - q_0) = \begin{cases} s_0 + t_0, & s_0 + t_0 < 0 \text{ und } s + t \geq 0, \\ -s_0 - t_0, & s_0 + t_0 > 0 \text{ und } s + t < 0, \\ 0, & \text{sonst.} \end{cases}$$

und folgern Sie

$$|\chi(q) - \chi(q_0) - D\chi(q)(q - q_0)| = o(\|q - q_0\|^2) \quad \text{für } q \rightarrow q_0.$$

Aufgabe 4.103

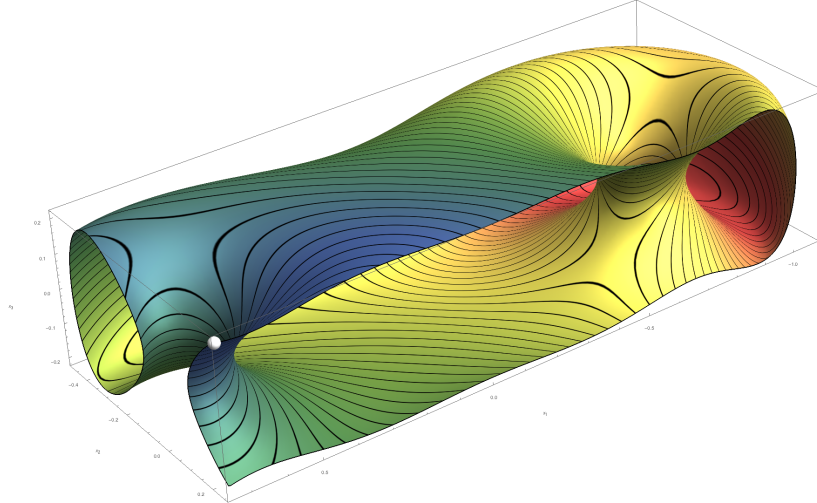
Im Setting von [Aufgabe 3.75](#), [Aufgabe 4.93](#) können wir auch nach Ruhelagen von elastischen oder starren Ketten suchen, die innerhalb eines vorgegebenen Halbraumes (zum Beispiel oberhalb des als x - y -Ebene idealisierten Erdbodens) liegen. Formulieren Sie dies als Ungleichheitsbedingungen und verwenden Sie das Semi-glatte Newton-Verfahren, um einige Beispiele numerisch zu lösen.

Aufgabe 4.104 (Primal-Dual-Active-Set-Algorithmus)

Implementieren Sie den Primal-Dual-Active-Set-Algorithmus [Algorithmus 9](#) und eine Routine zum Testen (i) der notwendigen Bedingungen zweiter Ordnung und (ii) der hinreichenden Bedingungen zweiter Ordnung. Testen Sie diese an folgendem Optimierungsproblem:

$$\begin{aligned} &\text{Minimiere} \quad (x_1 - 1)^2 + (x_2 + 1)^2 + (x_3 - 2)^2 \\ &\text{u. d. N.} \quad \left((x_1^2 + x_2^2)^2 - x_1^2 + x_2^2 \right)^2 + x_3^2 - \frac{1}{20} = 0, \\ &\quad x_1 - \frac{3}{4} \leq 0, \\ &\quad x_2 - \frac{1}{4} \leq 0, \\ &\quad (x_1, x_2, x_3) \in \mathbb{R}^3. \end{aligned}$$

Die globale Lösung dieses Problems lautet $x_* = \left(\frac{3}{4}, \frac{1}{4}, \sqrt{\frac{779}{20480}}\right)$. Der zulässige Bereich mit Höhenlinien der Zielfunktion und die globale Lösung sind in folgender Abbildung skizziert:



Hinweis: Wählen Sie Startpunkte in der Nähe von x_* . Wählen Sie für das backtracking die Parameter $\sigma = 0.01$ und $\gamma = 0.5$ und beschränken Sie die maximale Anzahl an backtracking-Iterationen auf 30.

Beachten Sie außerdem, dass $D\Psi(x)$ singulär werden kann! Startet man nicht gerade in der Nähe einer singulären Stelle, macht das in Regel aber nichts.

4.6 Penalty-Verfahren

Sei X ein normierter Raum, $U \subset X$ eine offene Menge, $F: U \rightarrow \mathbb{R}$ eine stetige Funktion und $M \subset X$ eine abgeschlossene Menge. Wir betrachten das Optimierungsproblem

$$\begin{array}{ll} \text{Minimiere} & F(x) \\ \text{u. d. N.} & x \in M \cap U. \end{array}$$

Die Idee von Penalty-Verfahren ist, ein restringiertes Optimierungsproblem in ein unrestringiertes umzuwandeln, dass die Unzulässigkeit eines Punktes $x \in X$ in geeigneter Weise quantifiziert wird und einer neuen Zielfunktion zugeschlagen wird.

Definition 4.105 Eine *Straffunktion* oder *penalty function* für M ist eine stetige Funktion $P: U \rightarrow [0, \infty[$ mit

$$P(x) = 0 \quad \text{genau dann, wenn} \quad x \in M.$$

Für einen *Strafparameter* (*penalty parameter*) $\alpha \in [0, \infty[$ definieren wir die *bestrafte Zielfunktion* (*penalized objective function*) durch

$$F_\alpha(x) := F(x) + \alpha P(x).$$

Beispiel 4.106 Im Optimierungsproblem

$$\begin{array}{ll} \text{Minimiere} & F(x) \\ \text{u. d. N.} & \Psi(x) \in K, \\ & x \in C, \\ & x \in U. \end{array} \tag{4.56}$$

(vgl. (4.18)) sei $P_C: U \rightarrow [0, \infty[$ eine Straffunktion für die abgeschlossene, konvexe Menge $C \subset X$ und $P_K: Z \rightarrow [0, \infty[$ eine Straffunktion für den abgeschlossenen, konvexen Kegel $K \subset Z$. Dann ist $P(x) := P_C(x) + P_K(\Psi(x))$ eine Straffunktion für die Menge $M = \{x \in C \mid \Psi(x) \in K\}$.

Ist $x \in M$ ein zulässiger Punkt, so gilt $F_\alpha(x) = F(x)$ für alle $\alpha \in [0, \infty[$. Im Allgemeinen hat man

$$\lim_{\alpha \rightarrow \infty} F_\alpha(x) = \begin{cases} F(x), & x \in M, \\ \infty, & x \in U \setminus M. \end{cases}$$

Die Hoffnung beim Anwenden von Penalty-Verfahren ist nun, dass die globalen Minimierer von F_α noch in einem geeigneten Sinne gegen die globalen Minimierer von (4.6) konvergieren:

$$\arg \min(F_\alpha) \xrightarrow{\alpha \rightarrow \infty} \arg \min(F),$$

zum Beispiel in Hausdorff-Distanz.¹⁰

¹⁰Konvergenz in Hausdorff-Distanz ist so ziemlich der schönste Konvergenzbegriff, den man sich hier wünschen kann. Oft kann man aber nur Konvergenz bezüglich schwächerer Begriffe wie zum Beispiel *Kuratowski-Konvergenz* (siehe z. B. Kapitel 4 in [12]) sicherstellen.

Beispiel 4.107 Im Optimierungsproblem (4.10) seien (X, g_X) , (Y, g_Y) und (Z, g_Z) Hilberträume. Dann ist

$$P(x) := \frac{1}{2} \|\Phi(x)\|_{g_Y}^2 + \frac{1}{2} \|\max(0, \Psi(x))\|_{g_Z}^2$$

eine Straffunktion für die Menge $M = \{x \in X \mid \Phi(x) = 0 \text{ und } \Psi(x) \leq 0\}$. Hierbei verstehen wir $\max(0, \Psi(x))$ wieder komponentweise, also

$$\max(0, \Psi(x)) := \left(\max(0, \Psi_1(x)), \dots, \max(0, \Psi_m(x)) \right)^T.$$

Satz 4.108 Über das Optimierungsproblem (4.6) nehmen wir lediglich an, dass M nichtleer ist. Sei $(\alpha_k)_{k \in \mathbb{N}}$ eine monoton steigende Folge positiver reeller Zahlen mit $\lim_{k \rightarrow \infty} \alpha_k = \infty$ und sei x_k jeweils ein globaler Minimierer von F_{α_k} . Dann gilt:

1. Die Folge $(F_{\alpha_k}(x_k))_{k \in \mathbb{N}}$ ist monoton wachsend.
2. Die Folge $(P(x_k))_{k \in \mathbb{N}}$ ist monoton fallend.
3. Die Folge $(F(x_k))_{k \in \mathbb{N}}$ ist monoton wachsend.
4. $\lim_{k \rightarrow \infty} P(x_k) = 0$.
5. Jeder Häufungspunkt der Folge $(x_k)_{k \in \mathbb{N}}$ ist ein globaler Minimierer von (4.6).¹¹
6. Gilt für endliches $k \in \mathbb{N}$ schon $P(x_k) = 0$, so ist x_k auch ein globaler Minimierer von (4.6).

BEWEIS. Zu Punkt 1: Da x_k das Minimum von F_{α_k} ist und wegen $\alpha_k < \alpha_{k+1}$ gilt

$$F_{\alpha_k}(x_k) \leq F_{\alpha_k}(x_{k+1}) \leq F_{\alpha_{k+1}}(x_{k+1}).$$

Zu Punkt 2: Die Addition der Ungleichungen

$$F_{\alpha_k}(x_k) \leq F_{\alpha_k}(x_{k+1}) \quad \text{und} \quad F_{\alpha_{k+1}}(x_{k+1}) \leq F_{\alpha_{k+1}}(x_k)$$

ergibt

$$\left(\cancel{F(x_k)} + \alpha_k P(x_k) \right) + \left(\cancel{F(x_{k+1})} + \alpha_{k+1} P(x_{k+1}) \right) \leq \left(\cancel{F(x_{k+1})} + \alpha_k P(x_{k+1}) \right) + \left(\cancel{F(x_k)} + \alpha_{k+1} P(x_k) \right).$$

¹¹In der Sprache der Kuratowski-Konvergenz formuliert man dies auch als $\limsup_{k \rightarrow \infty} \arg \min(F_{\alpha_k}) \subset \arg \min(\widetilde{F})$ für die Funktion

$$\widetilde{F}(x) = \begin{cases} F(x), & x \in M, \\ \infty, & x \notin M. \end{cases}$$

Das ist gewissermaßen schon die halbe Miete bei der Kuratowski-Konvergenz. Mehr dazu in Kapitel 4 in [12].

Algorithm 10 Abstraktes Penalty-Verfahren

Require: $\alpha_0 > 0, S > 1, \text{TOL}_F > 0, \text{TOL}_P > 0$.

```

1: function PENALTYMETHOD( $x_0, F, P, \alpha_0$ )
2:    $k \leftarrow 0$ 
3:   while  $P(x_k) > \text{TOL}_P$  do
4:     Find  $x_{k+1}$  with  $\|DF(x_{k+1}) + \alpha_k DP(x_{k+1})\| < \text{TOL}_F / \alpha_k$ 
       by using an iterative method for unconstrained optimization
       with  $x_k$  as starting point (e.g. BFGS).
5:      $\alpha_{k+1} \leftarrow S \alpha_k$ 
6:      $k \leftarrow k + 1$ 
7:   end while
8:   return  $x_k$ 
9: end function

```

Durch Umformung erhalten wir

$$(\alpha_{k+1} - \alpha_k) (P(x_{k+1}) - P(x_k)) \leq 0.$$

Wegen $(\alpha_{k+1} - \alpha_k) \geq 0$ gilt also $P(x_{k+1}) \leq P(x_k)$.

