Unrestringierte Optimierungsaufgaben

Jochen Werner

Vorlesung im Sommersemester 2002

Inhaltsverzeichnis

Kapitel 1

Einführung durch Beispiele

Unter einer *unrestringierten Optimierungsaufgabe*[1] versteht man das Problem

$$\text{(P)} \qquad \text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n,$$

wobei die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ gegeben ist und geeignete Eigenschaften besitzt (z. B. kann vorausgesetzt sein, dass f stetig differenzierbar ist und $f(x)$ sowie der Gradient $\nabla f(x)$ für alle $x \in \mathbb{R}^n$ in "analytischer Form" zur Verfügung stehen). Eine *globale Lösung von (P)* ist ein $x^* \in \mathbb{R}^n$ mit der Eigenschaft, dass $f(x^*) \leq f(x)$ für alle $x \in \mathbb{R}^n$, dagegen ist $x^* \in \mathbb{R}^n$ eine *lokale Lösung von (P)*, wenn eine Umgebung U^* von x^* mit $f(x^*) \leq f(x)$ für alle $x \in U^*$ existiert. Wir werden uns i. Allg. mit der näherungsweisen Berechnung einer lokalen Lösung von (P) begnügen müssen (es sei denn, die Zielfunktion ist so beschaffen, dass lokale und globale Lösungen übereinstimmen), da wir Iterationsverfahren betrachten werden, die ihrer Natur nach lokaler Natur sind.

Beispiel: Sei (siehe C. Geiger, C. Kanzow (1999, S. 4)) $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$ gegeben durch

$$f(x) := -x_1^2 x_2 + \frac{1}{4}(2x_1^2 - x_2^2) - \frac{1}{2}(2 - x_1^2 - x_2^2)^2.$$

[1] Wir geben einige Lehrbuchliteratur an:

- J. E. Dennis, R. B. Schnabel (1983) *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs.
- R. Fletcher (1987) *Practical Methods of Optimization. Second Edition*. John Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore.
- C. Geiger, C. Kanzow (1999) *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. Springer, Berlin-Heidelberg-New York.
- C. T. Kelley (1999) *Iterative Methods for Optimization*. SIAM, Philadelphia.
- J. Nocedal, S. J. Wright (1999) *Numerical Optimization*. Springer, Berlin-Heidelberg-New York.
- P. Spelluci (1992) *Numerische Verfahren der nichtlinearen Optimierung*. Birkhäuser, Basel-Boston-Berlin.
- J. Werner (1992) *Numerische Mathematik 2*. Vieweg, Braunschweig-Wiesbaden.

Die Frage ist: Wo hat f (lokale) Minima, wo (lokale) Maxima, wie sieht f aus? Die Benutzung von MATLAB wird eine ganz wichtige Rolle in der Vorlesung spielen und hier ist eine gute Gelegenheit, dieses wunderbare Programmsystem ein erstes Mal anzuwenden. Hier benutzen wir zunächst die Möglichkeiten zur Visualisierung, die MATLAB bietet. In Abbildung 1.1 links geben wir einen Flächenplot von $(x, f(x))$ mit $x \in [-2, 2] \times [-2, 2]$ wieder. Rechts findet man zugehörige Höhenlinien. Den linken

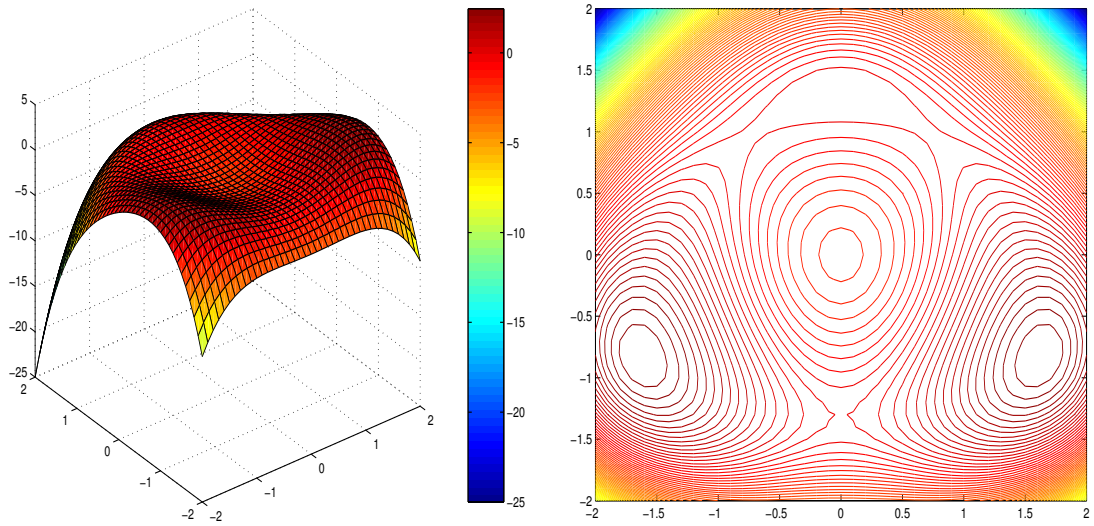

Abbildung 1.1: Flächenplot, Höhenlinienplot

Plot haben wir z. B. erhalten durch

```
x_1=-2:0.1:2;x_2=x_1;
[X_1,X_2]=meshgrid(x_1,x_2);
F=-(X_1.^2).*X_2+0.25*(2*X_1.^2-X_2.^2)-0.5*(2-X_1.^2-X_2.^2).^2;
surf(X_1,X_2,F);colorbar
```

den rechten durch anschließendes

```
contour(X_1,X_2,F,150);
```

Es sieht hier so aus, als wenn f mindestens drei lokale Extrema besitzt. Dies werden wir später nachprüfen. □

Jetzt geben wir noch einige weitere Beispiele von unrestringierten Optimierungsaufgaben an, auf die wir zum Teil später zurückkommen werden.

Beispiel: Die Konzentration[2] $z(t)$ eines Stoffes in einem chemischen Prozess gehorcht dem Gesetz

$$z(t) = a_1 + a_2 e^{\alpha_1 t} + a_3 e^{\alpha_2 t}$$

[2]Dieses Beispiel haben wir aus

H. R. Schwarz (1988) *Numerische Mathematik*. B. G. Teubner, Stuttgart.

mit noch unbekannten Parametern a_1, a_2, a_3 und α_1, α_2. Zur Bestimmung dieser Parameter liefern Messungen zu Zeiten t_i die Messwerte z_i, $i = 1, \ldots, 9$, die gegeben sind durch.

t_i	0.0	0.5	1.0	1.5	2.0	3.0	5.0	8.0	10.0
z_i	3.85	2.95	2.63	2.33	2.24	2.05	1.82	1.80	1.75

Die Parameter sind nach der Methode der kleinsten Quadrate als Lösung von

$$\left\{ \begin{array}{c} \text{Minimiere} \quad f(a_1, a_2, a_3, \alpha_1, \alpha_2) := \displaystyle\sum_{i=1}^{9} [a_1 + a_2 e^{\alpha_1 t_i} + a_3 e^{\alpha_2 t_i} - z_i]^2, \\ (a_1, a_2, a_3, \alpha_1, \alpha_2) \in \mathbb{R}^5 \end{array} \right.$$

zu bestimmen. Dies ist ein spezielles *nichtlineares Ausgleichsproblem* bzw. *nonlinear least squares problem*. Hier ist eine Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ gegeben (n ist hierbei die Anzahl der zu bestimmenden Parameter, m die Anzahl der Beobachtungen, wobei i. Allg. $m \geq n$), gesucht ist eine Lösung der Aufgabe

$$\text{Minimiere} \quad f(x) := \frac{1}{2}\|F(x)\|_2^2 = \frac{1}{2}\sum_{i=1}^{m} F_i(x)^2, \quad x \in \mathbb{R}^n,$$

wobei der Vorfaktor $\frac{1}{2}$ nur kosmetischer Art ist. Es kann sinnvoll sein, die euklidische Norm $\|\cdot\|_2$ durch eine andere Norm zu ersetzen, etwa durch die Maximumnorm $\|\cdot\|_\infty$. Die Aufgabe

$$\text{Minimiere} \quad f(x) := \|F(x)\|_\infty = \max_{i=1,\ldots,m} |F_i(x)|, \quad x \in \mathbb{R}^n$$

nennt man eine *diskrete, nichtlineare Tschebyscheffsche Approximationsaufgabe*. □

Beispiel: Das folgende Problem scheint 1629 zum ersten Mal von Fermat formuliert worden zu sein:

- Gegeben seien drei Punkte in der Ebene. Man finde einen Punkt in der Ebene derart, dass die Summe der Abstände dieses Punktes zu den drei vorgegebenen Punkten minimal ist.

Die Verallgemeinerung auf m Punkte im $\mathbb{R}^n$ heißt das Fermat-Weber-Problem:

- Gegeben seien $m \geq 3$ paarweise verschiedene Punkte $a_1, \ldots, a_m \in \mathbb{R}^n$ und positive reelle Zahlen $w_1, \ldots, w_m$. Man bestimme eine Lösung $x^* \in \mathbb{R}^n$ von

  $$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \sum_{i=1}^{m} w_i \, \|x - a_i\| \quad \text{auf} \quad M := \mathbb{R}^n,$$

  wobei $\|\cdot\|$ in diesem Abschnitt die *euklidische Norm* auf dem $\mathbb{R}^n$ bedeutet.

Verglichen mit anderen Optimierungsaufgaben ist das Fermat-Weber-Problem einfach in der Hinsicht, dass es sich hierbei um eine unrestringierte, konvexe Optimierungsaufgabe handelt. Schwierig ist es es vor allem deshalb, weil die Zielfunktion nicht überall differenzierbar ist.

Die ökonomische Interpretation (man spricht in den Wirtschaftswissenschaften auch von dem "Standortproblem") könnte die folgende sein: Eine Warenhauskette mit Filialen in $a_1, \ldots, a_k$ und Zulieferern in $a_{k+1}, \ldots, a_m$ will den Standort eines zusätzlichen Lagers bestimmen. Dieser soll so gewählt werden, dass eine gewichtete Summe der Abstände vom Lager zu den Filialen und von den Zulieferern zum Lager minimal wird.

Auch hier wollen wir uns die Aufgabenstellung durch Visualisierung der Fläche und der Höhenlinien veranschaulichen. Im ersten Beispiel ist

$$w := (2,4,5)^T, \qquad \begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix} = \begin{pmatrix} 2 & 90 & 43 \\ 42 & 11 & 88 \end{pmatrix}.$$

In Abbildung 1.2 links findet man einen Flächenplot, rechts einen Plot der Höhenlinien. Man erkennt, dass es hier genau ein lokales und globales Minimum gibt. Wesentlich

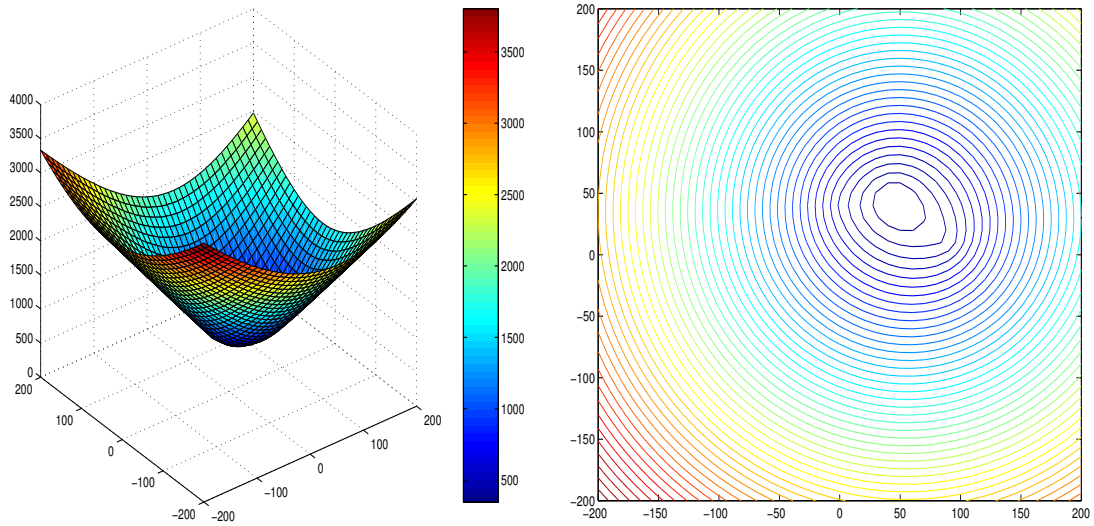

Abbildung 1.2: Flächenplot, Höhenlinienplot beim (konvexen) Fermat-Weber Problem

anders werden die Verhältnisse, wenn auch negative Gewichte auftreten, siehe C. T. Kelley (1999, S. 118 ff.). Im zweiten Beispiel setzen wir

$$w := (2,-4,2,1)^T, \qquad \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \end{pmatrix} := \begin{pmatrix} -10 & 0 & 5 & 25 \\ -10 & 0 & 8 & 30 \end{pmatrix}.$$

Die entsprechenden Plots (links haben wir allerdings die Fläche $(x, -f(x))$ aufgetragen, weil man dann besser sieht, dass man bei negativen Gewichten auch mit lokalen, nicht globalen Lösungen rechnen muss) findet man in Abbildung 1.3. □

Beispiel: Wir folgen jetzt C. T. Kelley (1999, S. 10 ff.) und schildern ein *diskretes optimales Steuerungsproblem*, das auf eine unrestringierte Optimierungsaufgabe führt. Das diskrete optimale Steuerungsproblem entsteht durch Diskretisierung eines *kontinuierlichen optimalen Steuerungsproblems*, daher schildern wir zunächst dieses.

Man stelle sich vor, ein Prozess könne durch Wahl einer Steuerungsfunktion $u(\cdot)$, einer Funktion auf dem Intervall $[0,T]$, beeinflusst werden. Die zugehörige Trajektorie

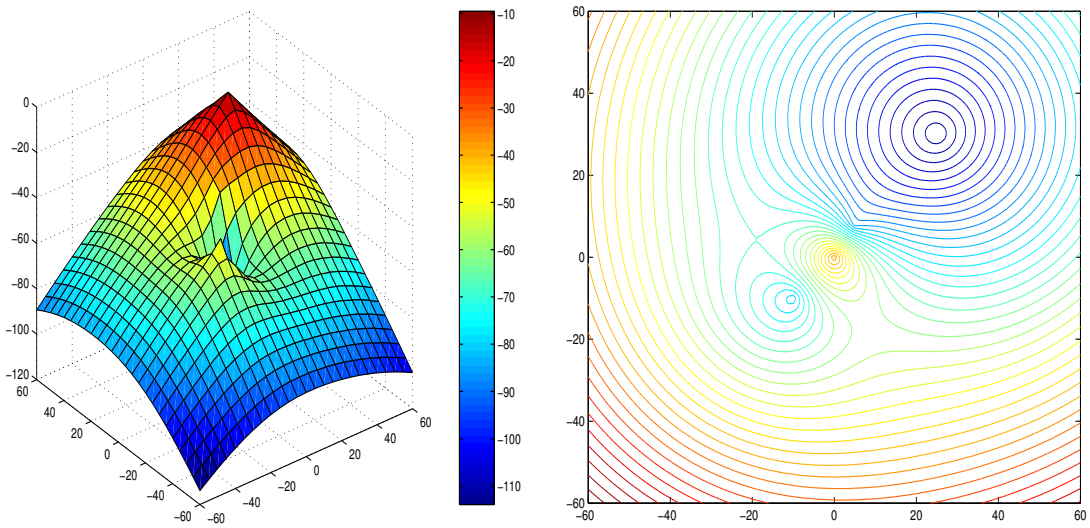

Abbildung 1.3: Plots beim nicht-konvexen Fermat-Weber Problem

$x(\cdot)$ ist die Lösung der gewöhnlichen Anfangswertaufgabe

$$x' = \phi(x, u(t), t), \qquad x(0) = x_0,$$

wobei $x_0 \in \mathbb{R}$ ein von u unabhängiger Anfangszustand ist und die Funktion $\phi\colon \mathbb{R} \times \mathbb{R} \times [0,T] \longrightarrow \mathbb{R}$ so beschaffen sei, dass die obige Anfangswertaufgabe für "vernünftig glattes" u eine eindeutige Lösung $x(\cdot; u)$ besitzt. Zu minimieren sei die Zielfunktion

$$f(u) := \int_0^T F(x(t;u), u(t), t)\, dt.$$

- Wir wollen ein Beispiel im Beispiel angeben. Sei etwa (siehe C. T. Kelley (1999, S. 35)) $T = 1$, $x_0 = 0$ und

  $$F(x,u,t) := (x-3)^2 + \frac{1}{2}u^2, \qquad \phi(x,u,t) := ux + t^2.$$

  Die Lösung von

  $$x' = u(t)x + t^2, \qquad x(0) = 0$$

  ist gegeben durch (es handelt sich hier um eine lineare Differentialgleichung erster Ordnung und eine solche ist im Prinzip exakt lösbar)

  $$x(t;u) = \int_0^t \exp\Big(\int_\tau^t u(s)\, ds\Big)\tau^2\, d\tau,$$

  die eigentliche Zielfunktion ist also

  $$f(u) := \int_0^1 \Big[(x(t;u)-3)^2 + \frac{1}{2}u(t)^2\Big]\, dt.$$

  Ohne Vorkenntnisse wird man Schwierigkeiten haben, die eindeutig existierende Lösung dieser kontinuierlichen (es ist eine Funktion und kein Vektor gesucht) unrestringierten Optimierungsaufgabe zu bestimmen.

Durch Diskretisierung erhält man aus dem kontinuierlichen ein diskretes optimales Steuerungsproblem bzw. eine endlichdimensionale unrestringierte Optimierungsaufgabe. Z. B. kann man die Anfangswertaufgabe mit Hilfe des expliziten Euler-Verfahrens näherungsweise lösen und das Integral in der Zielfunktion mit der zusammengesetzten Trapezformel approximieren. Man wähle also etwa $N \in \mathbb{N}$ mit $N \geq 2$, definiere die Maschenweite $h := T/N$ und $t_j := jh$, $j = 0, \ldots, N$. Zu $u = (u_j)_{j=0,\ldots,N} \in \mathbb{R}^{N+1}$ bestimme man den Vektor $x = (x_j)_{j=0,\ldots,N} \in \mathbb{R}^{N+1}$ aus

$$x_{j+1} := x_j + h\phi(x_j, u_j, t_j), \qquad j = 0, \ldots, N-1,$$

wobei x_0 natürlich der gegebene Anfangszustand ist. Die Zielfunktion ist dann gegeben durch

$$f(u) := h\Big[\frac{1}{2}F(x_0, u_0, 0) + \sum_{j=1}^{N-1} F(x_j, u_j, t_j) + \frac{1}{2}F(x_N, u_N, T)\Big],$$

diese ist auf dem $\mathbb{R}^{N+1}$ zu minimieren. Wir haben hier ein Beispiel kennengelernt, bei welchem die Berechnung des Gradienten der Zielfunktion keineswegs trivial ist. □

Beispiel: Auch dieses Beispiel findet man bei C. T. Kelley (1999, S. 11), es ist ein spezielles *Parameter Identifikations Problem*. Ziel ist es, den Dämpfungsfaktor c und die Federkonstante k in einem schwingenden System nach der Methode der kleinsten Quadrate zu bestimmen, indem man zu Zeiten t_i Auslenkungen u_i, $i = 1, \ldots, m$, misst. Genauer sei $u(\cdot; (c, k))$ bei gegebenem $u_0 \in \mathbb{R}$ die Lösung der Anfangswertaufgabe

$$u'' + cu' + ku = 0, \qquad u(0) = u_0, \quad u'(0) = 0.$$

Zur Identifikation der zu den beobachteten Werten gehörigen Parameter ist die Aufgabe

$$\text{Minimiere} \quad f(c, k) := \frac{1}{2}\sum_{i=1}^{m}[u(t_i; (c, k)] - u_i]^2, \quad (c, k) \in \mathbb{R}^2,$$

zu lösen. Wieder ist keineswegs klar, wie der Gradient von f zu berechnen ist. □

Wir werden in dieser Vorlesung Verfahren zur numerischen Behandlung unrestringierter Optimierungsaufgaben entwickeln und analysieren. Wie üblich in der Numerischen Mathematik kann man nicht hoffen, mit einem Superverfahren alle auftretenden Probleme zu lösen. Vielmehr kommt es auf Eigenschaften der Zielfunktion f an, welche Verfahren erfolgversprechend sind.