Zu [Punkt 3](#): Aus $F_{\alpha_k}(x_k) \leq F_{\alpha_k}(x_{k+1})$ und $P(x_{k+1}) \leq P(x_k)$ folgt

$$\begin{aligned}
F(x_k) &= F_{\alpha_k}(x_k) - \alpha_k P(x_k) \\
&\leq F_{\alpha_k}(x_{k+1}) - \alpha_k P(x_k) \\
&= F(x_{k+1}) + \alpha_k (P(x_{k+1}) - P(x_k)) \leq F(x_{k+1}).
\end{aligned}$$

Zu [Punkt 4](#): Sei $m_* = \inf_{x \in M} F(x)$ der Minimalwert. Weil M nichtleer ist gilt $m_* < \infty$. Aus [Punkt 3](#) folgt

$$F(x_0) + \alpha_k P(x_k) \leq F(x_k) + \alpha_k P(x_k) = F_{\alpha_k}(x_k) \leq \inf_{x \in M} F_{\alpha_k}(x) = \inf_{x \in M} F(x) = m_*. \quad (4.57)$$

Wegen $\alpha_k \rightarrow \infty$ muss also $P(x_k) \rightarrow 0$ gelten.

Zu [Punkt 5](#): Sei x_* Häufungspunkt der Folge $(x_k)_{k \in \mathbb{N}}$. Dann gibt es eine Teilfolge, die gegen x_* konvergiert. Ohne Einschränkung der Allgemeinheit bezeichnen wir diese Folge wieder mit $(x_k)_{k \in \mathbb{N}}$. Weil P stetig ist, gilt wegen [Punkt 4](#) $P(x_*) = \lim_{k \rightarrow \infty} P(x_k) = 0$. Weil aber P eine Straffunktion für M ist, muss $x_* \in M$ zulässig sein. Außerdem gilt wegen (4.57) auch $F(x_k) \leq F_{\alpha_k}(x_k) \leq m_*$ und daher $F(x_*) = \lim_{k \rightarrow \infty} F(x_k) \leq m_*$. Es ist damit x_* eine Lösung von (4.6).

Zu [Punkt 6](#): Ist schon x_k zulässig, so gilt ebenfalls $F(x_k) \leq F_{\alpha_k}(x_k) \leq m_*$. □

Unter Rückgriff auf unrestringierte Optimierungsverfahren für das Minimieren von F_{α_k} erhält man nun auf natürliche Weise [Algorithmus 10](#).

Algorithm 11 Penalty-Verfahren mit Approximation der Lagrange-Multiplikatoren

Require: $\alpha_0 > 0, S > 1, \text{TOL}_F > 0, \text{TOL}_P > 0$.

```

1: function PENALTYMETHODLAGRANGE( $x_0, F, \Phi, \Psi, \alpha_0$ )
2:    $k \leftarrow 0$ 
3:   while  $\|\Phi(x_k)\|^2 + \|\max(0, \Psi(x_k))\|^2 > \text{TOL}_P$  do
4:     Find  $x_{k+1}$  with  $\|DF(x_{k+1}) + \alpha DP(x_{k+1})\| < \text{TOL}_F / \alpha_k$  using an iterative method for
       unconstrained optimization
       with  $x_k$  as starting point (e.g. BFGS). ▷ vgl. (4.58)
5:      $\alpha_{k+1} \leftarrow S \alpha_k$ 
6:      $k \leftarrow k + 1$ 
7:   end while
8:    $\lambda_k \leftarrow \alpha_k J_{g_Y} \Phi(x_k)$ 
9:    $\mu_k \leftarrow \alpha_k J_{g_Z} \max(0, \Psi(x_k))$ 
10:  return  $(x_k, \lambda_k, \mu_k)$ 
11: end function

```

Bemerkung 4.109 (Approximation der Lagrange-Multiplikatoren)

Wir wollen das Setting von [Beispiel 4.107](#) noch einmal genauer betrachten. Nehmen wir an, dass F, Φ und Ψ stetig differenzierbar sind. Die bestrafte Zielfunktion lautet dann

$$F_\alpha(x) = F(x) + \frac{\alpha}{2} \|\Phi(x)\|_{g_Y}^2 + \frac{\alpha}{2} \|\max(0, \Psi(x))\|_{g_Z}^2$$

und ihre Ableitung ist gegeben durch

$$DF_\alpha(x)u = DF(x)u + \alpha g_Y(\Phi(x), D\Phi(x)u) + \alpha g_Z(\max(0, \Psi(x)), D\Psi(x)u) \quad \text{für } u \in X.$$

Das heißt, es gilt

$$DF_\alpha(x) = DF(x) + \alpha (J_{g_Y} \Phi(x)) \cdot D\Phi(x) + \alpha (J_{g_Z} \max(0, \Psi(x)) \cdot D\Psi(x). \quad (4.58)$$

Ist x_k ein Optimum von F_{α_k} so lauten die notwendigen Bedingungen erster Ordnung mit $\lambda_k := \alpha_k J_{g_Y}(\Phi(x_k))$ und $\mu_k := \alpha_k J_{g_Z}(\max(0, \Psi(x_k))) \geq 0$ gerade so:

$$DF(x_k) + \lambda_k D\Phi(x_k) + \mu_k D\Psi(x_k) = DF_{\alpha_k}(x_k) = 0.$$

Konvergieren die Folgen $(x_k)_{k \in \mathbb{N}}$, $(\lambda_k)_{k \in \mathbb{N}}$ und $(\mu_k)_{k \in \mathbb{N}}$ gegen x_* , λ_* bzw. μ_* , so gilt

$$DF(x_*) + \lambda_* D\Phi(x_*) + \mu_* D\Psi(x_*) = 0 \quad \text{und} \quad \mu_* \geq 0.$$

[Satz 4.108](#) entnehmen wir, dass x_* zulässig ist, dass also $\Phi(x_*) = 0$ und $\Psi(x_*) \leq 0$ gelten. Außerdem hat man

$$\begin{aligned}
0 &\geq \mu_* \Psi(x_*) = \lambda_* \Phi(x_*) + \mu_* \Psi(x_*) \\
&= \lim_{k \rightarrow \infty} (\lambda_k \Phi(x_k) + \mu_k \Psi(x_k)) \\
&= \lim_{k \rightarrow \infty} (\alpha_k \|\Phi(x_k)\|_{g_Y}^2 + \alpha_k \|\max(0, \Psi(x_k))\|_{g_Z}^2) \geq 0,
\end{aligned}$$

also $\mu_* \Psi(x_*) = 0$. Das Tripel (x_*, λ_*, μ_*) erfüllt also die KKT-Bedingungen. Es wird also nicht nur das Optimum x_* durch $(x_k)_{k \in \mathbb{N}}$ approximiert, sondern auch die dazugehörigen Lagrange-Multiplikatoren λ_* und μ_* durch die Folgen $(\lambda_k)_{k \in \mathbb{N}}$ und $(\mu_k)_{k \in \mathbb{N}}$. Das heißt, im Prinzip kann man Penalty-Verfahren nutzen, um geeignete Anfangsbedingungen für Newton-artige Verfahren zu generieren (zum Beispiel für das semi-glatte Newton-Verfahren aus [Satz 4.99](#)).

Bemerkung 4.110 (Fehlende Glattheit)

Die bestrafte Zielfunktion F_{α_k} aus [Beispiel 4.107](#) ist zwar noch stetig differenzierbar, aber nicht mehr zweimal stetig differenzierbar. Theoretisch ist daher nicht unbedingt garantiert, dass unrestringierte Optimierungsverfahren mit potentiell superquadratischer Konvergenzrate wie z. B. das BFGS-Verfahren oder das Newton-Verfahren auch tatsächlich schnell konvergieren.¹² In der Praxis ist das BFGS-Verfahren trotzdem oft das Mittel der Wahl, denn wie wir gleich sehen werden, sind Gradientenverfahren ob der schlechten Konditionierung von F_{α} für große α deutlich zu langsam.

Bemerkung 4.111 (Schlechte Konditionierung)

Für gleichungsrestringierte Probleme der Form

$$\begin{array}{ll} \text{Minimiere} & F(x) \\ \text{u. d. N.} & \Phi(x) = 0 \end{array}$$

ist $F_{\alpha}(x) = F(x) + \frac{\alpha}{2} \|\Phi(x)\|_{g_Y}^2$ noch genauso oft differenzierbar wie F und Φ . Sind diese noch zweimal differenzierbar, so können wir auch $D^2 F_{\alpha}(x)$ berechnen: Wir haben

$$DF_{\alpha}(x) u = DF(x) u + \alpha g_Y(\Phi(x), D\Phi(x) u)$$

als erste Ableitung. Nochmaliges Differenzieren ergibt die zweite Ableitung:

$$D^2 F_{\alpha}(x)(u, v) = D^2 F(x)(u, v) + \alpha g_Y(\Phi(x), D^2 \Phi(x)(u, v)) + \alpha g_Y(D\Phi(x) u, D\Phi(x) v).$$

Ist x_k eine Folge wie in [Bemerkung 4.109](#), so haben wir mit $\lambda_k := \alpha_k J_{g_Y}(\Phi(x_k))$

$$D^2 F_{\alpha_k}(x_k) = D^2 F(x_k) + \lambda_k D^2 \Phi(x_k) + \alpha_k D\Phi(x_k)' J_{g_Y} D\Phi(x_k).$$

Wir haben daher Konvergenz

$$D\Phi(x_k)' J_{g_Y} D\Phi(x_k) \xrightarrow{k \rightarrow \infty} D\Phi(x_*)' J_{g_Y} D\Phi(x_*).$$

Ist die Folge $\lambda_k := \alpha_k J_{g_Y}(\Phi(x_k))$ auch konvergent gegen ein λ_* , so konvergiert

$$D^2 F(x_k) + \lambda_k D^2 \Phi(x_k) \xrightarrow{k \rightarrow \infty} D^2 F(x_*) + \lambda_* D^2 \Phi(x_*).$$

¹²Allerdings ist die max-Funktion noch Newton-differenzierbar, weshalb das Newton-Verfahren mit winzigen Modifikation zu einem semi-glatte Newtonverfahren wird.

Ist $D\Phi(x_*)$ surjektiv und nicht trivial, so gibt es ein $u \in X$ und ein $c > 0$ mit $\|D\Phi(x_*)u\| \geq c\|u\|$. Dann gilt aber auch

$$\langle D\Phi(x_*)' J_{g_Y} D\Phi(x_*) u, u \rangle = g_Y(D\Phi(x_*) u, D\Phi(x_*) u) \geq c^2 \|u\|^2$$

und wir erhalten

$$D^2 F_{\alpha_k}(x_k)(u, u) \geq (c^2 \alpha_k - \|D^2 F(x_k) + \lambda_k D^2 \Phi(x_k)\|) \|u\|^2 \xrightarrow{k \rightarrow \infty} \infty.$$

Für $v \in \ker(D\Phi(x_*))$ gilt aber

$$D^2 F_{\alpha_k}(x_k)(v, v) = D^2 F(x_k)(v, v) \xrightarrow{k \rightarrow \infty} D^2 F(x_*)(v, v) \leq \|D^2 F(x_*)\| \|v\|^2.$$

Folglich gilt für die *Konditionszahl* von $D^2 F_{\alpha_k}(x_k)$ bezüglich g_X :

$$\text{cond}(D^2 F_{\alpha_k}(x_k)) = \frac{\sup_{\|u\|=1} |D^2 F_{\alpha_k}(x_k)(u, u)|}{\inf_{\|v\|=1} |D^2 F_{\alpha_k}(x_k)(v, v)|} \xrightarrow{k \rightarrow \infty} \infty.$$

Das heißt, die in [Algorithmus 10](#) zu lösenden Hilfsprobleme für F_{α_k} werden sukzessive schwieriger und fehleranfälliger. Dies ist der wohl gravierendste Nachteil der Penalty-Verfahren. Da bei Gradientenverfahren die Konvergenzgeschwindigkeit wesentlich von der Konditionszahl abhängt, sehen wir hieran auch, dass man diese besser nicht für große Werte von α_k zur Anwendung bringen und stattdessen Newton-artige Verfahren verwenden sollte.