Kapitel 2

Grundlagen

2.1 Optimalitätsbedingungen

2.1.1 Notwendige Optimalitätsbedingungen erster Ordnung

Wir betrachten bei vorgegebener Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ die unrestringierte Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n.$$

Unser Ziel in diesem Unterabschnitt ist es, notwendige Bedingungen erster Ordnung dafür anzugeben, dass ein $x^* \in \mathbb{R}^n$ eine lokale Lösung von (P) ist. "Erster Ordnung" bedeutet in diesem Zusammenhang, dass lediglich Ableitungen erster Ordnung von f auftreten.

Definition 1.1 Ist $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ in einem Punkt $x \in \mathbb{R}^n$ stetig partiell differenzierbar, existieren also die partiellen Ableitungen $\partial F_i/\partial x_j$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, aller Komponenten F_i von F in einer Umgebung von x und sind diese in x stetig, so heißt F in x *stetig differenzierbar*. Man nennt

$$F'(x) := \Big(\frac{\partial F_i}{\partial x_j}(x)\Big)_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}} \in \mathbb{R}^{m\times n}$$

die *Funktionalmatrix*[1] von F in x. Ist $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ in $x \in \mathbb{R}^n$ stetig differenzierbar, so heißt

$$\nabla f(x) := \Big(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x)\Big)^T$$

der *Gradient* von f in x. Die Funktion $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ heißt *stetig differenzierbar auf einer offenen Menge* $D \subset \mathbb{R}^n$, wofür wir kürzer $F \in C^1(D;\mathbb{R}^n)$ schreiben werden, wenn F in jedem Punkt $x \in D$ stetig differenzierbar ist. Statt $C^1(D;\mathbb{R})$ schreiben wir kürzer $C^1(D)$.

[1] Im englischen wird oft von "the Jacobian" gesprochen.

Beispiel: Bei nichtlinearen Ausgleichsproblemen nach der Methode der kleinsten Quadrate handelt es sich um die Aufgabe, die Zielfunktion

$$f(x) := \frac{1}{2}\sum_{i=1}^{m} F_i(x)^2$$

zu minimieren, wobei $F_i\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$. Sind alle F_i in x stetig differenzierbar, so auch die Zielfunktion f und es ist

$$\frac{\partial f}{\partial x_j}(x) = \sum_{i=1}^{m} \frac{\partial F_i}{\partial x_j}(x) F_i(x), \qquad j = 1, \ldots, n.$$

Fasst man die F_i als Komponenten einer Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ auf (die Zielfunktion lässt sich dann als $f(x) = \frac{1}{2}\|F(x)\|_2^2$ schreiben), so ist der Gradient von f in x durch

$$\nabla f(x) = F'(x)^T F(x)$$

gegeben. □

Von Zielfunktionen der Form

$$f(x) = \max_{i=1,\ldots,m} F_i(x), \qquad f(x) = \max_{i=1,\ldots,m} |F_i(x)|, \qquad f(x) = \sum_{i=1}^{m} |F_i(x)|$$

kann man nicht erwarten, dass sie stetig differenzierbar sind, selbst dann, wenn es die F_i sind. Daher ist es sinnvoll, auch einen schwächeren Ableitungsbegriff einzuführen.

Definition 1.2 Die Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ heißt in $x \in \mathbb{R}^n$ *in die Richtung* $p \in \mathbb{R}^n$ *richtungsdifferenzierbar*, wenn

$$F'(x; p) := \lim_{t \to 0+} \frac{F(x + tp) - F(x)}{t}$$

existiert. In diesem Fall heißt $F'(x; p)$ die *Richtungsableitung von F in x in Richtung p*. Die Funktion F heißt im Punkte x *richtungsdifferenzierbar*, wenn F in *jede* Richtung $p \in \mathbb{R}^n$ richtungsdifferenzierbar ist. In diesem Falle nennt man die Abbildung $F'(x; \cdot)\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ die *Gateaux-Variation* von F in x.

Natürlich ist eine in einem Punkt x stetig differenzierbare Funktion dort auch richtungsdifferenzierbar. Dies notieren wir in dem folgenden Lemma.

Lemma 1.3 *Ist $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ in $x \in \mathbb{R}^n$ stetig differenzierbar, so ist F in x richtungsdifferenzierbar und die Gateaux-Variation $F'(x; \cdot)\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ ist durch $F'(x; p) = F'(x)p$ gegeben.*

Der folgende Satz liefert eine erste notwendige Optimalitätsbedingung.

Satz 1.4 *Sei $x^* \in \mathbb{R}^n$ eine lokale Lösung von (P). Ist f in x^* richtungsdifferenzierbar, so ist $f'(x^*; p) \geq 0$ für jedes $p \in \mathbb{R}^n$. Ist f in x^* sogar stetig differenzierbar, so ist $\nabla f(x^*) = 0$.*

Beweis: Da x^* eine lokale Lösung von (P) ist, existiert eine Umgebung U^* von x^* mit $f(x^*) \leq f(x)$ für alle $x \in U^*$. Bei vorgegebenem $p \in \mathbb{R}^n$ existiert ein $t_0 > 0$ derart, dass $x^* + tp \in U^*$ und daher $f(x^*) \leq f(x^* + tp)$ für alle $t \in [0, t_0]$. Folglich ist

$$f'(x^*; p) = \lim_{t \to 0+} \frac{f(x^* + tp) - f(x^*)}{t} \geq 0 \qquad \text{für alle } p \in \mathbb{R}^n.$$

Ist f in x^* sogar stetig differenzierbar, so ist $f'(x^*; p) = \nabla f(x^*)^T p \geq 0$ für alle $p \in \mathbb{R}^n$ und (setze z. B. $p := -\nabla f(x^*)$) daher $\nabla f(x^*) = 0$. □ □

Die nächste Definition ist grundlegend.

Definition 1.5 Ist $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ in $x^* \in \mathbb{R}^n$ richtungsdifferenzierbar mit Gateaux-Variation $f'(x^*; \cdot)$ (bzw. stetig differenzierbar mit Gradienten $\nabla f(x^*)$), so heißt x^* ein *stationärer Punkt* von f oder eine *stationäre Lösung* von (P), wenn $f'(x^*; p) \geq 0$ für alle $p \in \mathbb{R}^n$ (bzw. $\nabla f(x^*) = 0$).

Bemerkung: Sei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ in $x \in \mathbb{R}^n$ richtungsdifferenzierbar mit Gateaux-Variation $f'(x; \cdot)$. Ein $p \in \mathbb{R}^n$ heißt eine *Abstiegsrichtung* in x (natürlich bezüglich der Zielfunktion f), wenn $f'(x; p) < 0$. Nach Definition der Richtungsableitung impliziert dies nämlich, dass $f(x + tp) < f(x)$ für alle hinreichend kleinen $t > 0$. Man kann daher sagen, dass ein $x^* \in \mathbb{R}^n$ eine stationäre Lösung von (P) ist, wenn es in x^* keine Abstiegsrichtung gibt. □

Wegen Satz 1.4 ist eine lokale Lösung von (P), in der die Zielfunktion richtungsdifferenzierbar ist, notwendig eine stationäre Lösung.

Beispiel: Eine beliebte Testfunktion bei unrestringierten Optimierungsaufgaben ist die sogenannte *Rosenbrock-Funktion*

$$f(x) := 100(x_2 - x_1^2)^2 + (1 - x_1)^2.$$

In Abbildung 2.1 zeichnen wir einen einen Flächen- und Höhenlinienplot der Rosenbrock-Funktion, wobei wir allerdings den Faktor 100 durch 2 ersetzen (andernfalls würde man nichts erkennen). Man erkennt, dass die Suche nach einem Minimum der Suche nach dem tiefsten Punkt in einem langgestreckten, "bananenförmigen" Tal entspricht. Mit dem Faktor 100 statt 2, ist dies in noch weitaus größerem Maße der Fall. Als Gradienten von f berechnet man

$$\nabla f(x) = \begin{pmatrix} -400x_1(x_2 - x_1^2) - 2(1 - x_1) \\ 200(x_2 - x_1^2) \end{pmatrix}.$$

Der einzige stationäre Punkt von f ist offenbar $x^* = (1, 1)$. Dieser Punkt ist sogar die eindeutige globale Lösung der zugehörigen unrestringierten Optimierungsaufgabe. Dies ist ein besonders glücklicher Umstand und keineswegs die Regel. □

Beispiel: Gegeben sei das nichtlineare Ausgleichsproblem, $f(x) := \frac{1}{2}\|F(x)\|_2^2$ auf dem $\mathbb{R}^n$ zu minimieren, wobei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ stetig differenzierbar sei. Dann ist x^* eine stationäre Lösung des nichtlinearen Ausgleichsproblems, wenn $F'(x^*)^T F(x^*) = 0$.

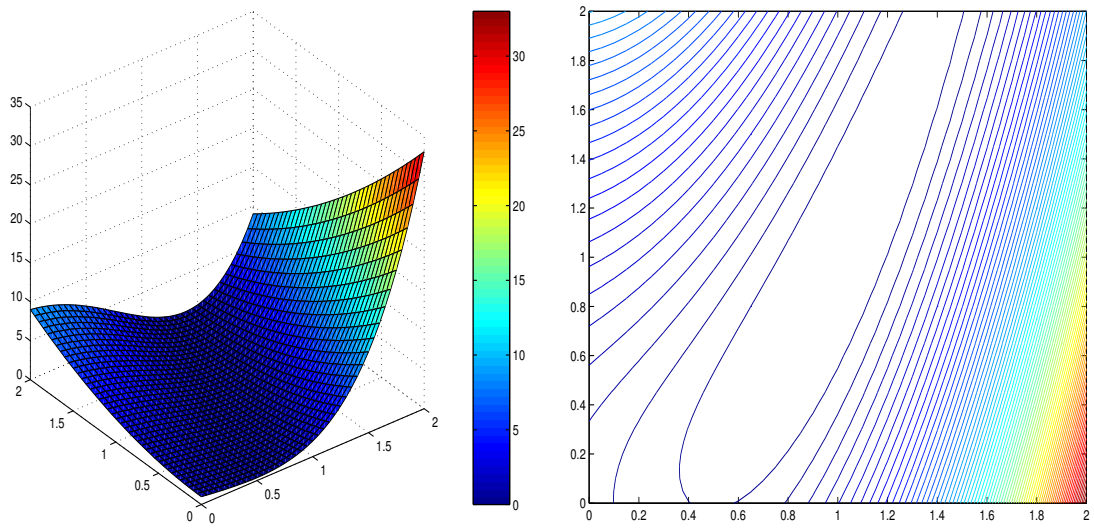

Abbildung 2.1: Flächenplot, Höhenlinienplot zur Rosenbrock-Funktion

Ist $F(x) := Ax - b$ mit $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^n$, so spricht man von einem *linearen Ausgleichsproblem*. Ein x^* ist genau dann zugehörige stationäre Lösung, wenn $A^T(Ax^* - b) = 0$ bzw. x^* den sogenannten *Normalgleichungen* genügt. □

Konvexe Funktione spielen in der (unrestringierten) Optimierung eine wichtige Rolle. Hierbei heißt bekanntlich eine Funktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ *konvex* auf einer konvexen Menge $D \subset \mathbb{R}^n$, wenn

$$f((1-t)x + ty) \le (1-t)f(x) + tf(y) \qquad \text{für alle } x, y \in D \text{ und alle } t \in [0,1].$$

Wir wollen zwar im folgenden den Schwerpunkt auf "glatte" unrestringierte Optimierungsaufgaben legen, aber doch einige Grundlagen für "nichtglatte" (oder besser: "halbglatte") Aufgaben bereitstellen. Interessant in diesem Zusammenhang ist das folgende Lemma.

Lemma 1.6 *Sei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ (auf dem $\mathbb{R}^n$) konvex. Dann ist f in jedem $x \in \mathbb{R}^n$ richtungsdifferenzierbar und es ist*

$$f'(x;p) \le f(x+p) - f(x) \qquad \text{für alle } x, p \in \mathbb{R}^n.$$

Die Gateaux-Variation $f'(x;\cdot)\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ von f in x ist nichtnegativ homogen, d. h.

$$f'(x;\alpha p) = \alpha f'(x;p) \qquad \text{für alle } \alpha \ge 0 \text{ und alle } p \in \mathbb{R}^n,$$

subadditiv, d. h.

$$f'(x;p+q) \le f'(x;p) + f'(x;q) \qquad \text{für alle } p, q \in \mathbb{R}^n,$$

und konvex.

Beweis: Zu $x, p \in \mathbb{R}^n$ definiere man $\phi\colon (0,1] \longrightarrow \mathbb{R}$ durch

$$\phi(t) := \frac{f(x+tp) - f(x)}{t}.$$

Wir zeigen die Existenz von $\lim_{t\to 0+} \phi(t)$ bzw. der Richtungsableitung von f in Richtung p, indem wir nachweisen, dass ϕ auf $(0,1]$ nach unten beschränkt und monoton nicht fallend ist. Wegen der Konvexität von f ist

$$f(x) = f\Big(\frac{1}{1+t}(x+tp) + \frac{t}{1+t}(x-p)\Big) \le \frac{1}{1+t} f(x+tp) + \frac{t}{1+t} f(x-p)$$

für alle $t \in (0,1]$, woraus $f(x) - f(x-p) \le \phi(t)$ für alle $t \in (0,1]$ folgt. Ist ferner $0 < s \le t \le 1$, so ist

$$f(x+sp) - f(x) = f\Big(\frac{s}{t}(x+tp) + \frac{t-s}{t}x\Big) - f(x) \le \frac{s}{t}[f(x+tp) - f(x)]$$

wieder wegen der Konvexität von f, woraus $\phi(s) \le \phi(t)$ folgt. Die Existenz von $f'(x;p) = \lim_{t\to 0+} \phi(t)$ ist damit gesichert. Wegen $\phi(t) \le \phi(1)$ erhalten wir insbesondere $f'(x;p) \le f(x+p) - f(x)$. Der Nachweis der restlichen Behauptungen ist einfach und wird hier übergangen. □ □

Bemerkung: Insbesondere ist bei konvexer Zielfunktion f ein stationärer Punkt von f eine globale Lösung der Aufgabe, f auf dem $\mathbb{R}^n$ zu minimieren. Denn aus

$$0 \le f'(x^*;p) \le f(x^*+p) - f(x^*) \qquad \text{für alle } p \in \mathbb{R}^n$$

folgt (setze $p := x - x^*$) natürlich $f(x^*) \le f(x)$ für alle $x \in \mathbb{R}^n$. Für eine konvexe Zieldunktion fallen also die Begriffe "lokale Lösung", "stationäre Lösung" und "globale Lösung" zusammen. □

Wegen Lemma 1.6 wissen wir, dass konvexe Funktionen eine Gateaux-Variation besitzen. Diese aber im konkreten Fall auszurechnen, kann schwieriger sein. Vektornormen auf dem $\mathbb{R}^n$ sind prominente Vertreter konvexer Funktionen. Im folgenden satz wird die Gateaux-Variation der Maximumnorm berechnet.

Satz 1.7 *Sei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ durch*

$$f(x) := \|x\|_\infty = \max_{j=1,\dots,n} |x_j|$$

definiert. Dann ist f in jedem $x \in \mathbb{R}^n$ richtungsdifferenzierbar, die Gateaux-Variation ist durch

$$f'(x;p) = \begin{cases} \|p\|_\infty, & x = 0, \\ \max\limits_{j\in J(x)} \operatorname{sign}(x_j) p_j, & x \neq 0 \end{cases}$$

mit

$$J(x) := \{j \in \{1,\dots,n\} : |x_j| = \|x\|_\infty\}$$

gegeben.

Beweis: Wir nehmen o. B. d. A. $x \neq 0$ an. Sei eine Richtung $p \in \mathbb{R}^n$ gegeben. Aus Stetigkeitsgründen ist $|x_j + tp_j| < \|x + tp\|_\infty$ für alle $j \notin J(x)$ und alle hinreichend kleinen $t > 0$. Daher ist

$$\frac{f(x+tp) - f(x)}{t} = \max_{j \in J(x)} \frac{|x_j + tp| - |x_j|}{t} = \max_{j \in J(x)} \operatorname{sign}(x_j) p_j$$

für alle hinreichend kleinen $t > 0.$, woraus nach dem Grenzübergang $t \to 0+$ die Behauptung folgt. □ □

Bemerkung: Ähnlich kann die Gateaux-Variation der durch

$$f(x) := \max_{j=1,\ldots,n} x_j, \qquad f(x) := \|x\|_1 = \sum_{j=1}^{n} |x_j|$$

definierten konvexen Funktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ berechnet werden, siehe Aufgabe 2. □

Die unrestringierte Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n,$$

bei der $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ eine i. Allg. nichtlineare glatte Abbildung und $\|\cdot\|$ eine Norm auf dem $\mathbb{R}^n$ ist, nennt man eine *diskrete nichtlineare Approximationsaufgabe*. Wichtigste Spezialfälle sind $\|\cdot\| = \|\cdot\|_2$ (nichtlineares Ausgleichsproblem), $\|\cdot\| = \|\cdot\|_\infty$ (dann spricht man von einer *diskreten Tschebyscheffschen Approximationsaufgabe* und $\|\cdot\| = \|\cdot\|_1$ (*diskrete L_1-Approximationsaufgabe*). I. Allg. wird hier $m > n$ sein, so dass man (P) als die Aufgabe auffassen kann, den Defekt des überbestimmten und daher gewöhnlich nicht lösbaren nichtlinearen Gleichungssystems $F(x) = 0$ bezüglich der Norm $\|\cdot\|$ zu minimieren. Hierdurch und durch die sogenannte (unrestringierte) *Min-Max-Optimierungsaufgabe*, bei der die Zielfunktion f die Form $f(x) := \max_{i=1,\ldots,m} F_i(x)$ hat, sind die wohl wichtigsten "halbglatten" Optimierungsaufgaben gegeben. Wenn die Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ bzw. die Komponentenfunktionen F_i, $i = 1, \ldots, m$, hinreichend glatt sind, kann man immer noch hoffen, dass die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) richtungsdifferenzierbar ist. Diese Vermutung soll im folgenden Satz für die bei der diskreten Tschebyscheff-Approximation auftretende Zielfunktion exemplarisch nachgewiesen werden.

Satz 1.8 *Mit der Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ sei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ durch*

$$f(x) := \|F(x)\|_\infty = \max_{i=1,\ldots,m} |F_i(x)|$$

definiert. Dann gilt: Ist F in $x^* \in \mathbb{R}^n$ stetig differenzierbar, so ist f in x^* richtungsdifferenzierbar mit der Gateaux-Variation

$$f'(x^*; p) = \begin{cases} \max\limits_{i=1,\ldots,m} |\nabla F_i(x^*)^T p|, & F(x^*) = 0, \\ \max\limits_{i \in I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)^T p, & F(x^*) \neq 0, \end{cases}$$

wobei

$$I(x^*) := \{i \in \{1, \ldots, m\} : |F_i(x^*)| = \|F(x^*)\|_\infty\}.$$

Beweis: Mit $g\colon \mathbb{R}^m \longrightarrow \mathbb{R}$, definiert durch

$$g(y) := \|y\|_\infty = \max_{i=1,\ldots,m} |y_i|,$$

ist $f = g \circ F$. Die Abbildung F ist nach Voraussetzung in x^* stetig differenzierbar, während g eine konvexe Funktion ist, deren Gateaux-Variation wir wegen Satz 1.7 kennen. Behauptet wird, dass f in x^* die Gateaux-Variation

$$f'(x^*; p) = g'(F(x^*); F'(x^*)p)$$

besitzt. Hieran erkennt man sehr deutlich, dass der Aussage des Satzes eine *Kettenregel* zugrunde liegt. Wir wollen den Beweis so führen, dass dieser Zusammenhang deutlich wird.