Trotzdem sind Penalty-Verfahren oft eine gute Strategie, sich einem neuen Optimierungsproblem zu nähern: Sie sind einfach zu implementieren und verhalten sich für nicht zu große Strafparameter sehr robust. Dies reicht oft schon, sich einen Überblick zu verschaffen und Startwerte für andere, effizientere Verfahren zu erzeugen. Gegenüber projizierten Gradientenverfahren und anderen Verfahren, die Zulässigkeit für alle Iterierten x_k erzwingen, haben sie den Vorteil, dass sie nicht so leicht oder zumindest nicht gleich zu Anfang in einem lokalen Minimum stecken bleiben, da die Nebenbedingungen aufgeweicht werden.

4.6.1 Übungsaufgaben

Aufgabe 4.112

Implementieren Sie das Penalty-Verfahren aus [Algorithmus 10](#) für Optimierungsprobleme der Form (4.10) mit Straffunktion

$$P(x) := \frac{1}{2} \|\Phi(x)\|_{g_Y}^2 + \frac{1}{2} \|\max(0, \Psi(x))\|_{g_Z}^2.$$

Verwenden Sie dabei das BFGS-Verfahren, um die bestraften Optimierungsprobleme näherungsweise zu lösen. Testen Sie Ihre Implementierung am Optimierungsproblem

$$\begin{aligned} &\text{Minimiere} \quad (x_1 - 1)^2 + (x_2 + 1)^2 + (x_3 - 2)^2 \\ &\text{u. d. N.} \quad \left((x_1^2 + x_2^2)^2 - x_1^2 + x_2^2 \right)^2 + x_3^2 - \frac{1}{20} = 0, \\ &\quad x_1 - \frac{3}{4} \leq 0, \\ &\quad x_2 - \frac{1}{4} \leq 0, \\ &\quad (x_1, x_2, x_3) \in \mathbb{R}^3. \end{aligned}$$

Hinweis: Dies Problem wurde schon in [Aufgabe 4.104](#) verwendet. Sie können also gegebenenfalls Code wiederverwerten.

Aufgabe 4.113

Im Setting von [Aufgabe 3.75](#) suchen wir nun die Ruhelage einer *starren Kette*, bei dem jedes Kettenglied die feste Länge $\|x_{i+1} - x_i\| = \ell$ hat (vgl. auch [Aufgabe 4.93](#)). Lösen Sie dieses Problem mit eine Penalty-Verfahren, das die Streckungsenergie als Straffunktion benutzt.

Man kann auch Kontakt mit dem Erdboden (vgl. [Aufgabe 4.103](#) und [Aufgabe 4.123](#)) durch einen geeigneten Strafterm bestrafen.

Aufgabe 4.114

Seien (X, g_X) und (Z, g_Z) Hilberträume, sei $U \subset X$ eine offene Menge und $\Psi: U \rightarrow Z$ eine stetig differenzierbare Abbildungen. Seien $C \subset X$ eine konvexe, abgeschlossene Menge mit $C \cap U \neq \emptyset$ und $K \subset Z$ ein abgeschlossener, konvexer Kegel. Zeigen Sie, dass

$$P: U \rightarrow [0, \infty[, \quad P(x) := \frac{1}{2} d(x, C)^2 + \frac{1}{2} d(\Psi(x), K)^2$$

eine *stetig differenzierbare* Penalty-Funktion für $M := \{x \in C \cap U \mid \Psi(x) \in K\}$ ist. Hier sei $d(x, C)$ der minimale Abstand zu C .

Hinweis: Verwenden Sie [Proposition 2.31](#). Für die Differenzierbarkeit unterscheiden sie die Fälle $x \in C^\circ$, $x \in X \setminus C$ und $x \in \partial C := C \setminus C^\circ$.

Aufgabe 4.115

Seien (X, g_X) und (Y, g_Y) Hilberträume, sei $Z = \mathbb{R}^m$ mit dem euklidischen Skalarprodukt g_Z , sei $U \subset X$ eine offene Menge, $\Psi: U \rightarrow Z$ eine stetig differenzierbare Abbildungen und sei

$$K := \{z = (z_1, \dots, z_m) \mid \forall j \in \{1, \dots, m\} : z_j \leq 0\}.$$

Zeigen Sie, dass die quadratische Penalty-Funktion

$$P: U \rightarrow [0, \infty[, \quad P(x) := \frac{1}{2} \|\Phi(x)\|^2 + \frac{1}{2} \|\max(0, \Psi(x))\|^2$$

aus [Beispiel 4.107](#) exakt der Penalty-Funktion aus [Aufgabe 4.114](#) entspricht für die abgeschlossene konvexe Menge $\tilde{C} := X$, den abgeschlossenen konvexen Kegel $\tilde{K} := \{0\} \times K$ und für die Abbildung

$$\tilde{\Psi}: U \rightarrow \tilde{Z} := Y \times Z, \quad \tilde{\Psi}(x) := (\Phi(x), \Psi(x)).$$

4.7 Barriere-Verfahren

Sei X ein normierter Raum, $U \subset X$ eine offene Menge, $F: U \rightarrow \mathbb{R}$ eine stetige Funktion und $M \subset X$ eine abgeschlossene Menge mit nichtleerem Inneren und mit $M = \overline{M^\circ}$. Wieder betrachten wir das Optimierungsproblem

$$\begin{array}{ll} \text{Minimiere} & F(x) \\ \text{u. d. N.} & x \in M \cap U. \end{array}$$

Die Idee bei Barriere-Methoden ist ähnlich wie bei Penalty-Methoden, aber hier möchte man Unzulässigkeit verbieten und nahezu unlässige Punkte bestrafen.

Definition 4.116 Eine *Barrierefunktion* oder *barrier function* für M ist eine stetige Funktion $B: M^\circ \rightarrow [0, \infty[$ so dass für jede Folge $(x_k)_{k \in \mathbb{N}}$ in M° und jeden Randpunkt $x \in \partial M := \overline{M} \setminus M^\circ$ gilt:

$$x_k \rightarrow x \Rightarrow B(x_k) \rightarrow \infty.$$

Für einen *Strafparameter* (*penalty parameter*) $\alpha \in [0, \infty[$ definieren wir die *modifizierte Zielfunktion* durch

$$F_\alpha(x) := F(x) + \alpha B(x).$$

Beispiel 4.117 Wir betrachten das Optimierungsproblem (4.18):

$$\begin{array}{ll} \text{Minimiere} & F(x) \\ \text{u. d. N.} & \Psi(x) \in K, \\ & x \in C, \\ & x \in U. \end{array} \tag{4.59}$$

Sei $B_C: C^\circ \rightarrow [0, \infty[$ eine Barrierefunktion für die abgeschlossene, konvexe Menge $C \subset X$ und sei $B_K: K^\circ \rightarrow [0, \infty[$ eine Barrierefunktion für den abgeschlossenen, konvexen Kegel $K \subset Z$. Diese Funktionen kann es freilich nur gegeben, wenn C° und K° nichtleer sind. Dann ist aber $B(x) := B_C(x) + B_K(\Psi(x))$ eine Barrierefunktion für die Menge $M = \{x \in C \mid \Psi(x) \in K\}$, falls M° nichtleer ist und $M^\circ = \{x \in C^\circ \mid \Psi(x) \in K^\circ\}$ gilt.

Beispiel 4.118 Sei $K \subset \mathbb{R}^m$ der abgeschlossene, konvexe Kegel

$$K = \{z \in \mathbb{R}^m \mid z \leq 0\}.$$

Die am meisten verwendeten Barrierefunktionen sind *inverse Barrierefunktion*

$$B_{\text{inv}}(z) := - \sum_{j=1}^m \frac{1}{z_j}$$

und die *logarithmische Barrierefunktion*

$$B_{\log}(z) := - \sum_{j=1}^m \log(-z_j).$$

Zum Beispiel erhält man für logarithmische Barrieren als modifizierte Zielfunktion:

$$F_{\alpha}(x) = F(x) - \alpha \sum_{j=1}^m \log(-\Psi_j(x)).$$

Für Optimierungsverfahren braucht man auch die ersten Ableitungen. Bei der logarithmischen Barrierefunktion verhalten sich diese am Rand des Zulässigkeitsbereiches etwas weniger singulär als bei der inversen Barrierefunktion:

$$DB_{\text{inv}}(z) := (z_1^{-2}, \dots, z_m^{-2}) \quad DB_{\log}(z) := -(z_1^{-1}, \dots, z_m^{-1}).$$

Gleiches gilt für die zweiten Ableitungen:

$$D^2 B_{\text{inv}}(z) := -2 \operatorname{diag}(z_1^{-3}, \dots, z_m^{-3}) \quad D^2 B_{\log}(z) := \operatorname{diag}(z_1^{-2}, \dots, z_m^{-2}).$$

Die erklärt, warum häufig logarithmische Barrierefunktionen vorgezogen werden.

Bei Barriere-Verfahren hofft man, dass globale Minimierer x_{α} von F_{α} gegen globale Minimierer von F konvergieren, wenn man $\alpha \rightarrow 0$ gehen lässt. Die Überlegungen dazu sind denen bei Penalty-Verfahren ganz ähnlich. Ebenfalls lassen sich Approximationen von Lagrange-Multiplikatoron durch Barriere-Verfahren ermitteln. Für logarithmische Barrieren hat man zum Beispiel

$$0 = DF_{\alpha}(x_{\alpha}) = DF(x_{\alpha}) + \alpha DB(x_{\alpha}) = DF(x_{\alpha}) + \sum_{j=1}^m \frac{\alpha}{\Psi_j(x_{\alpha})} D\Psi_j(x_{\alpha}),$$

was auf folgende Approximation führt:

$$\mu_{\alpha} := \left(\frac{\alpha}{\Psi_1(x_{\alpha})}, \dots, \frac{\alpha}{\Psi_m(x_{\alpha})} \right).$$

Für den Startwert muss immer ein zulässiger Punkt gefunden werden, was nicht immer einfach ist. Dafür wird aber garantiert, dass die erzeugten Iterierten ebenfalls stets zulässig sind. Das ist besonders dann attraktiv, wenn die Zielfunktion nicht überall außerhalb des Zulässigkeitsbereichs definiert ist oder dort Singularitäten oder auch nur Unstetigkeiten besitzt. Jedoch ist es bei Barriere-Verfahren etwas schwieriger, geeignete Abbruchkriterien zu formulieren.

Mit Penalty-Verfahren haben Barriere-Verfahren gemein, dass die Konditionszahlen der zweiten Ableitungen beliebig groß werden für hinreichend extreme Wert des Parameters α .

4.7.1 Innere-Punkte-Verfahren für lineare Optimierungsprobleme

Sind im Optimierungsproblem (4.10) auch die Räume X und Y endlichdimensional und sind F , Φ und Ψ affin-linear, so spricht man von einem *linearen Optimierungsproblem*. Dies ist eine überaus wichtige Klasse von Optimierungsproblemen mit vielen Anwendungen, insbesondere in Ökonomie und Logistik.

Das allgemeinste lineare Optimierungsproblem hat folgende Form

$$\begin{array}{ll} \text{Minimiere} & c^\top x + c_0 \\ \text{u. d. N.} & A_G x = b_G, \\ & A_U x \leq b_U, \\ & x \in \mathbb{R}^n. \end{array} \quad (4.60)$$

Dabei seien $X = \mathbb{R}^n$, $Y = \mathbb{R}^k$ endlichdimensional und $Z = \mathbb{R}^m$, $c \in \mathbb{R}^n$, $c_0 \in \mathbb{R}$, $A_G \in L(\mathbb{R}^n; \mathbb{R}^k)$, $b_G \in \mathbb{R}^k$, $A_U \in L(\mathbb{R}^n; \mathbb{R}^m)$, $b_U \in \mathbb{R}^m$. Oft lässt man das $c_0 \in \mathbb{R}$ einfach weg, weil dies nichts an der Lösung ändert.

Definition 4.119 Wir sagen, ein lineares Optimierungsproblem ist in *Schlupfform*, wenn es folgende Gestalt hat:

$$\begin{array}{ll} \text{Minimiere} & c^\top x + c_0 \\ \text{u. d. N.} & A x = b, \\ & x \geq 0, \\ & x \in \mathbb{R}^n. \end{array} \quad (4.61)$$

Dabei können wir gegebenenfalls durch Streichen von Zeilen in A erreichen, dass A immer vollen Rang hat, also eine surjektive lineare Abbildung ist.