Seien $p \in \mathbb{R}^n$ und eine Nullfolge $\{t_k\} \subset \mathbb{R}_+$ gegeben. Definiert man

$$r_k := F(x^* + t_k p) - F(x^*) - t_k F'(x^*)p,$$

so ist $r_k = o(t_k)$ bzw. $\lim_{k\to\infty} r_k/t_k = 0$, da F als in x^* stetig differenzierbare Funktion insbesondere in x^* (total) differenzierbar ist. Damit erhält man

$$\begin{aligned}
\frac{f(x^* + t_k p) - f(x^*)}{t_k} &= \frac{g \circ F(x^* + t_k p) - g \circ F(x^*)}{t_k} \\
&= \frac{g(F(x^*) + t_k F'(x^*)p + r_k) - g(F(x^*))}{t_k} \\
&= \frac{g(F(x^*) + t_k F'(x^*)p) - g(F(x^*))}{t_k} \\
&\qquad + \frac{g(F(x^*) + t_k F'(x^*)p + r_k) - g(F(x^*) + t_k F'(x^*)p)}{t_k}.
\end{aligned}$$

Da g in $F(x^*)$ eine Gateaux-Variation $g'(F(x^*); \cdot)$ besitzt, konvergiert der erste Summand gegen $g'(F(x^*); F'(x^*)p)$. Greifen wir auf die Definition von g als Maximumnorm auf dem $\mathbb{R}^m$ zurück, so erhalten wir für den Betrag des zweiten Summanden

$$\frac{|\,\|F(x^*) + t_k F'(x^*)p + r_k\|_\infty - \|F(x^*) + t_k F'(x^*)p\|_\infty\,|}{t_k} \leq \frac{\|r_k\|_\infty}{t_k}.$$

Wegen $\lim_{k\to\infty} r_k/t_k = 0$ konvergiert dieser zweite Summand gegen Null. Daher ist $f'(x^*; p) = g'(F(x^*); F'(x^*)p)$ bewiesen, woraus die Behauptung des Satzes folgt. □ □

Bemerkung: Eine etwas genauere Inspektion des Beweise von Satz 1.8 zeigt, dass neben der stetigen Differenzierbarkeit von F die Existenz der Gateaux-Variation und die Lipschitzstetigkeit der durch $g(y) := \max_{i=1,\ldots,m} |y_i|$ definierten Abbildung $g\colon \mathbb{R}^m \longrightarrow \mathbb{R}$ entscheidend für die Existenz der Gateaux-Variation von $f = g \circ F$ in x^* sowie die Gültigkeit der Kettenregel $f'(x^*; p) = g'(F(x^*); F'(x^*)p)$ sind. Analog kann die Existenz der Gateaux-Variation weiterer halbglatter Zielfunktionen $f = g \circ F$ nachgewiesen werden. □

Wir fassen die notwendigen Optimalitätsbedingungen (erster Ordnung) für eine lokale Lösung einer diskreten Tschebyscheffschen Approximationsaufgabe in dem folgenden Satz zusammen.

Satz 1.9 *Gegeben sei die diskrete Tschebyscheffsche Approximationsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) := \|F(x)\|_\infty = \max_{i=1,\dots,m} |F_i(x)|, \quad x \in \mathbb{R}^n.$$

Die Funktionen $F_i\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \dots, m$, seien in $x^* \in \mathbb{R}^n$ stetig differenzierbar und es sei $F(x^*) \neq 0$. Dann gilt: Ist x^* eine lokale Lösung von (P), so ist x^* auch eine stationäre Lösung von (P), d. h. es gilt

$$(*) \qquad f'(x^*; p) = \max_{i \in I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)^T p \geq 0 \qquad \textit{für alle } p \in \mathbb{R}^n,$$

wobei

$$I(x^*) := \{i \in \{1, \dots, m\} : |F_i(x^*)| = \|F(x^*)\|_\infty\}.$$

Ferner ist $(*)$ äquivalent zu der Existenz von reellen Zahlen λ_i^*, $i \in I(x^*)$, mit

$$(**) \quad \lambda_i^* \geq 0 \quad (i \in I(x^*)), \qquad \sum_{i \in I(x^*)} \lambda_i^* = 1, \qquad \sum_{i \in I(x^*)} \lambda_i^* \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*) = 0.$$

Beweis: Eine lokale Lösung von (P) ist notwendig eine stationäre Lösung (Satz 1.4 zusammen mit Definition 1.5). Wegen Satz 1.8 ist für eine lokale Lösung x^* also notwendig $(*)$ erfüllt. Die Äquivalenz von $(*)$ und $(**)$ ist daher die eigentliche Aussage des Satzes.

Zunächst wollen wir die einfache Richtung des Satzes beweisen und nehmen an, $(**)$ würde gelten. Für alle $p \in \mathbb{R}^n$ ist dann

$$0 = \sum_{i \in I(x^*)} \lambda_i^* \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)^T p \leq \max_{i \in I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)^T p,$$

es gilt also $(*)$.

Der Beweis der umgekehrten Richtung ist schwieriger, hier kommt man nicht ohne etwas tiefere Hilfsmittel aus. Jetzt nehmen wir also an, $(*)$ würde gelten bzw. x^* sei eine stationäre Lösung von (P). Zum Beweis benutzen wir das *Farkas-Lemma*[2]:

- *Seien $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$. Dann besitzt das System $Ax = b$, $x \geq 0$ genau dann keine Lösung, wenn das System $A^T y \leq 0$, $b^T y > 0$ eine Lösung besitzt. Hierbei ist die $\leq$- bzw. $\geq$-Beziehung zwischen Vektoren komponentenweise zu verstehen.*

Sei $q := \#(I(x^*))$ die Anzahl der Elemente in der nichtleeren Indexmenge $I(x^*)$ und B die $n \times q$-Matrix, deren Spalten $\operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)$, $i \in I(x^*)$, sind. Ferner sei $e := (1, \dots, 1)^T \in \mathbb{R}^q$. Die Annahme, dass $(**)$ nicht gilt, bedeutet gerade, dass

$$\begin{pmatrix} B \\ e^T \end{pmatrix} \lambda = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \lambda \geq 0$$

[2] Einen elementaren, aber nicht sehr anschaulichen Beweis findet man z. B. bei

J. Werner (1984, S. 37 ff.) *Optimization. Theory and Applications*. Vieweg, Braunschweig-Wiesbaden.

Anschaulichere Beweise basieren auf der Abgeschlossenheit endlich erzeugter Kegel und einem starken Trennungssatz. Siehe aber auch

J. Werner (1992, S. 116) *Numerische Mathematik 2*. Vieweg, Braunschweig-Wiesbaden.

nicht lösbar ist. Das Farkas-Lemma liefert, dass das System

$$B^T p + \gamma e = \begin{pmatrix} B^T & e \end{pmatrix} \begin{pmatrix} p \\ \gamma \end{pmatrix} \leq 0, \qquad \gamma = \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} p \\ \gamma \end{pmatrix} > 0$$

lösbar ist. Die Zeilen von B^T sind $\operatorname{sign}(F_i(x^*))\nabla F_i(x^*)^T$, $i \in I(x^*)$. Daher erhalten wir die Existenz eines $p \in \mathbb{R}^n$ und einer positiven Zahl γ mit $B^T p \leq -\gamma e < 0$ bzw.

$$\operatorname{sign}(F_i(x^*))\nabla F_i(x^*)^T p \leq -\gamma < 0 \qquad \text{für alle } i \in I(x^*),$$

ganz offensichtlich ein Widerspruch zu $(*)$. Der Satz ist damit bewiesen. □ □

Bemerkung: Entsprechend zu Satz 1.9 können auch stationäre Lösungen für die Min-Max-Aufgabe sowie die diskrete L_1-Approximationsaufgabe charakterisiert werden, siehe Aufgaben 3 und 5. □

Beispiel: Wir betrachten[3] das diskrete Tschebyscheffsche Approximationsproblem

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \max_{i=1,2} |F_i(x)|, \quad x \in \mathbb{R}^2,$$

wobei

$$\begin{aligned} F_1(x) &:= x_1 - x_2^3 + 5x_2^2 - 2x_2 - 13, \\ F_2(x) &:= x_1 + x_2^3 + x_2^2 - 14x_2 - 29. \end{aligned}$$

Zunächst geben wir in Abbildung 2.2 einen Flächen- und einen Höhenlinienplot machen, wobei wir uns den (x_1, x_2)-Bereich $[7, 15] \times [-1.3, -9.3]$ beschränken. Letzteren haben wir z. B. durch

```
x_1=7:0.2:15;x_2=-1.3:0.05:-0.3;
[X_1,X_2]=meshgrid(x_1,x_2);
F_1=X_1-X_2.^3+5*X_2.^2-2*X_2-13;
F_2=X_1+X_2.^3+X_2.^2-14*X_2-29;
F=max(abs(F_1),abs(F_2));
contour(X_1,X_2,F,30);
```

erhalten. Man erkennt, dass man bei der numerischen Behandlung dieser Aufgabe mit Schwierigkeiten rechnen muss, da man ähnlich wie bei der Rosenbrock-Funktion in einem langgestreckten Tal nach einem tiefsten Punkt sucht. Für $x^* := (5, 4)$ ist $F(x^*) = 0$, also x^* trivialerweise eine globale Lösung von (P). Uns interessieren weitere lokale Lösungen x^*. Offenbar ist notwendigerweise $I(x^*) = \{1, 2\}$. Daher existieren nichtnegative λ_1^*, λ_2^* mit $\lambda_1^* + \lambda_2^* = 1$ und

$$\lambda_1^* \operatorname{sign}(F_1(x^*))\nabla F_1(x^*) + \lambda_2^* \operatorname{sign}(F_2(x^*))\nabla F_2(x^*) = 0.$$

[3] Siehe

R. Gonin, A. H. Money (1989, S. 125) *Nonlinear L_p-Norm Estimation*. M. Dekker, New York-Basel

und

R. Fletcher, G. A. Watson (1980) First and second order conditions for a class of nondifferentiable optimization problems. Mathematical Programming 18, 291–307.

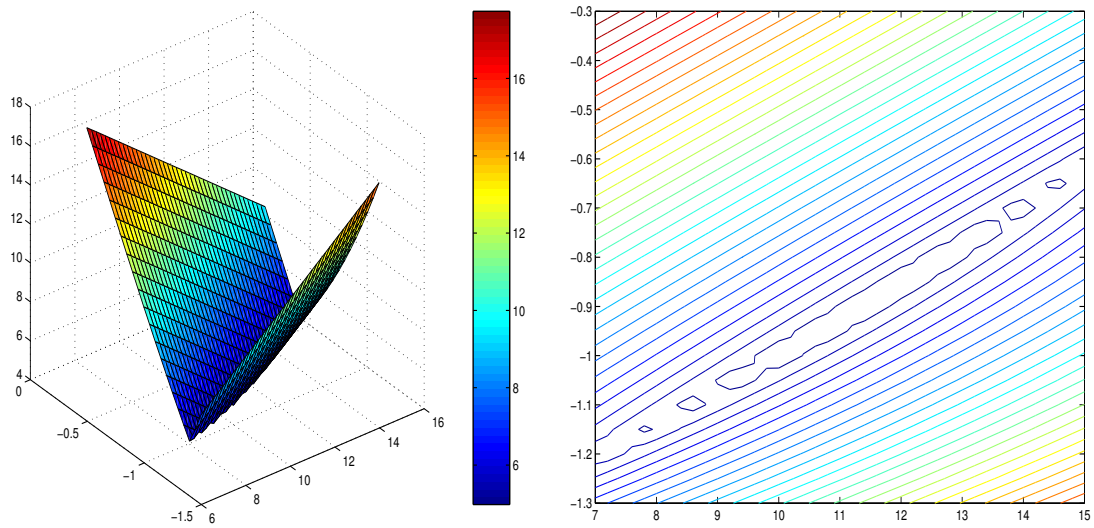

Abbildung 2.2: Flächenplot, Höhenlinienplot bei einer diskreten Tschebyscheffschen Approximationsaufgabe