Lemma 4.120 (Schlupfform)

Jedes lineare Optimierungsproblem (4.60) kann in Schlupfform (4.61) gebracht werden .

BEWEIS. Zunächst wandeln wir die Ungleichungen durch Einführen von *Schlupfvariablen* (*slack variables*) $s \in \mathbb{R}^m$ um in Gleichungen:

$$\begin{array}{ll} \text{Minimiere} & c^\top x + c_0 \\ \text{u. d. N.} & A_G x = b_G, \\ & A_U x + s = b_U, \\ & s \geq 0, \\ & x \in \mathbb{R}^n, s \in \mathbb{R}^m. \end{array}$$

Um nun Ungleichungsbedingungen für *alle* Variablen zu erhalten, splitten wir x auf in $x = x^+ - x^-$ mit $x^+ \geq 0, x^- \geq 0$. Wir erhalten dann das folgende Optimierungsproblem in Schlupfform:

$$\begin{aligned} \text{Minimiere} \quad & c^\top x^+ - c^\top x^- + c_0 \\ \text{u. d. N.} \quad & A_G x^+ - A_G x^- = b_G, \\ & A_U x^+ - A_U x^- + s = b_U, \\ & x^+ \geq 0, \\ & x^- \geq 0, \\ & s \geq 0, \\ & x^+ \in \mathbb{R}^n, x^- \in \mathbb{R}^n, s \in \mathbb{R}^m. \end{aligned} \quad \square$$

Für ein lineares Problem in Schlupfform lässt sich eine sehr einfache Barrierefunktion finden:

$$B(x) = - \sum_{i=1}^n \log(x_i).$$

Die modifizierte Zielfunktion ist also

$$F_\alpha(x) = c^\top x + c_0 - \alpha \sum_{i=1}^n \log(x_i)$$

und wir erhalten das Optimierungsproblem

$$\begin{aligned} \text{Minimiere} \quad & c^\top x + c_0 - \alpha \sum_{i=1}^n \log(x_i) \\ \text{u. d. N.} \quad & A x = b, \\ & x > 0, \\ & x \in \mathbb{R}^n. \end{aligned} \quad (4.62)$$

Das ist ein Problem der Form (4.16) und wie wir gelernt haben, sind KKT-Paare (x_*, λ_*) dieses Problems gerade die Nullstellen der Abbildung

$$\Theta_\alpha(x, \lambda) := \begin{pmatrix} DF_\alpha(x) + \lambda A \\ A x - b \end{pmatrix}.$$

Als Ableitungen von F_α erhalten wir für $u, v \in \mathbb{R}^n$

$$DF_\alpha(x) u = c^\top u - \alpha \sum_{i=1}^n x_i^{-1} u_i \quad \text{und} \quad D^2 F_\alpha(x) (u, v) = \alpha \sum_{i=1}^n x_i^{-2} u_i v_i.$$

Das heißt, wir haben $DDF_\alpha(x) = \alpha \operatorname{diag}(x_1^{-2}, \dots, x_n^{-2})$. Folglich gilt

$$D\Theta_\alpha(x, \lambda) = \begin{pmatrix} \alpha \operatorname{diag}(x_1^{-2}, \dots, x_n^{-2}) & A' \\ A & 0 \end{pmatrix}.$$

Algorithm 12 Innere-Punkte-Verfahren mit log-Barriere für lineare Optimierungsprobleme

Require: $\gamma \in]0, 1[, s \in]0, 1[, \alpha_0 > 0, \text{TOL}_F > 0, \text{TOL}_B > 0, R_{\max} > 0$.

```

1: function INTERIORPOINTMETHOD( $x_0, \lambda_0, c, A, b, \alpha_0$ )
2:    $k \leftarrow 0$ 
3:   while  $\neg \sum_{i=1}^n (-\log(x_{ki})) \geq 1/\text{TOL}_B$  do                                ▶ äußere Schleife
4:      $(x, \lambda) \leftarrow (x_k, \lambda_k)$                                 ▶ nutze alten Wert als Startwert
5:     while  $\|\Theta_{\alpha_k}(x, \lambda)\| > \alpha_k \text{TOL}_F$  do                    ▶ innere Schleife für Newton-Verfahren
6:       Solve
          
$$\begin{pmatrix} \alpha \text{diag}(x^{-2}) & A \\ A & 0 \end{pmatrix} \begin{pmatrix} \delta x \\ \delta \lambda \end{pmatrix} = - \begin{pmatrix} c^\top - \alpha (x^{-1})^\top + \lambda A \\ A x - b \end{pmatrix}$$

7:        $t_{\max} \leftarrow \min \{ -(x/\delta x)_i \mid (x/\delta x)_i < 0 \}$                 ▶ größte zulässige Schrittweite
8:        $t \leftarrow \min(1, \gamma t_{\max})$ 
9:        $(x, \lambda) \leftarrow (x, \lambda) + t(\delta x, \delta \lambda)$ 
10:    end while
11:     $(x_{k+1}, \lambda_{k+1}) \leftarrow (x, \lambda)$                                 ▶  $\|\Theta_{\alpha_k}(x_{k+1}, \lambda_{k+1})\| \leq \alpha_k \text{TOL}_F$ 
12:     $\alpha_{k+1} \leftarrow s \alpha_k$ 
13:     $k \leftarrow k + 1$ 
14:    if  $\|x_k\| > R_{\max}$  then
15:      return Failed
16:    break
17:    end if
18:  end while
19:   $\mu_k \leftarrow -(\alpha_k/x_k)^\top$ 
20:  return  $(x_k, \lambda_k, \mu_k)$ 
21: end function
    
```

Ist x zulässig, so gilt $D^2 F_\alpha(x)(u, u) = \alpha \sum_{i=1}^n x_i^{-2} u_i^2 \geq \alpha \min(x_1^{-2}, \dots, x_n^{-2}) \|u\|^2$. Hat also A vollen Rang, so ist wegen [Lemma 4.85](#) die lineare Abbildung $D\Theta(x, \lambda)$ für $x \in M^\circ$ aus dem Inneren des Zulässigkeitsbereiches stets invertierbar und das (gedämpfte) Newton-Verfahren

$$(x, \lambda) \mapsto (x, \lambda) - \tau(x, \lambda) D\Theta_\alpha(x, \lambda)^{-1} \Theta_\alpha(x, \lambda)$$

ist anwendbar. Hierbei sollte man eine geeignete Schrittweitenbegrenzung für $\tau(x, \lambda)$ verwenden, damit die Iterierten des Newton-Verfahrens auch im Inneren von M bleiben. Eine algorithmische Umsetzung könnte also so aussehen wie in [Algorithmus 12](#). Man beachte, dass lineare Probleme nicht zwangsläufig Lösungen besitzen und diese im Falle von Existenz in der Regel nicht eindeutig sind. Jede Lösung x_* muss aber auf dem Rand des Zulässigkeitsbereiches liegen, das heißt, es gilt $B(x_*) = \infty$. Das erklärt die etwas seltsame Abbruchbedingungen.

Bemerkung 4.121 Bei linearen Problemen aus der Praxis sind die Matrizen A_G und A_U aus (4.60) in der Regel dünn besetzt. Deshalb ist die Matrix A in (4.61) und $D\Theta_\alpha(x, \lambda)$ in der Regel

ebenfalls dünn besetzt. Dies kann man ausnutzen, indem man zum Berechnen der Newton-Schritte direkte Löser für dünnbesetzte Gleichungssysteme benutzt. Manche direkten Löser¹³ besitzen eine symbolische Faktorisierungsphase, deren Ergebnis nur von der *Besetzungsstruktur* (*sparsity pattern*) der Matrix $D\Theta_\alpha(x, \lambda)$ abhängt. Da die symbolische Faktorisierungsphase in der Regel teurer ist als die sogenannte numerische Faktorisierungsphase, sollte man – wenn man die Möglichkeit dazu hat – die symbolische Faktorisierung für nachfolgende Newton-Iterationen wiederverwerten.

4.7.2 Übungsaufgaben

Aufgabe 4.122

Implementieren Sie das Innere-Punkte-Verfahren mit log-Barriere für lineare Optimierungsprobleme aus [Algorithmus 12](#) und eine Routine, die ein allgemeines lineares Optimierungsproblem (4.60) in Schlupfform (vgl. (4.61)) überführt. Testen Sie Ihre Implementierung für zufällige $c \in \mathbb{R}^3$ an folgendem Optimierungsproblem

$$\begin{array}{ll}\text{Minimiere} & c^\top x \\ \text{u. d. N.} & (1, 1, 1) x = 1, \\ & -x \leq 0, \\ & x \leq (2, 2, 2)^\top, \\ & x \in \mathbb{R}^3.\end{array}$$

Aufgabe 4.123

Im Setting von [Aufgabe 3.75](#) suchen wir nun die Ruhelage einer Gummikette, die durch eine vorgegebene Kontaktebene auf einen Halbraum beschränkt ist (vgl. [Aufgabe 4.103](#)). Lösen Sie dieses Problem numerisch mit einem Barriereverfahren durch Wahl einer geeigneten Barrierefunktion.

¹³Wie zum Beispiel die $LDL^\top$ -Faktorisierung des PARDISO-Solvers. Für PARDISO gibt es auch ein MATLAB-Interface, siehe [9].

4.8 Augmented-Lagrangian-Verfahren

Penalty- und Barriereverfahren haben gemeinsam, dass man die Lösung des eigentlichen Problems nur für einen extremen Grenzwert ($\alpha \rightarrow \infty$ bei Penaltyverfahren, $\alpha \rightarrow 0$ bei Barriereverfahren) erwarten kann. Wir erinnern uns, dass dies zur Folge hatte, dass man dazu immer schlechter konditionierte Optimierungsprobleme lösen muss.

Durch eine unscheinbare Modifikation des Penalty-Verfahrens, kann das Augmented-Lagrangian-Verfahren diese Problematik beheben.

4.8.1 Die Lagrange-Funktion

Definition 4.124 Die *Lagrange-Funktion* (der *Lagrangian*) für das Optimierungsproblem

$$\begin{aligned} &\text{Minimiere} && F(x) \\ &\text{u. d. N.} && \Phi(x) = 0, \\ &&& \Psi(x) \in K, \\ &&& x \in U. \end{aligned} \tag{4.63}$$

ist definiert als

$$\mathcal{L}: U \times Y' \times Z' \rightarrow \mathbb{R}, \quad \mathcal{L}(x, \lambda, \mu) := F(x) + \lambda \Phi(x) + \mu \Psi(x).$$

Wir haben $D\mathcal{L}(x, \lambda, \mu) \in (X \times Y' \times Z')' \cong X' \times Y'' \times Z''$, so dass wir $D\mathcal{L}(x, \lambda, \mu)$ mit den kanonischen Abbildungen ψ_Y und ψ_Z in die Bidualräume wie folgt darstellen können:¹⁴

$$D\mathcal{L}(x, \lambda, \mu) \cong \begin{pmatrix} DF(x) + \lambda D\Phi(x) + \mu D\Psi(x) \\ \psi_Y(\Phi(x)) \\ \psi_Z(\Psi(x)) \end{pmatrix}.$$

Die zweite Ableitung $DD\mathcal{L}(x, \mu)$ ist – wenn sie existiert – eine stetige, lineare Abbildung von $X \times Y' \times Z'$ nach $X' \times Y'' \times Z''$ und kann durch die folgende Blockmatrix dargestellt werden:

$$DD\mathcal{L}(x, \lambda, \mu) \cong \begin{pmatrix} DDF(x) + \lambda DD\Phi(x) + \mu DD\Psi(x) & D\Phi(x)' & D\Psi(x)' \\ \psi_Y D\Phi(x) & 0 & 0 \\ \psi_Z D\Psi(x) & 0 & 0 \end{pmatrix}.$$

Bis auf die Bidualabbildungen sind uns diese Darstellungen schon vertraut; wir haben sie wiederholt bei Newton-Verfahren für restringierte Optimierungsprobleme gefunden.

Ist $x \in M := \{x \in U \mid \Phi(x) = 0, \Psi(x) \in K\}$, $\lambda \in Y'$ und $\mu \in \text{pol}(K)$, so gilt stets

$$\mathcal{L}(x, \lambda, \mu) \leq F(x)$$

mit Gleichheit genau dann, wenn $\mu \Psi(x) = 0$ gilt.

¹⁴Statt den Gleichheitszeichen nutzen wir hier “ $\cong$ ”, da wir streng genommen die Notation ein wenig verbiegen.

4.8.2 Augmented-Lagrangian-Verfahren für Gleichungsnebenbedingungen

In diesem Abschnitt seien (X, g_X) und (Y, g_Y) Hilberträume. Wer mag, darf auch gerne annehmen, dass sie endlich-dimensional sind.

Definition 4.125 Die *erweiterte Lagrange-Funktion* (den *augmented Lagrangian*) für das gleichungsrestringierte Optimierungsproblem

$$\begin{aligned} &\text{Minimiere} && F(x) \\ &\text{u. d. N.} && \Phi(x) = 0, \\ &&& x \in U \end{aligned} \tag{4.64}$$

zum Strafparameter $\alpha > 0$ ist gegeben durch

$$\mathcal{L}_\alpha: U \times Y' \rightarrow \mathbb{R}, \quad \mathcal{L}_\alpha(x, \lambda) := F(x) + \frac{\alpha}{2} \|\Phi(x)\|_{g_Y}^2 + \lambda \Phi(x).$$

Das ist gerade die Lagrange-Funktion zum modifizierten Optimierungsproblem

$$\begin{aligned} &\text{Minimiere} && F(x) + \frac{\alpha}{2} \|\Phi(x)\|_{g_Y}^2 \\ &\text{u. d. N.} && \Phi(x) = 0, \\ &&& x \in U. \end{aligned} \tag{4.65}$$

Satz 4.126 (Augmented Lagrangian mit Orakel)

Seien (X, g_X) und (Y, g_Y) Hilberträume. Sei (x_*, λ_*) ein Punkt, der die hinreichenden Bedingungen zweiter Ordnung aus [Satz 4.79](#) erfüllt. Das heißt, es gilt mit $h := D^2F(x_*) + \lambda_* D^2\Phi(x_*)$:

1. $D\Phi(x_*)$ ist surjektiv (LICQ).
2. (x_*, λ_*) erfüllt die KKT-Bedingungen.
3. Es gibt ein $c_0 > 0$ mit $h(u, u) \geq c_0 \|u\|_{g_X}^2$ für alle $u \in \ker(D\Phi(x_*))$.

Dann gibt es ein $\alpha_* \in [0, \infty[$ mit folgender Eigenschaft: Für alle $\alpha > \alpha_*$ ist x_* ein lokales Minimum der Funktion $x \mapsto \mathcal{L}_\alpha(x, \lambda_*)$.

BEWEIS. Wegen der KKT-Bedingung $DF(x_*) + \lambda_* D\Phi(x_*) = 0$ und der Zulässigkeit $\Phi(x_*) = 0$ ist x_* ein kritischer Punkt von $x \mapsto \mathcal{L}_\alpha(x, \lambda_*)$:

$$\partial_x \mathcal{L}_\alpha(x, \lambda_*) = DF(x_*) + \lambda_* D\Phi(x_*) + \alpha J_{g_Y}(\Phi(x_*)) D\Phi(x_*) = 0.$$

Die zweite Ableitung nach x hat die Form

$$\partial_x^2 \mathcal{L}_\alpha(x_*, \lambda_*) = h + J_{g_Y}(\Phi(x_*)) D^2\Phi(x_*) + \alpha D\Phi(x_*)' J_{g_Y} D\Phi(x_*) = h + \alpha D\Phi(x_*)' J_{g_Y} D\Phi(x_*).$$

Auch dies kommt uns bekannt vor (vgl. [Satz 4.83](#)).

Wir können X orthogonal zerlegen in $X = X_0 \oplus X_1$ mit $X_0 := \ker(D\Phi(x_*))$ und $X_1 := \ker(D\Phi(x_*))^\perp$. Weil die Einschränkung von $D\Phi(x_*)$ auf X_1 ein stetiger Isomorphismus auf Y ist, gibt es ein $c_1 > 0$ mit $\|D\Phi(x_*) u_1\|_{g_Y} \geq \sqrt{c_1} \|u_1\|_{g_X}^2$ für alle $u_1 \in X_1$. Für $u \in X$ sei $u = u_0 + u_1$ die orthogonale Zerlegung mit $u_0 \in X_0$, $u_1 \in X_1$. Mit quadratischer Ergänzung erhalten wir

$$\begin{aligned} \partial_x^2 \mathcal{L}_\alpha(x_*, \lambda_*)(u, u) &= h(u_0, u_0) + 2h(u_0, u_1) + h(u_1, u_1) + \alpha g_Y(D\Phi(x_*) u_1, D\Phi(x_*) u_1) \\ &\geq c_0 \|u_0\|^2 - 2\|h\| \|u_0\| \|u_1\| - \|h\| \|u_1\|^2 + \alpha c_1 \|u_1\|^2 \\ &= \frac{c_0}{2} \|u_0\|^2 + \frac{(c_0 \|u_0\| - 2\|h\| \|u_1\|)^2}{2c_0} + \left(\alpha c_1 - \|h\| - \frac{2}{c_0} \|h\|^2\right) \|u_1\|^2 \\ &\geq \frac{c_0}{2} \|u_0\|^2 + \left(\alpha c_1 - \|h\| - \frac{2}{c_0} \|h\|^2\right) \|u_1\|^2. \end{aligned}$$

Für $\alpha \geq \alpha_* := \frac{c_0^2 + 2c_0 \|h\| + 4\|h\|^2}{2c_0 c_1}$ ergibt dies mit dem Satz des Pythagoras:

$$\partial_x^2 \mathcal{L}_\alpha(x_*, \lambda_*)(u, u) \geq \frac{c_0}{2} (\|u_0\|^2 + \|u_1\|^2) = \frac{c_0}{2} \|u\|^2.$$

Damit sind bei x_* für $\alpha \geq \alpha_*$ die hinreichenden Bedingungen zweiter Ordnung für strikte lokale Minima aus der unrestringierten Optimierung für die Funktion $x \mapsto \mathcal{L}_\alpha(x, \lambda_*)$ erfüllt (vgl. [Satz 2.23](#)). $\square$

In der Regel ist der Lagrange-Multiplikator λ_* nicht im Voraus bekannt. Wie wir aber schon in [Bemerkung 4.109](#) gesehen haben, können wir ihn bei Penalty-Verfahren schätzen. Die Idee beim Augmented-Lagrangian-Verfahren ist, für $x_k \in U$, $\lambda_k \in Y'$ und $\alpha_k > 0$ eine (beinahe) optimale Lösung x_{k+1} von $x \mapsto \mathcal{L}_{\alpha_k}(x, \lambda_k)$ zu finden und anschließend λ_k und α_k geeignet aufzudatieren. Wegen

$$0 \approx \partial_x \mathcal{L}_{\alpha_k}(x_{k+1}, \lambda_k) = DF(x_{k+1}) + (\alpha_k J_{g_Y}(\Phi(x_{k+1})) + \lambda_k) D\Phi(x_{k+1})$$

und

$$0 = DF(x_*) + \lambda_* D\Phi(x_*)$$

verwenden wir $\lambda_{k+1} := \lambda_k + \alpha_k J_{g_Y} \Phi(x_{k+1})$ als neuen Schätzer für den Lagrange-Multiplikator λ_* . Wenn das Residuum der Nebenbedingung nicht kleiner geworden ist, das heißt, wenn $\|\Phi(x_{k+1})\| \geq \|\Phi(x_k)\|$ gilt, so nehmen wir dies als Hinweis darauf, dass α_k noch nicht groß genug ist (d. h., es gilt $\alpha_k < \alpha_*$ ist mit dem α_* aus [Satz 4.126](#)) und wählen $\alpha_{k+1} > \alpha_k$. Andernfalls wählen wir $\alpha_{k+1} := \alpha_k$, um α_{k+1} nicht unnötig groß werden zu lassen und die Konditionierungsprobleme aus [Bemerkung 4.111](#) zu vermeiden. Das Verfahren wird als erfolgreich beendet betrachtet, wenn sowohl $\|DF(x_k) + \lambda_k D\Phi(x_k)\|$ als auch $\|\Phi(x_{k+1})\|$ (beinahe) verschwinden, denn das sind die KKT-Bedingungen für Optimierungsprobleme mit Gleichungsnebenbedingungen erfüllt (vgl. (4.37)). In Pseudocode, der noch genug Freiraum für Parameter-Tuning gibt, könnte dies in etwa wie in [Algorithmus 13](#) aussehen.

Algorithm 13 Augmented-Lagrangian-Verfahren für Gleichungsnebenbedingungen

Require: $\varrho: [0, \infty[\rightarrow]0, 1[$, $R: [0, \infty[\rightarrow]1, \infty[$ $\alpha_0 > 0$, $\text{TOL}_{\text{KKT}} > 0$, $\text{TOL}_{\mathcal{L}} > 0$.

```

1: function AUGMENTEDLAGRANGIANMETHODEqCONSTRAINED( $x_0, \lambda_0, \alpha_0, F, \Phi$ )
2:    $k \leftarrow 0$ 
3:    $r_0 \leftarrow \|\Phi(x_0)\|$ 
4:   while  $\|DF(x_k) + \lambda_k D\Phi(x_k)\|^2 + r_k^2 > \text{TOL}_{\text{KKT}}^2$  do
5:     Find  $x_{k+1}$  with  $\|\partial_x \mathcal{L}_{\alpha_k}(x_{k+1}, \lambda_k)\| \leq r_k \text{TOL}_{\mathcal{L}}$ 
        by using an iterative method for unconstrained
        optimization with  $x_k$  as starting point.
6:      $\lambda_{k+1} \leftarrow \lambda_k + \alpha_k J_{g_X} \Phi(x_{k+1})$ 
7:      $r_{k+1} \leftarrow \|\Phi(x_{k+1})\|$ 
8:     if  $r_{k+1} \leq \varrho(\alpha_k) r_k$  then
9:        $\alpha_{k+1} \leftarrow \alpha_k$ 
10:    else
11:       $\alpha_{k+1} \leftarrow R(\alpha_k) \alpha_k$ 
12:    end if
13:     $k \leftarrow k + 1$ 
14:  end while
15:  return  $(x_k, \lambda_k)$ 
16: end function

```

4.8.3 Quadrierte Schlupfvariablen

Wir haben jetzt schon mehrmals gesehen, dass Optimierungsprobleme mit Gleichungsnebenbedingungen einfacher zu behandeln sind als solche mit Ungleichungsbedingungen. Ist $Z = \mathbb{R}^m$, so kann es unter Umständen sehr nützlich sein, das Optimierungsproblem

$$\begin{aligned}
 &\text{Minimiere} && F(x) \\
 &\text{u. d. N.} && \Phi(x) = 0, \\
 &&& \Psi(x) \leq 0, \\
 &&& x \in U
 \end{aligned} \tag{4.66}$$

durch Einführen von *quadrierten Schlupfvariablen* in das folgende gleichungsrestringierte Optimierungsproblem umzuwandeln:

$$\begin{aligned}
 &\text{Minimiere} && F(x) \\
 &\text{u. d. N.} && \Phi(x) = 0, \\
 &&& \Psi(x) + s^2 = 0, \\
 &&& x \in U, \quad s \in Z,
 \end{aligned} \tag{4.67}$$

wobei $s^2 = (s_1^2, \dots, s_m^2)^\top$ komponentenweise zu interpretieren ist. Dies kann man durch $\tilde{U} := U \times Z$, $\tilde{x} := (x, s)$,

$$\tilde{F}: \tilde{U} \rightarrow \mathbb{R}, \quad \tilde{F}(\tilde{x}) = \tilde{F}(x, s) := F(x) \quad \text{und} \quad \tilde{Y} := Y \times Z, \quad \tilde{\Phi}(x, s) := \begin{pmatrix} \Phi(x) \\ \Psi(x) + s^2 \end{pmatrix}$$

wieder in die gewohnte Form für gleichungsrestringierte Optimierungsprobleme pressen:

$$\begin{aligned} &\text{Minimiere} && \tilde{F}(\tilde{x}) \\ &\text{u. d. N.} && \tilde{\Phi}(\tilde{x}) = 0, \\ &&& \tilde{x} \in \tilde{U}. \end{aligned} \tag{4.68}$$

Diese Umwandlung kommt nicht ohne ihren Preis daher und mache Autoren warnen davor (vgl. z. B. die Bemerkungen nach Corollary 3 in [11]). Dafür gibt es mehrere Gründe:

Zum ersten führt man etliche zusätzliche Variablen ein. Der Flaschenhals beim Lösen von nichtlinearen Gleichungssystemen sind die vielen linearen Gleichungssysteme und diese werden durch das Hinzufügen und Koppeln von bzw. mit neuen Variablen deutlich größer. Da die Rechenkomplexität des Lösens von (generischen, dicht besetzten) linearen Gleichungssystemen superlinear in der Anzahl der Variablen N wächst (in der Regel mit $O(N^3)$), schlägt es sich wesentlich auf die Gesamtperformanz der verwendeten Algorithmen nieder, wenn man die Anzahl der Variablen unbedacht aufbläst. Dies ist insbesondere deshalb so gravierend, weil ungleichungsrestringierte Probleme durchaus ein Vielfaches an Ungleichungen als an Variablen enthalten können.