Betrachtet man in der letzten Gleichung die erste Komponente, so lautet die entsprechende Gleichung

$$\lambda_1^* \operatorname{sign}(F_1(x^*)) + \lambda_2^* \operatorname{sign}(F_2(x^*)) = 0.$$

Zusammen mit $\lambda_1^* + \lambda_2^* = 1$ erkennt man hieran, dass notwendigerweise $\operatorname{sign}(F_1(x^*)) = -\operatorname{sign}(F_2(x^*))$, also o. B. d. A. $\operatorname{sign}(F_1(x^*)) = 1$, $\operatorname{sign}(F_2(x^*)) = -1$. Notwendigerweise ist also $\lambda_1^* = \lambda_2^* = \frac{1}{2}$. Folglich ist $x^* = (x_1^*, x_2^*)$ genau dann eine stationäre Lösung von (P), wenn x^* Lösung des nichtlinearen Gleichungsystems

$$F_1(x) + F_2(x) = 0, \qquad \nabla F_1(x) - \nabla F_2(x) = 0.$$

Dies ist gleichbedeutend mit

$$x_1 + 3x_2^2 - 8x_2 - 21 = 0, \qquad -3x_2^2 + 4x_2 + 6 = 0.$$

Als Lösungen erhält man $x^* = (x_1^*, x_2^*)$ mit

$$x_1^* = 15 + 4\Big(\frac{2}{3} \pm \frac{1}{3}\sqrt{22}\Big), \qquad x_2^* = \frac{2}{3} \pm \frac{1}{3}\sqrt{22}$$

als stationäre Lösungen. Die stationären Lösungen sind also

$$(x_1^*, x_2^*) = (23.92055434643124, 2.23013858660781)$$

und

$$(x_1^*, x_2^*) = (11.41277898690209, -0.89680525327448).$$

Im nächsten Unterabschnitt stellen wir Aussagen bereit, mit deren Hilfe wir entscheiden können, ob es sich bei diesen stationären Lösungen auch wirklich um lokale Lösungen von (P) handelt. □

2.1.2 Notwendige und hinreichende Optimalitätsbedingungen zweiter Ordnung

Nachdem wir uns bisher mit *notwendigen Optimalitätsbedingungen erster Ordnung* (es gehen nur Ableitungen erster Ordnung ein) beschäftigt haben, kommen wir nun zu notwendigen und hinreichenden Optimalitätsbedingungen zweiter Ordnung für lokale Lösungen der unrestringierten Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n.$$

Zunächst betrachten wir den "glatten" Fall und definieren analog zu 1.1:

Definition 1.10 Ist $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ in einem Punkt $x \in \mathbb{R}^n$ zweimal stetig partiell differenzierbar, existieren also alle partiellen Ableitungen $\partial^2 f/\partial x_i \partial x_j$, $i, j = 1, \ldots, n$, in einer Umgebung von x und sind diese in x stetig, so heißt f *in x zweimal stetig differenzierbar*. Man nennt die symmetrische Matrix

$$\nabla^2 f(x) := \Big(\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\Big)_{1 \le i,j \le n} \in \mathbb{R}^{n \times n}$$

die *Hessesche* von f in x. Die Funktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ heißt *zweimal stetig differenzierbar auf der offenen Menge $D \subset \mathbb{R}^n$*, wofür wir kürzer $f \in C^2(D)$ schreiben werden, wenn f in jedem Punkt $x \in D$ zweimal stetig differenzierbar ist.

Beispiel: Für

$$f(x) := \frac{1}{2} \sum_{k=1}^{m} F_k(x)^2 = \frac{1}{2} \|F(x)\|_2^2$$

mit in $x \in \mathbb{R}^n$ zweimal stetig differenzierbaren $F_k\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $k = 1, \ldots, m$, erhält man

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \sum_{k=1}^{m} \frac{\partial F_k}{\partial x_i}(x) \frac{\partial F_k}{\partial x_j}(x) + \sum_{k=1}^{m} \frac{\partial^2 F_k}{\partial x_i \partial x_j}(x) F_k(x),$$

so dass

$$\nabla^2 f(x) = F'(x)^T F'(x) + \sum_{k=1}^{m} F_k(x) \nabla^2 F_k(x)$$

die Hessesche von f ist. □

Aus der Analysis sind die folgenden notwendigen und hinreichenden Optimalitätsbedingungen zweiter Ordnung bekannt.

Satz 1.11 *Die Funktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ sei auf einer offenen Umgebung von $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbar. Dann gilt:*

1. *Ist x^* eine lokale Lösung von (P), so ist $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ ist (symmetrisch und) positiv semidefinit.*
2. *Ist $\nabla f(x^*) = 0$ und ist $\nabla^2 f(x^*)$ positiv definit, so ist x^* eine isolierte, lokale Lösung von (P), d. h. es gibt eine Umgebung U^* von x^* mit $f(x^*) < f(x)$ für alle $x \in U^* \setminus \{x^*\}$.*

Beispiel: Wir kommen auf ein Beispiel aus der Einführung zurück und betrachten die Funktion

$$f(x) := -x_1^2 x_2 + \frac{1}{4}(2x_1^2 - x_2^2) - \frac{1}{2}(2 - x_1^2 - x_2^2)^2.$$

Es ist

$$\nabla f(x) = \begin{pmatrix} -2x_1x_2 + x_1 + 2x_1(2 - x_1^2 - x_2^2) \\ -x_1^2 - \frac{1}{2}x_2 + 2x_2(2 - x_1^2 - x_2^2) \end{pmatrix}.$$

Wir bestimmen zunächst alle stationären Punkte von f und entscheiden dann, welche lokale Minima und welche lokale Maxima von f sind. Offenbar sind $(0,0)$ und $(0, \pm\sqrt{3/2})$ stationäre Lösung. Zur Bestimmung stationärer Punkte, bei denen die erste Komponente nicht verschwindet, hat man das nichtlineare Gleichungssystem

$$\begin{aligned} -2x_2 + 1 + 2(2 - x_1^2 - x_2^2) &= 0, \\ -x_1^2 - \frac{1}{2}x_2 + 2x_2(2 - x_1^2 - x_2^2) &= 0 \end{aligned}$$

zu lösen. Mit Unterstützung von Maple erhalten wir die weiteren stationären Lösungen $(\pm 1/\sqrt{2}, 1)$ und $(\pm\sqrt{95}/6, -5/6)$. Als Hessesche von f berechnet man

$$\nabla^2 f(x) = \begin{pmatrix} 5 - x_2 - 6x_1^2 - 2x_2^2 & -2x_1(1 + 2x_2) \\ -2x_1(1 + 2x_2) & \frac{7}{2} - x_2 - 2x_1^2 - 6x_2^2 \end{pmatrix}.$$

In der folgenden Tabelle geben wir das Resultat unserer Berechnungen an, wobei wir wieder massiv Maple eingesetzt haben:

Stationärer Punkt x^*	Eigenwerte von $\nabla^2 f(x^*)$	Typ
$(0,0)$	$5, \frac{7}{2}$	Lokales Minimum
$(0, \sqrt{\frac{3}{2}})$	$2 - \frac{1}{2}\sqrt{6}, -\frac{11}{2} - \frac{1}{2}\sqrt{6}$	Sattelpunkt
$(0, -\sqrt{\frac{3}{2}})$	$-\frac{11}{2} + \frac{1}{2}\sqrt{6}, 2 + \frac{1}{2}\sqrt{6}$	Sattelpunkt
$(\sqrt{\frac{1}{2}}, 1)$	$-\frac{11}{4} + \frac{1}{4}\sqrt{337}, -\frac{11}{4} - \frac{1}{4}\sqrt{337}$	Sattelpunkt
$(-\sqrt{\frac{1}{2}}, 1)$	$-\frac{11}{4} + \frac{1}{4}\sqrt{337}, -\frac{11}{4} - \frac{1}{4}\sqrt{337}$	Sattelpunkt
$(\frac{1}{6}\sqrt{95}, -\frac{5}{6})$	$-\frac{33}{4} + \frac{1}{36}\sqrt{18849}, -\frac{33}{4} - \frac{1}{36}\sqrt{18849}$	Lokales Maximum
$(-\frac{1}{6}\sqrt{95}, -\frac{5}{6})$	$-\frac{33}{4} + \frac{1}{36}\sqrt{18849}, -\frac{33}{4} - \frac{1}{36}\sqrt{18849}$	Lokales Maximum

Hierbei nennen wir einen stationären Punkt einen Sattelpunkt, wenn die Hessesche in diesem Punkt indefinit ist. □

Vom mathematischen Standpunkt aus interessanter sind notwendige und hinreichende Optimalitätsbedingungen zweiter Ordnung für "nichtglatte" unrestringierte Optimierungsaufgaben. Exemplarisch wollen wir die diskrete Tschebyscheffsche Approximationsaufgabe betrachten und eine hinreichende optimalitätsbedingung zweiter Ordnung im folgenden Satz formulieren und anschließend beweisen.

Satz 1.12 *Gegeben sei die diskrete Tschebyscheffsche Approximationsaufgabe*

(P) *Minimiere* $f(x) := \|F(x)\|_\infty = \max\limits_{i=1,\ldots,m} |F_i(x)|, \quad x \in \mathbb{R}^n.$

Die Funktionen $F_i\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$, seien auf einer offenen Umgebung von $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbar und es sei $F(x^*) \neq 0$. Sei

$$I(x^*) := \{i \in \{1, \ldots, m\} : |F_i(x^*)| = \|F(x^*)\|_\infty\}.$$

Es wird vorausgesetzt, dass reelle Zahlen λ_i^*, $i \in I(x^*)$, existieren mit:

1. *Es ist*

$$\lambda_i^* \geq 0 \quad (i \in I(x^*)), \qquad \sum_{i \in I(x^*)} \lambda_i^* = 1, \qquad \sum_{i \in I(x^*)} \lambda_i^* \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*) = 0,$$

   d. h. in x^* ist die notwendige Optimalitätbedingung erster Ordnung $(**)$ aus Satz 1.9 erfüllt.