Zum zweiten ist die wesentliche Regularitätsbedingung für gleichungsrestringierte Optimierung (also die Bedingung, dass $D\tilde{\Phi}(\tilde{x})$ surjektiv ist) nur an solchen zulässigen Punkten (x, s) erfüllt, für die die LICQ (vgl. Definition 4.53) bei x erfüllt ist:

Lemma 4.127 *Sei (x, s) ein zulässiger Punkt des Optimierungsproblems (4.68). Dann sind äquivalent*

1. *Am Punkt (x, s) ist die KRZCQ für (4.68) erfüllt.*
2. *Am Punkt x ist die LICQ für (4.66) erfüllt.*

BEWEIS. Wir haben

$$D\tilde{\Phi}(x, s) = \begin{pmatrix} D\Phi(x) & 0 \\ D\Psi(x) & 2 \operatorname{diag}(s) \end{pmatrix}$$

und durch Zeilen- und Spaltenvertauschungen erhalten wir die dazu konjugierte Blockmatrix

$$\begin{pmatrix} D\Phi(x) & 0 & 0 \\ D\Psi_{\mathcal{A}(x)}(x) & 0 & 0 \\ D\Psi_{\mathcal{I}(x)}(x) & 0 & 2 \operatorname{diag}(s_{\mathcal{I}(x)}) \end{pmatrix}.$$

Man beachte dabei, dass $s_{\mathcal{A}(x)} = 0$ und $s_j \neq 0$ für alle $j \in \mathcal{I}(x)$ gelten muss, wenn $\Psi(x) + s^2 = 0$ erfüllt ist. Das heißt aber, dass $D\tilde{\Phi}(x, s)$ genau dann surjektiv ist, wenn der Block

$$\begin{pmatrix} D\Phi(x) \\ D\Psi_{\mathcal{A}(x)}(x) \end{pmatrix}$$

surjektiv ist; genau das ist aber die LICQ bei x . $\square$

Ist die LICQ beim einem Optimum x_* also nicht erfüllt, so kann das Stabilitätsprobleme in seiner Umgebung, z. B. beim Newton-Verfahren, verursachen.¹⁵

Drittens führt man fast zwangsläufig “künstliche” KKT-Punkte (x_*, s_*) für das Optimierungsproblem (4.67) ein, für die x_* kein KKT-Punkt von (4.66) ist. (Für lokale Minimierer gilt das aber nicht.) Das führt Algorithmen, die vornehmlich auf Abstieg bedacht sind (wie z. B. das projizierte Gradientenverfahren), nicht unbedingt in die Irre; das Newton-Verfahren allerdings schon.

Wir wollen diese letzte Angelegenheit etwas genauer untersuchen. Sei $(x_*, s_*, \lambda_*, \mu_*)$ ein KKT-Tupel für das Optimierungsproblem (4.67). Wegen

$$D\tilde{F}(x, s) = \begin{pmatrix} DF(x) & 0 \end{pmatrix} \quad \text{und} \quad D\tilde{\Phi}(x, s) := \begin{pmatrix} D\Phi(x) & 0 \\ D\Psi(x) & 2 \operatorname{diag}(s) \end{pmatrix}$$

lauten die KKT-Bedingungen wie folgt

$$\begin{aligned} DF(x_*) + \lambda_* D\Phi(x_*) + \mu_* D\Psi(x_*) &= 0, \\ \mu_* \operatorname{diag}(s_*) &= 0, \\ \Phi(x_*) &= 0, \\ \Psi(x_*) + s_*^2 &= 0. \end{aligned} \tag{4.69}$$

Aus $\Psi(x_*) + s_*^2 = 0$ leiten wir ab: $\Psi(x_*) \leq 0$, $s_{*\mathcal{A}(x_*)} = 0$ und $s_{*j} \neq 0$ für alle inaktiven Indizes $j \in \mathcal{I}(x_*)$. Wegen $\mu_* \operatorname{diag}(s_*) = 0$ gilt daher $\mu_{*j} = 0$ für alle inaktiven Indizes $j \in \mathcal{I}(x_*)$. Hieraus folgt direkt

$$\mu_* \Psi(x_*) = \mu_{*\mathcal{A}(x_*)} \Psi_{\mathcal{A}(x_*)}(x_*) + \mu_{*\mathcal{I}(x_*)} \Psi_{\mathcal{I}(x_*)}(x_*) = 0.$$

und wir erhalten

$$\begin{aligned} DF(x_*) + \lambda_* D\Phi(x_*) + \mu_* D\Psi(x_*) &= 0, \\ \Phi(x_*) &= 0, \\ \Psi(x_*) &\leq 0, \\ \mu_{*\mathcal{I}(x_*)} &= 0, \\ \mu_* \Psi(x_*) &= 0. \end{aligned}$$

¹⁵Für so gut wie alle anderen iterativen Verfahren müsste sich dann aber auch die Konvergenzgeschwindigkeit gegen x_* verringern.

An den vollen KKT-Bedingungen für das Problem (4.66) fehlt also nur $\mu_{*\mathcal{A}(x_*)} \geq 0$. Dies kann man auch nicht anders erwarten, denn das Einführen von quadratischen Schlupfvariablen hat die Tendenz, am Rand des Zulässigkeitsbereiches neue KKT-Punkte einzuführen. Der Grund dafür ist, dass die KKT-Bedingungen bei Gleichungsrestringtionen überhaupt nicht in der Lage sind, zwischen lokalen Minima, lokalen Maxima oder sonstigen kritischen Punkten zu unterscheiden: Zum Beispiel lauten die KKT-Bedingungen für das Minimieren und das Maximieren unter Gleichungsrestringtionen genau gleich. Folgendes einfaches Beispiel soll das illustrieren:

Beispiel 4.128 Wir betrachten das Optimierungsproblem:

$$\begin{aligned} &\text{Minimiere } x \\ &\text{u. d. N. } \quad x \leq 0, \\ &\quad \quad \quad x \in \mathbb{R}. \end{aligned} \tag{4.70}$$

Mit einer quadrierten Schlupfvariablen erhalten wir das Optimierungsproblem

$$\begin{aligned} &\text{Minimiere } x \\ &\text{u. d. N. } \quad x + s^2 = 0, \\ &\quad \quad \quad x, s \in \mathbb{R}. \end{aligned} \tag{4.71}$$

und der Punkt $(x_*, s_*, \mu_*) := (0, 0, -1/2d)$ ist hiervon ein KKT-Punkt, obwohl (x_*, μ_*) kein KKT-Punkt des Ausgangsproblems ist.

Lemma 4.129 Zu einem Punkt $(x_*, \lambda_*, \mu_*) \in U \times Y' \times Z'$ betrachten wir folgende Aussagen.

1. (x_*, λ_*, μ_*) erfüllt die KKT-Bedingungen für das Optimierungsproblem (4.66).
2. Es gibt einen Punkt $s_* \in Z$, so dass $(x_*, s_*, \lambda_*, \mu_*)$ die KKT-Bedingungen für das Optimierungsproblem (4.67) erfüllt.

Dann ist [Punkt 2](#) notwendig für [Punkt 1](#) aber nicht hinreichend.

BEWEIS. Zu “[Punkt 2](#) $\Rightarrow$ [Punkt 1](#)”: Sei (x_*, λ_*, μ_*) ein KKT-Tripel für das Optimierungsproblem (4.66). Mit $\mathcal{A}_* := \mathcal{A}(x_*)$ und $\mathcal{I}_* := \mathcal{I}(x_*)$ sind die KKT-Bedingungen gleichbedeutend mit

$$\begin{aligned} DF(x_*) + \lambda_* D\Phi(x_*) + \mu_* D\Psi(x_*) &= 0, \\ \Phi(x_*) &= 0, \\ \Psi_{\mathcal{A}_*}(x_*) &= 0, \\ \Psi_j(x_*) &< 0 \quad \text{für alle } j \in \mathcal{I}_*, \\ \mu_{*\mathcal{I}_*} &= 0, \\ \mu_{*\mathcal{A}_*} &\geq 0. \end{aligned}$$

Mit

$$s_{*j} := \begin{cases} \pm \sqrt{-\Psi_j(x_*)}, & j \in \mathcal{I}_*, \\ 0, & j \in \mathcal{A}_*, \end{cases}$$

gilt $\Psi(x_*) + s_*^2 = 0$ und $\mu_* \text{diag}(s_*) = 0$. Wir erhalten also die KKT-Bedingung aus (4.69) zum Optimierungsproblem (4.67). Im Übrigen hatten wir bei der Wahl von s_* genau $2^{\text{card}(I_*)}$ Möglichkeiten.

Zu “Punkt 1 $\Rightarrow$ Punkt 2”: Dies haben wir schon in Beispiel 4.128 gezeigt. $\square$

Im Prinzip kann man aber quadrierte Schlupfvariablen in Verbindung mit dem Newton-Verfahren anwenden, wenn der Startwert $(x_0, s_0, \lambda_0, \mu_0)$ hinreichend nahe an einer Lösung $(x_*, s_*, \lambda_*, \mu_*)$ liegt, die die Bedingungen aus Satz 4.83 für $\tilde{F}$ und $\tilde{\Phi}$ erfüllt.

Wir wollen hier quadrierte Schlupfvariablen aber zunächst für das Augmented-Lagrangian-Verfahren anwenden.

4.8.4 Augmented-Lagrangian-Verfahren für Ungleichungsnebenbedingungen

Die erweiterte Lagrange-Funktion $\tilde{\mathcal{L}}_\alpha$ für das durch Einführen von quadrierten Schlupfvariablen entstandene Optimierungsproblem (4.67) hat die Form

$$\tilde{\mathcal{L}}_\alpha(x, s, \lambda, \mu) := F(x) + \frac{\alpha}{2} \|\Phi(x)\|^2 + \frac{\alpha}{2} \|\Psi(x) + s^2\|^2 + \lambda \Phi(x) + \mu (\Psi(x) + s^2).$$

Falls (x_*, s_*) ein Minimierer von (4.67) mit Lagrange-Multiplikatoren λ_* und μ_* ist, läuft wegen Satz 4.126 wieder alles darauf hinaus, $(x, s) \mapsto \tilde{\mathcal{L}}_\alpha(x, s, \lambda_*, \mu_*)$ zu minimieren. Der entscheidende Trick ist hier, dass man die Minimierung über die Schlupfvariable s analytisch durchführen und so die Problematik der überschüssigen Variablen umgehen kann.

Lemma 4.130 Für festes $z \in Z = \mathbb{R}^m$ und $\mu \in Z' = \mathbb{R}^m$ mit $\mu \geq 0$ gilt

$$\min_{s \in Z} \left(\frac{\alpha}{2} \|z + s^2\|^2 + \mu (z + s^2) \right) = \frac{1}{2\alpha} \left(\|\max(0, \mu + \alpha z)\|^2 - \|\mu\|^2 \right).$$

BEWEIS. Weil die Normquadrate sich als Quadratsummen darstellen lassen, können wir dies für jede der m Komponenten von s unabhängig voneinander minimieren. Für feste $z \in \mathbb{R}$ und $\mu \in \mathbb{R}$ mit $\mu \geq 0$ betrachten wir daher die Funktion $f: \mathbb{R} \rightarrow \mathbb{R}$

$$f(s) := \frac{\alpha}{2} (z + s^2)^2 + \mu (z + s^2).$$

Die Funktion f ist von vierter Ordnung mit $f(s) \rightarrow \infty$ für $|s| \rightarrow \infty$; nach dem Satz von Weierstrass gibt es also immer ein Minimum. Die notwendige Bedingung erster Ordnung für lokale Minima ist

$$f'(s) = 2s(\alpha(s^2 + z) + \mu) = 0.$$

Wir erhalten also als Kandidaten für Minima:

$$s_0 := 0, \quad s_+ := \frac{\sqrt{-\alpha z - \mu}}{\sqrt{\alpha}} \quad \text{und} \quad s_- := -\frac{\sqrt{-\alpha z - \mu}}{\sqrt{\alpha}}.$$

Algorithm 14 Augmented-Lagrangian-Verfahren für Ungleichungsnebenbedingungen

Require: $\varrho: [0, \infty[\rightarrow]0, 1]$, $R: [0, \infty[\rightarrow]1, \infty[$ $\alpha_0 > 0$, $\text{TOL}_{\text{KKT}} > 0$, $\text{TOL}_{\mathcal{L}} > 0$.

```

1: function AUGMENTEDLAGRANGIANMETHOD( $x_0, \lambda_0, \mu_0, \alpha_0, F, \Phi, \Psi$ )
2:    $k \leftarrow 0$ 
3:    $r_0^2 \leftarrow \|\Phi(x_0)\|^2 + \|\max(0, \Psi(x_0))\|^2$ 
4:   while  $\|DF(x_k) + \lambda_k D\Phi(x_k) + \mu_k D\Psi(x_k)\|^2 + r_k^2 > \text{TOL}_{\text{KKT}}^2$  do
5:     Find  $x_{k+1}$  with  $\|\partial_x \mathcal{L}_{\alpha_k}(x_{k+1}, \lambda_k, \mu_k)\| \leq r_k \text{TOL}_{\mathcal{L}}$ 
       by using an iterative method for unconstrained
       optimization with  $x_k$  as starting point.
6:      $\lambda_{k+1} \leftarrow \lambda_k + \alpha_k J_{g_Y} \Phi(x_{k+1})$ 
7:      $\mu_{k+1} \leftarrow \max(0, \mu_k + \alpha_k J_{g_Z} \Psi(x_{k+1}))$ 
8:      $r_{k+1}^2 \leftarrow \|\Phi(x_{k+1})\|^2 + \|\max(0, \Psi(x_{k+1}))\|^2$ 
9:     if  $r_{k+1} \leq \varrho(\alpha_k) r_k$  then
10:        $\alpha_{k+1} \leftarrow \alpha_k$ 
11:     else
12:        $\alpha_{k+1} \leftarrow R(\alpha_k) \alpha_k$ 
13:     end if
14:      $k \leftarrow k + 1$ 
15:   end while
16:   return  $(x_k, \lambda_k, \mu_k)$ 
17: end function
    
```

Fall 1: $\alpha z + \mu < 0$. Wegen $f''(0) = 2(\alpha z + \mu) < 0$ kann $s = 0$ kein lokales Minimum sein. Der Minimalwert ist also $f(s_+) = f(s_-)$.

Fall 2: $\alpha z + \mu \geq 0$. Im diesem Fall sind die anderen Ausdrücke entweder auch gleich 0 oder gar nicht definiert. Das Minimum liegt also bei $s = 0$ und der Minimalwert ist $f(0)$.

Setzt man das ein und führt die Fallunterscheidung gewissenhaft durch, so erhält man

$$\min_{s \in \mathbb{R}} \left(\frac{\alpha}{2} (z + s^2)^2 + \mu (z + s^2) \right) = \frac{1}{2\alpha} \left(\max(0, \mu + \alpha z)^2 - \mu^2 \right).$$

Aufsummieren der Ergebnisse für die einzelnen Komponenten $s_1, \dots, s_m$ ergibt dann die Behauptung des Lemmas. $\square$

Die *erweiterte Lagrange-Funktion* (der *augmented Lagrangian*) für das allgemeine restringierte Optimierungsproblem(4.66) definiert man daher durch

$$\begin{aligned} \mathcal{L}_\alpha(x, \lambda, \mu) &:= \min_{s \in Z} \tilde{\mathcal{L}}_\alpha(x, s, \lambda, \mu) \\ &= F(x) + \frac{\alpha}{2} \|\Phi(x)\|^2 + \lambda \Phi(x) + \frac{1}{2\alpha} \left(\|\max(0, \mu + \alpha J_{g_Z} \Psi(x))\|^2 - \|\mu\|^2 \right). \end{aligned}$$

Wir übernehmen in groben Zügen die Idee von [Algorithmus 13](#) und erzeugen Folgen $(x_k)_{k \in \mathbb{N}}$, $(\lambda_k)_{k \in \mathbb{N}}$, $(\mu_k)_{k \in \mathbb{N}}$ und $(\alpha_k)_{k \in \mathbb{N}}$. Dabei soll wieder $0 \approx \partial_x \mathcal{L}_{\alpha_k}(x_{k+1}, \lambda_k, \mu_k)$ gelten. Wegen

$$\begin{aligned} \partial_x \mathcal{L}_{\alpha_k}(x_{k+1}, \lambda_k, \mu_k) = & DF(x_{k+1}) + \left(\alpha_k J_{g_Y}(\Phi(x_{k+1})) + \lambda_k \right) D\Phi(x_{k+1}) \\ & + \max(0, \mu_k + \alpha_k J_{g_Z}(\Psi(x_{k+1}))) D\Psi(x_{k+1}) \end{aligned}$$

verwenden wir nun zusätzlich

$$\mu_{k+1} := \max(0, \mu_k + \alpha_k J_{g_Z}(\Psi(x_{k+1})))$$

als Aufdatierungsformel und ergänzen das Residuum für die Nebenbedingungen durch einen geeigneten Beitrag von $\max(0, \Psi(x_{k+1}))$. Als Pseudocode fassen wir dies noch einmal in [Algorithmus 14](#) zusammen.

Literatur

- [1] Michele Benzi, Gene H. Golub und Jörg Liesen. „Numerical solution of saddle point problems“. In: *Acta Numer.* 14 (2005), S. 1–137. doi: [10.1017/S0962492904000212](https://doi.org/10.1017/S0962492904000212).
- [2] Jonathan M. Borwein und Warren B. Moors. „Stability of closedness of convex cones under linear mappings“. In: *J. Convex Anal.* 16.3-4 (2009), S. 699–705.
- [3] O. P. Ferreira und B. F. Svaiter. „Kantorovich’s Theorem on Newton’s Method“. Version 1. In: (25. Sep. 2012). arXiv: [1209.5704](https://arxiv.org/abs/1209.5704).
- [4] Wilfried Gansterer, J. Schneid und C. W. Ueberhuber. *Mathematical properties of equilibrium systems*. Technical Report. 2003. URL: <http://eprints.cs.univie.ac.at/1100/>.
- [5] Michael Hintermüller. „Semismooth Newton Methods and Applications“. In: (2010). URL: http://www.math.uni-hamburg.de/home/hinze/Psfiles/Hintermueller_OWNotes.pdf.
- [6] H. Maurer und J. Zowe. „First and second order necessary and sufficient optimality conditions for infinite-dimensional programming problems“. In: *Math. Programming* 16.1 (1979), S. 98–110. doi: [10.1007/BF01582096](https://doi.org/10.1007/BF01582096).
- [7] *Methode der Kleinsten Quadrate*. URL: https://de.wikipedia.org/wiki/Methode_der_kleinsten_Quadrate#Geschichte.
- [8] Jorge Nocedal und Stephen J. Wright. *Numerical optimization*. Second. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006, S. xxii+664.
- [9] *PARDISO 5.0.0 Solver Project*. URL: <http://www.pardiso-project.org>.
- [10] Boris T. Polyak. *Introduction to optimization*. Translations Series in Mathematics and Engineering. Translated from the Russian, With a foreword by Dimitri P. Bertsekas. Optimization Software, Inc., Publications Division, New York, 1987, S. xxvii+438.
- [11] Stephen M. Robinson. „Stability theory for systems of inequalities. II. Differentiable nonlinear systems“. In: *SIAM J. Numer. Anal.* 13.4 (1976), S. 497–513. doi: [10.1137/0713043](https://doi.org/10.1137/0713043).
- [12] R. Tyrrell Rockafellar und Roger J.-B. Wets. *Variational analysis*. Bd. 317. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Berlin: Springer-Verlag, 1998, S. xiv+733. doi: [10.1007/978-3-642-02431-3](https://doi.org/10.1007/978-3-642-02431-3).
- [13] Dirk Werner. *Funktionalanalysis*. 5. Aufl. Springer-Verlag, Berlin, 2000, S. xii+501.

- [14] Robert Whitley. „An elementary proof of the Eberlein-Šmulian theorem“. In: *Math. Ann.* 172 (1967), S. 116–118. doi: [10.1007/BF01350091](https://doi.org/10.1007/BF01350091). URL: <https://doi.org/10.1007/BF01350091>.
- [15] Eberhard Zeidler. *Nonlinear functional analysis and its applications. I. Fixed-point theorems*, Translated from the German by Peter R. Wadsack. Springer-Verlag, New York, 1986, S. xxi+897. doi: [10.1007/978-1-4612-4838-5](https://doi.org/10.1007/978-1-4612-4838-5). URL: <http://dx.doi.org/10.1007/978-1-4612-4838-5>.
- [16] J. Zowe und S. Kurcyusz. „Regularity and stability for the mathematical programming problem in Banach spaces“. In: *Appl. Math. Optim.* 5.1 (1979), S. 49–62. doi: [10.1007/BF01442543](https://doi.org/10.1007/BF01442543).