2. *Mit*

$$T^* := \{p \in \mathbb{R}^n : \nabla F_i(x^*)^T p = 0 \text{ für alle } i \in I(x^*) \text{ mit } \lambda_i^* > 0\}$$

   ist

$$p^T \Big\{ \sum_{i \in I(x^*)} \lambda_i^* \operatorname{sign}(F_i(x^*)) \nabla^2 F_i(x^*) \Big\} p > 0 \qquad \text{für alle } p \in T^* \setminus \{0\}.$$

Dann ist x^* eine isolierte lokale Lösung von (P), d. h. es gibt eine Umgebung U^* von x^* mit $\|F(x^*)\|_\infty < \|F(x)\|_\infty$ für alle $x \in U^* \setminus \{x^*\}$.

Beweis: Angenommen, die Behauptung sei falsch. Dann gibt es eine gegen x^* konvergierende Folge $\{x_k\}$ mit $x_k \neq x^*$ und $f(x_k) \leq f(x^*)$ für alle k. Es ist $x_k = x^* + t_k p_k$ mit

$$t_k := \|x_k - x^*\|, \qquad p_k := \frac{x_k - x^*}{\|x_k - x^*\|}.$$

Da wir notfalls zu einer Teilfolge übergehen können wir annehmen, dass $\lim_{k\to\infty} p_k = p \neq 0$.

Wegen $x_k = x^* + t_k p + r_k$ mit $r_k := t_k(p_k - p)$ und $\lim_{k\to\infty} r_k/t_k = 0$ kann leicht analog zum Beweis von Satz 1.8 gezeigt werden, dass

$$\lim_{k\to\infty} \frac{f(x^* + t_k p_k) - f(x^*)}{t_k} = f'(x^*; p).$$

Wegen $f(x^* + t_k p_k) \leq f(x^*)$ und $t_k > 0$ ist $f'(x^*; p) \leq 0$, wegen Satz 1.8 ist also

$$f'(x^*; p) = \max_{i \in I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)^T p \leq 0.$$

Aus der ersten Voraussetzung erhalten wir

$$\sum_{i \in I(x^*)} \lambda_i^* \underbrace{\operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)^T p}_{\leq 0} = 0$$

und hieraus $p \in T^*$. Für $i \in I(x^*)$ ist

$$|F_i(x_k)| \leq \|F(x_k)\|_\infty \leq \|F(x^*)\|_\infty = |F_i(x^*)|,$$

ferner ist $\text{sign}\,(F_i(x_k)) = \text{sign}\,(F_i(x^*))$ für alle hinreichend großen k. Für alle $i \in I(x^*)$ und alle hinreichend großen k ist daher

$$\begin{aligned} 0 &\geq |F_i(x_k)| - |F_i(x^*)| \\ &= \text{sign}\,(F_i(x^*))[F_i(x^* + t_k p_k) - F_i(x^*)] \\ &= t_k \text{sign}\,(F_i(x^*))\nabla F_i(x^*)^T p_k + \frac{1}{2} t_k^2 p_k^T \{\text{sign}\,(F_i(x^*))\nabla^2 F_i(z_{ik})\} p_k \end{aligned}$$

mit $z_{ik} = x^* + \theta_{ik} t_k p_k$ und $\theta_{ik} \in (0,1)$. Eine Multiplikation dieser Ungleichung mit λ_i^*, $i \in I(x^*)$, und anschließendes Aufsummieren liefert unter erneuter Benutzung der ersten Voraussetzung, dass

$$0 \geq t_k \underbrace{\Big\{ \sum_{i \in I(x^*)} \lambda_i^* \text{sign}\,(F_i(x^*))\nabla F_i(x^*) \Big\}^T}_{=0} p_k + \frac{1}{2} t_k^2 p_k^T \Big\{ \sum_{i \in I(x^*)} \lambda_i^* \text{sign}\,(F_i(x^*))\nabla^2 F_i(z_{ik}) \Big\} p_k$$

bzw.

$$p_k^T \Big\{ \sum_{i \in I(x^*)} \lambda_i^* \text{sign}\,(F_i(x^*))\nabla^2 F_i(z_{ik}) \Big\} p_k \leq 0$$

für alle hinreichend großen k. Mit $k \to \infty$ folgt wegen $p_k \to p$ und $z_{ik} \to x^*$, dass

$$p^T \Big\{ \sum_{i \in I(x^*)} \lambda_i^* \text{sign}\,(F_i(x^*))\nabla^2 F_i(x^*) \Big\} p \leq 0,$$

was wegen $p^* \in T^* \setminus \{0\}$ ein Widerspruch zur zweiten Voraussetzung ist. □ □

Beispiel: Wir haben in einem früheren Beispiel nachgewiesen, dass $x^* = (x_1^*, x_2^*)$ mit

$$x_1^* := 15 + 4\Big(\frac{2}{3} \pm \frac{1}{3}\sqrt{22}\Big), \qquad x_2^* := \frac{2}{3} \pm \frac{1}{3}\sqrt{22}$$

stationäre Lösungen des diskreten Tschebyscheffschen Approximationsproblems

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \max_{i=1,2} |F_i(x)|, \quad x \in \mathbb{R}^2$$

sind, wobei

$$\begin{aligned} F_1(x) &:= x_1 - x_2^3 + 5x_2^2 - 2x_2 - 13, \\ F_2(x) &:= x_1 + x_2^3 + x_2^2 - 14x_2 - 29. \end{aligned}$$

Die zugehörigen "Multiplikatoren" sind $\lambda_1^* = \lambda_2^* = \frac{1}{2}$. Wir betrachten zunächst $x^* = (\frac{1}{3}(53 + 4\sqrt{22}), \frac{1}{3}(2 + \sqrt{22}))$. Es ist

$$\nabla F_1(x^*) = \nabla F_2(x^*) = \begin{pmatrix} 1 \\ -4 + 2\sqrt{22} \end{pmatrix}$$

und daher

$$T^* = \Big\{ \alpha \begin{pmatrix} 4 - 2\sqrt{22} \\ 1 \end{pmatrix} : \alpha \in \mathbb{R} \Big\}.$$

Weiter ist

$$\lambda_1^* \text{sign}\,(F_1(x^*))\nabla^2 F_1(x^*) + \lambda_2^* \text{sign}\,(F_2(x^*))\nabla^2 F_2(x^*) = \begin{pmatrix} 0 & 0 \\ 0 & -2\sqrt{22} \end{pmatrix},$$

diese Matrix ist negativ definit auf T^* und daher ist in x^* die hinreichende Bedingung zweiter Ordnung für ein lokales Minimum von f nicht erfüllt. Nun betrachten wir die andere stationäre Lösung $x^* = (\frac{1}{3}(53 - 4\sqrt{22}), \frac{1}{3}(2 - \sqrt{22})$. Diesmal ist

$$\nabla F_1(x^*) = \nabla F_2(x^*) = \begin{pmatrix} 1 \\ -4 - 2\sqrt{22} \end{pmatrix}$$

und daher

$$T^* = \Big\{\alpha \begin{pmatrix} 4 + 2\sqrt{22} \\ 1 \end{pmatrix} : \alpha \in \mathbb{R}\Big\}.$$

Weiter ist

$$\lambda_1^* \text{sign}\,(F_1(x^*))\nabla^2 F_1(x^*) + \lambda_2^* \text{sign}\,(F_2(x^*))\nabla^2 F_2(x^*) = \begin{pmatrix} 0 & 0 \\ 0 & 2\sqrt{22} \end{pmatrix},$$

diese Matrix ist positiv definit auf T^* und daher liegt bei x^* ein lokales Minimum von f. □

2.1.3 Aufgaben

1. Die Funktion $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$ sei definiert durch[4]

   $$f(x) := (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2.$$

   Für $(x_1, x_2) \in [-5, 5] \times [-5, 5]$ gebe man einen Flächen- und einen Höhenlinienplot an. Anschließend berechne man wenigstens einen stationären Punkt von f.

2. Man berechne die Gateaux-Variation der durch

   $$f(x) := \max_{j=1,\ldots,n} x_j, \qquad f(x) := \|x\|_1 = \sum_{j=1}^{n} |x_j|$$

   definierten konvexen Funktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$.

3. Ist $f(x) := \max_{i=1,\ldots,m} F_i(x)$ mit in $x^* \in \mathbb{R}^n$ stetig differenzierbaren $F_i\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$, so ist x^* genau dann stationärer Punkt von f bzw.

   $$(*) \qquad f'(x^*; p) = \max_{i \in I(x^*)} \nabla F_i(x^*)^T p \geq 0 \qquad \text{für alle } p \in \mathbb{R}^n,$$

   wobei

   $$I(x^*) := \{i \in \{1, \ldots, m\} : F_i(x^*) = f(x^*)\},$$

   wenn reelle Zahlen λ_i^*, $i \in I(x^*)$, existieren mit

   $$(**) \qquad \lambda_i^* \geq 0 \quad (i \in I(x^*)), \qquad \sum_{i \in I(x^*)} \lambda_i^* = 1, \qquad \sum_{i \in I(x^*)} \lambda_i^* \nabla F_i(x^*) = 0.$$

[4] Siehe

D. M. Himmelblau (1972) *Applied Nonlinear Programming*. McGraw-Hill, New York.

4. Gegeben[5] sei die Min-Max-Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \max_{t=1,2,3} F_i(x), \quad x \in \mathbb{R}^2,$$

wobei

$$F_1(x) := x_1^4 + x_2^2, \qquad F_2(x) := (2 - x_1)^2 + (2 - x_2)^2, \qquad F_3(x) := 2\exp(-x_1 + x_2).$$

Man zeige, dass $x^* = (1, 1)$ eine stationäre Lösung von (P) ist. Ferner mache man einen Flächen- und einen Höhenlinienplot auf $[0, 2] \times [0, 2]$.

5. Ist $f(x) := \sum_{i=1}^m |F_i(x)|$ mit in $x^* \in \mathbb{R}^n$ stetig differenzierbaren $F_i: \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$, so ist x^* genau dann stationärer Punkt von f bzw.

$$(*) \qquad \begin{cases} f'(x^*; p) = \displaystyle\sum_{i \in I(x^*)} |\nabla F_i(x^*)^T p| + \sum_{i \notin I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)^T p \geq 0 \\ \qquad\qquad \text{für alle } p \in \mathbb{R}^n, \end{cases}$$

wobei

$$I(x^*) := \{i \in \{1, \ldots, m\} : F_i(x^*) = 0\},$$

wenn reelle Zahlen λ_i^*, $i \in I(x^*)$, existieren mit

$$(**) \quad \lambda_i^* \in [-1, 1] \quad (i \in I(x^*)), \qquad \sum_{i \in I(x^*)} \lambda_i^* \nabla F_i(x^*) + \sum_{i \notin I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*) = 0.$$

6. Mit Hilfe von Aufgabe 5 zeige man: Ist x_2^* die reelle Lösung von $x_2^3 + x_2 - 9 = 0$ und $x_1^* := \sqrt{10 - x_2^*}$, so ist $x^* = (x_1^*, x_2^*)$ eine stationäre Lösung der Aufgabe[6]

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \sum_{i=1}^3 |F_i(x)|, \quad x \in \mathbb{R}^2,$$

wobei

$$F_1(x) := x_1^2 + x_2 - 10, \qquad F_2(x) = x_1 + x_2^2 - 7, \qquad F_3(x) := x_1^2 - x_2^3 - 1.$$

Auf $[0, 5] \times [0, 4]$ mache man einen Höhenlinienplot.