Index

- $\mathcal{A}(x)$, 125
- A' , 13
- A^* , 18
- $\mathcal{I}(x)$, 125
- $L(X; Y)$, 12
- $L_c^-(F)$, 15
- $\text{Met}(X)$, 93
- $\mathcal{R}(x, \mu)$, 158
- $\mathcal{S}(x, \mu)$, 158
- X' , 13
- $\text{conv}(A)$, 26
- $\overline{\text{conv}}(A)$, 26
- $\overline{\text{cone}}(A)$, 111
- $\text{cone}(A)$, 111
- $\text{pol}(A)$, 108, 110
- $\text{prepol}(A)$, 112
- $\widetilde{\text{cone}}(B)$, 111
- schwach-*-abgeschlossene, konische Hülle, 111
- Abadie-Bedingung, 107, 128
- Abbildung
 - adjungierte, 18
 - duale, 13
 - kanonische in den Bidualraum, 13
 - Rieszsche, 18
- Abbruchkriterium, 34
- abgeschlossene konische Hülle, 111
- abgeschlossene konvexe Hülle, 26
- Ableitung, 20
 - k -te, 20
 - Fréchet-, 20
 - Gâteaux-, 20
 - zweite, 20
- Abstand, 12
- Abstiegsrichtung, 34
- Adjungierte, 18
- adjungierte Abbildung, 18
- aktive Menge, 125
- aktive Nebenbedingung, 125
- Aktive-Mengen-Strategie, 156
- allgemeiner Abstiegsalgorithmus
 - der unrestringierten Optimierung, 35
- Approximation
 - beste, 27
- Armijo-Bedingung, 73
- Aufdatierungsmethode, 34
- augmented Lagrangian, 178
- Augmented-Lagrangian-Verfahren
 - für Gleichungsnebenbedingungen, 180
 - für Ungleichungsnebenbedingungen, 185
- Ausgleichsproblem, 82
- backtracking line search, 73
- Bairescher Kategoriensatz, 131
- Banachraum, 12
- barrier function, 171
- Barriere-Verfahren, 171
- Barrierefunktion, 171
 - inverse, 171
 - logarithmische, 172
- Bedingung
 - Armijo-, 73
 - Effizienz-, 36
 - Geschwindigkeits-, 36
 - Krümmungs-, 76

- Lipschitz-, 70
- Winkel-, 36
- Zoutendijk-, 92
- Bedingung der linearen Unabhängigkeit, 121
- Bedingungen
 - Karush-Kuhn-Tucker-, 108, 116, 128, 147
 - starke Wolfe-Powell-, 76
 - Wolfe-Powell-, 76
- Bereich
 - zulässiger, 5
- Besetzungsstruktur, 176
- bestrafte Zielfunktion, 164
- BFGS-Verfahren, 97
- Bidualraum, 13
- Broyden-Klasse, 93
- Cauchy-Schrittweite, 69
- Cauchy-Schwarz-Ungleichung, 18
- Cauchyfolge, 12
- CG method, preconditioned, 57
- CG-Verfahren
 - vorkonditioniertes, 57
- Curry-Schrittweite, 69
- DFP-Verfahren, 95
- Dido, 10
- differenzierbar, 20
 - Fréchet-, 20
 - Gâteaux-, 20
 - stetig, 20
- Differenzierbarkeit
 - k -te, 20
- direkte Methode der Variationsrechnung, 15, 17
- duale Abbildung, 13
- Dualraum, 13
- effizient, 36
 - semi-, 36
- Effizienzbedingung, 36
- endlich erzeugt, 115
- Erwartungswert, 83
- erweiterte Lagrange-Funktion, 178
- Euklidische Norm, 14
- exakte Schrittweiten, 70
- folgenkompakt, 15
- Fréchet-Ableitung, 20
- Fréchet-differenzierbar, 20
- Funktion
 - konkave, 31
 - konvexe, 31
 - strikt konkave, 31
 - strikt konvexe, 31
- Gauß-Glockenfunktion, 83
- Gauß-Newton-Verfahren, 85
- Geschwindigkeitsbedingung, 36
- Gradient, 39
- Grenzwert, 12
- Guignard-Bedingung, 108
- Gummikette, 103, 176
- Gâteaux-Ableitung, 20
- Gâteaux-differenzierbar, 20
- Hausdorff-Distanz, 164
- Hilbertraum, 18
- Hülle
 - schwach-*-abgeschlossene, konische, 111
 - abgeschlossene konische, 111
 - abgeschlossene konvexe, 26
 - konische, 111
 - konvexe, 26
- inaktive Menge, 125
- inaktive Nebenbedingung, 125
- inexaktes Newton-Verfahren, 89
- Innere-Punkte-Verfahren, 175
- inneres Produkt, 18
- inverse Barrierefunktion, 171
- Isomorphismus
 - Riesz-, 18

- kanonische Abbildung in den Bidualraum, 13
- Karush-Kuhn-Tucker-Bedingungen, 108, 114, 116, 128, 147
 - für polyhedrische Kegel, 118
- Karush-Kuhn-Tucker-Punkt, 108, 128
- Kegel, 104
 - endlich erzeugter, 115
 - polarer, 108, 110
 - polyhedrischer, 115
 - präpolarer, 112
 - punktierter, 104
 - spitzer, 104
 - stumpfer, 104
- Kette
 - Gummi-, 103, 176
 - starre, 151, 170
- Kettenlinie, 9
- kleinste Quadrate, 83
- koerziv, 41
- Konditionszahl, 42, 169
- konische Hülle, 111
- konkav
 - Funktion, 31
- Konvergenz, 12
 - superlineare, 89
- konvex, 26
 - Funktion, 31
 - strikt, 39
- konvexe Hülle, 26
- Konvexkombination, 26
- Kovarianzmatrix, 83
- KRZCQ, 128
- Krümmungsbedingung, 76
- Kuhhaut, 10
- Kuratowski-Konvergenz, 164
- Kurcysz-Robinson-Zowe-Regularitätsbedingung, 128
- L-BFGS-Verfahren, 99
- Lagrange-Funktion
 - erweiterte, 178
- erweiterte , 185
- Lagrange-Multiplikator, 108, 128
- Lagrange=Funktion, 177
- Lagrangian, 177
 - augmented, 178, 185
- Levenberg-Marquardt-Verfahren, 88
- LICQ, 121, 125
- Likelihood, 9, 83
- linear independence constraint
 - qualification, 121, 125
- lineares Optimierungsproblem, 173
- linearisierter Tangentialkegel, 107, 128
- Lipschitz-Bedingung, 70
- Log-Likelihood, 9, 83
- logarithmische Barrierefunktion, 172
- lokalkonvex, 109
- lokalkonvexe Topologie, 109
- Mangasarian-Fromovitz-Regularitätsbedingung, 141
- Mangasarian=Fromovitz=Regularitätsbedingung, 123
- Maximum-Likelihood-Schätzer, 9, 84
- Menge
 - konvexe, 26
- Menge, aktive, 125
- Menge, inaktive, 125
- method of steepest descent, 41
- Methode
 - der kleinsten Quadrate, 83
 - direkte, der Variationsrechnung, 17
- Methode des steilsten Abstiegs, 41
- metrischer Projektor, 27
- MFCQ, 123, 141
- Minimum
 - globales, 22
 - lokales, 22
 - striktes, 22
- Minkowski-Summe, 26
- Modell, 82
- Modellierung, 7
- modifizierte Zielfunktion, 171

- Mogeln, 158
- monotoner Operator, 33
- Moore-Penrose-Pseudoinverse, 85, 88
- Nebenbedingung
 - aktive, 125
 - inaktive, 125
- Newton-Suchrichtung, 89
- Newton-Verfahren, 62, 147
 - inexaktes, 89
 - semi-glattes, 159
- Newtonsuchrichtung, 68
- normierter Raum, 12
- Operator
 - monotoner, 33
- Operatornorm, 13
- optimale Schrittweite, 69
- Optimalitätsbedingung
 - hinreichende erster Ordnung, 105
 - hinreichende zweiter Ordnung, 23, 106, 143
 - notwendige erster Ordnung, 22, 105
 - notwendige zweiter Ordnung, 22, 142
- Optimierung
 - diskrete, 5
 - finite, 5
 - ganzzahlige, 5
 - glatte, 6
 - infinite, 5
 - kontinuierliche, 5
 - konvexe, 6
 - lineare, 6
 - nicht-glatte, 6
 - quadratische, 6
 - restringierte, 6
 - unrestringierte, 6
- Optimierungsproblem, 5
 - lineares, 173
 - quadratisches, 150, 154
- Parallelogrammgleichung, 18
- penalized objective function, 164
- penalty function, 164
- penalty parameter, 164, 171
- Penalty-Verfahren, 166
- polarer Kegel, 108, 110
- polyhedrisch, 115
- Polynom
 - Taylor, 21
- preconditioner, 57
- primal-dual active set, 158
- Problem
 - des Handlungsreisenden, 7
 - Isoperimetrisches, 10
- Programmierung
 - lineare, 6
 - quadratische, 6
- Projektor
 - metrischer, 27
- präpolaren Kegel, 112
- Pseudoinverse
 - Moore-Penrose-, 85
- Punkt
 - zulässiger, 5
- punktiert, 104
- QP, 156
- quadratic programs, 154
- quadratisches Optimierungsproblem, 150, 154
- Quasi-Newton-Verfahren, 91
- Rasch Modell, 8
- Raum
 - normierter, 12
 - reflexiver, 13
- reflexiv
 - Raum, 13
- Regularitätsbedingung
 - Abadie, 107, 128
 - der linearen Unabhängigkeit, 125
 - Guignard, 108
 - Kurcyusz-Robinson-Zowe, 128
 - Mangasarian-Fromovitz, 123, 141

- Regularitätsbedingung der linearen
 - Unabhängigkeit, 125
- Residualnorm, 35
- Richtung
 - des steilsten Abstiegs, 39
 - des steilsten Anstiegs, 39
- Riemannsche Metrik, 68, 81
- Riesz-Isomorphismus, 18
- Sattelpunktmatrix, 147
- Satz
 - Bairescher Kategoriensatz, 131
 - Darstellungssatz von Fréchet-Riesz, 30
 - des Pythagoras, 18
 - Hahn-Banachscher Trennungssatz, 29
 - Kompaktheitssatz von Riesz, 16
 - vom abgeschlossenen Bild, 151
 - von Baire, 131
 - von Banach-Alaoglu, 16
 - von Browder-Minty, 33
 - von der offenen Abbildung, 132, 133
 - von Eberlein-Šmulian, 17
 - von Hahn-Banach, 130
 - von Heine-Borel, 16
 - von Newton-Kantorovich, 62
 - von Weierstrass, 15
 - über die beste Approximation, 27
 - über implizite Funktionen, 137
 - über metrische Regularität, 137
- Schlupfvariablen, 173
 - quadrierte, 180
- Schlupfform, 173
- Schrittweite
 - Armijo-, 73
 - Cauchy-, 69
 - Curry-, 69
 - exakte, 70
 - optimale, 69
 - starke Wolfe-Powell-, 76
 - Wolfe-Powell-, 76
- Schrittweitenregel, 34
- Schur-Komplement, 151
- schwach- $\ast$ -Topologie, 110
- schwache Topologie, 109
- Sekantengleichung, 93
- semi-effizient, 36
- semi-glattes Newton-Verfahren, 159
- senkrecht, 18
- sequential quadratic programming, 150, 155
- Sherman-Morrison-Formel, 95
- Sherman-Morrison-Woodbury-Formel, 102
- Skalarprodukt, 18
 - Standard-, 18
- slack variables, 173
- sparsity pattern, 176
- spitz, 104
- SQP, 150, 155
- Standard-Skalarprodukt, 18
- starke Wolfe-Powell-Bedingungen, 76
- starre Kette, 151, 170
- stetig differenzierbar, 20
- Strafparameter, 164, 171
- Straffunktion, 164
- strikt konkav
 - Funktion, 31
- strikt konvex
 - Funktion, 31
 - Norm, 39
- stumpf, 104
- Subniveaumenge, 15
- Suchrichtungsstrategie, 34
- Suchstrategie, 36
- superlineare Konvergenz, 89
- Tangentialkegel, 104
 - linearisierter, 107, 128
- Tangentialvektor, 104
- Taylor-Polynom, 21
- Tikhonov-Regularisierung, 87
- Topologie
 - lokalconvexe, 109
 - schwach- $\ast$, 110
 - schwache, 109

INDEX

Transposition, [19](#)

unterhalbstetig, [15](#)

Update-Formeln, [99](#)

Verfahren

Augmented-Lagrangian-, für
Gleichungsnebenbedingungen, [180](#)

Augmented-Lagrangian-, für
Ungleichungsnebenbedingungen,
[185](#)

Barriere-, [171](#)

BFGS, [97](#)

DFP-, [95](#)

Gauß-Newton-, [85](#)

inexaktes Newton-, [89](#)

Innere-Punkte-, [175](#)

L-BFGS, [99](#)

Levenberg-Marquardt-, [88](#)

Newton-, [62](#), [147](#)

Newton-artige, [80](#)

Penalty-, [166](#)

Quasi-Newton-, [91](#)

vollständig, [12](#)

Vorkonditionierer, [57](#)

Winkelbedingung, [36](#)

Wolfe-Powell-Bedingungen, [76](#)
starke, [76](#)

Zielfunktion, [5](#)

bestrafte, [164](#)

modifizierte, [171](#)

Zoutendijk-Bedingung, [92](#)

zulässiger Bereich, [5](#)

zulässiger Punkt, [5](#)