7. Man bestimme alle stationären Punkte der durch

$$f(x) := x_1^2 x_2^2 + (x_2^2 - 1)^2$$

definierten Funktion $f: \mathbb{R}^2 \longrightarrow \mathbb{R}$. Welche dieser stationären Punkte sind lokale Minima, welche lokale Maxima?

8. Zeige[7], dass die durch $f(x) := 8x_1 + 12x_2 + x_1^2 - 2x_2^2$ definierte Funktion nur einen stationären Punkt besitzt, der weder ein lokales Minimum noch ein lokales Maximum ist. Man mache einen Höhenlinienplot von f in der Nähe des stationären Punktes.

---

[5] Siehe

R. Gonin, A. H. Money (1989, S. 127) *Nonlinear L_p-Norm Estimation*. M. Dekker, New York-Basel.

[6] Dieses Beispiel stammt aus R. Gonin, A. H. Money (1989, S. 49).

[7] Diese Aufgabe haben wir J. Nocedal, S. J. Wright (1999, S. 30) entnommen.

9. Gegeben sei die unrestringierte Min-Max-Optimierungsaufgabe

   (P) $$\text{Minimiere} \quad f(x) := \max_{i=1,\ldots,m} F_i(x), \quad x \in \mathbb{R}^n.$$

   Die Funktionen $F_i: \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$, seien auf einer offenen Umgebung von $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbar. Sei

   $$I(x^*) := \{i \in \{1, \ldots, m\} : F_i(x^*) = f(x^*)\}.$$

   Es wird vorausgesetzt, dass reelle Zahlen λ_i^*, $i \in I(x^*)$, existieren mit:

   (a) Es ist

   $$\lambda_i^* \geq 0 \quad (i \in I(x^*)), \qquad \sum_{i \in I(x^*)} \lambda_i^* = 1, \qquad \sum_{i \in I(x^*)} \lambda_i^* \nabla F_i(x^*) = 0.$$

   (b) Mit

   $$T^* := \{p \in \mathbb{R}^n : \nabla F_i(x^*)^T p = 0 \text{ für alle } i \in I(x^*) \text{ mit } \lambda_i^* > 0\}$$

   ist

   $$p^T \Big\{ \sum_{i \in I(x^*)} \lambda_i^* \nabla^2 F_i(x^*) \Big\} p > 0 \qquad \text{für alle } p \in T^* \setminus \{0\}.$$

   Man zeige, dass x^* eine isolierte lokale Lösung von (P) ist.

10. Gegeben sei die Min-Max-Optimierungsaufgabe aus Aufgabe 4, also

    (P) $$\text{Minimiere} \quad f(x) := \max_{t=1,2,3} F_i(x), \quad x \in \mathbb{R}^2,$$

    wobei

    $$F_1(x) := x_1^4 + x_2^2, \qquad F_2(x) := (2 - x_1)^2 + (2 - x_2)^2, \qquad F_3(x) := 2\exp(-x_1 + x_2).$$

    Man zeige, dass $x^* = (1, 1)$ eine isolierte lokale Lösung von (P) ist.

11. Gegeben sei die diskrete L_1-Approximationsaufgabe

    (P) $$\text{Minimiere} \quad f(x) := \sum_{i=1}^{m} |F_i(x)|, \quad x \in \mathbb{R}^n.$$

    Die Funktionen $F_i: \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$, seien auf einer offenen Umgebung von $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbar. Sei

    $$I(x^*) := \{i \in \{1, \ldots, m\} : F_i(x^*) = 0\}.$$

    Es wird vorausgesetzt, dass reelle Zahlen λ_i^*, $i \in I(x^*)$, existieren mit:

    (a) Es ist $\lambda_i^* \in [-1, 1]$, $i \in I(x^*)$, und

    $$\sum_{i \in I(x^*)} \lambda_i^* \nabla F_i(x^*) + \sum_{i \notin I(x^*)} \text{sign}\,(F_i(x^*)) \nabla F_i(x^*) = 0.$$

(b) Mit

$$T^* := \left\{ p \in \mathbb{R}^n : \nabla F_i(x^*)^T p \begin{cases} = 0, & i \in I(x^*) \text{ mit } |\lambda_i^*| < 1, \\ \geq 0, & i \in I(x^*) \text{ mit } \lambda_i^* = 1, \\ \leq 0, & i \in I(x^*) \text{ mit } \lambda_i^* = -1 \end{cases} \right\}$$

ist

$$p^T \Big\{ \sum_{i \in I(x^*)} \lambda_i^* \nabla^2 F_i(x^*) + \sum_{i \notin I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla^2 F_i(x^*) \Big\} p > 0 \text{ für alle } p \in T^* \setminus \{0\}.$$

Man zeige, dass x^* eine isolierte lokale Lösung von (P) ist.

2.2 Konvexe Funktionen

2.2.1 Glatte konvexe Funktionen

In diesem Unterabschnitt sollen glatte, d. h. einmal oder zweimal stetig differenzierbare Funktionen betrachtet und ihre Konvexität durch geeignete Bedingungen charakterisiert werden.

Definition 2.1 Eine Funktion $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ heißt *gleichmäßig konvex* auf der konvexen Menge $D \subset \mathbb{R}^n$, falls eine Konstante $c > 0$ existiert mit

$$(*) \qquad (1-t)f(x) + tf(y) - f((1-t)x + ty) \geq \frac{c}{2} t(1-t)\|x-y\|_2^2 \quad \text{für alle } x, y \in D,\ t \in [0,1].$$

(Dagegen heißt f bekanntlich *konvex* auf D, wenn $(*)$ mit $c = 0$ gilt.)

Bemerkung: Spezielle quadratische Funktionen sind die einfachsten Beispiele für glatte, nichtlineare und konvexe Funktionen. Genauer seien $c \in \mathbb{R}^n$ und eine symmetrische, positiv semidefinite Matrix $Q \in \mathbb{R}^{n \times n}$ gegeben und hiermit $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ durch

$$f(x) := c^T x + \frac{1}{2} x^T Q x$$

definiert. Für $x, y \in \mathbb{R}^n$ und $t \in [0,1]$ ist

$$(1-t)f(x) + tf(y) - f((1-t)x + ty) = \frac{t(1-t)}{2}(x-y)^T Q (x-y) \geq 0,$$

womit die Konvexität von f bewiesen ist. Ist Q sogar positiv definit, so ist f gleichmäßig konvex (als Konstante c kann der kleinste Eigenwert von Q gewählt werden). Als Gradienten von f in x erhält man $\nabla f(x) = c + Qx$. Damit ist ein $x^* \in \mathbb{R}^n$ genau dann ein stationärer Punkt von f bzw. (für positiv semidefinites Q) eine globale Lösung der zugehörigen unrestringierten Optimierungsaufgabe, wenn x^* dem linearen Gleichungssystem $c + Qx = 0$ genügt. Für positiv definites Q bzw. gleichmäßig konvexes f ist dieses lineare Gleichungssystem eindeutig lösbar. Offenbar ist $\nabla^2 f(x) = Q$. □

In den beiden folgenden Sätzen wird die Konvexität und die gleichmäßige Konvexität einer glatten Funktion f durch ihre ersten bzw. zweiten Ableitungen charakterisiert.

Satz 2.2 *Sei $D \subset \mathbb{R}^n$ konvex und $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ auf einer offenen Obermenge von D stetig differenzierbar. Dann gilt:*

1. *f ist genau dann auf D konvex, wenn*

$$\nabla f(x)^T(y-x) \leq f(y) - f(x) \qquad \text{für alle } x, y \in D.$$

2. *f ist genau dann auf D gleichmäßig konvex (mit einer Konstanten $c > 0$), wenn*

$$\frac{c}{2}\|y-x\|_2^2 + \nabla f(x)^T(y-x) \leq f(y) - f(x) \qquad \text{für alle } x, y \in D.$$

Beweis: Für alle $x, y \in D$ und $t \in [0,1]$ sei

$$(1-t)f(x) + tf(y) \geq f((1-t)x + ty) + \frac{c}{2}t(1-t)\|y-x\|_2^2$$

mit einer Konstanten $c \geq 0$. Es sei also f konvex ($c = 0$) bzw. gleichmäßig konvex ($c > 0$). Dann ist

$$f(y) - f(x) \geq \frac{f(x + t(y-x)) - f(x)}{t} + \frac{c}{2}(1-t)\|y-x\|_2^2 \qquad \text{für alle } t \in (0,1].$$

Mit $t \to 0+$ folgt

$$f(y) - f(x) \geq \nabla f(x)^T(y-x) + \frac{c}{2}\|y-x\|_2^2.$$

Damit ist eine Richtung (nämlich "$\Longrightarrow$") bewiesen. Für die andere Richtung "$\Longleftarrow$" nehmen wir an, mit einer Konstanten $c \geq 0$ sei

$$\frac{c}{2}\|y-x\|_2^2 + \nabla f(x)^T(y-x) \leq f(y) - f(x) \qquad \text{für alle } x, y \in D.$$

Seien $x, y \in D$ und $t \in [0,1]$ vorgegeben. Dann ist $z := (1-t)x + ty \in D$ wegen der Konvexität von D und daher nach Voraussetzung

$$\begin{aligned} f(x) - f(z) &\geq \nabla f(z)^T(x-z) + \frac{c}{2}\|x-z\|_2^2, \\ f(y) - f(z) &\geq \nabla f(z)^T(y-z) + \frac{c}{2}\|y-z\|_2^2. \end{aligned}$$

Eine Multiplikation dieser Ungleichungen mit $(1-t)$ bzw. t und anschließende Addition ergibt

$$\begin{aligned} (1-t)f(x) + tf(y) - f((1-t)x + ty) &\geq \frac{c}{2}[(1-t)\|x-z\|_2^2 + t\|y-z\|_2^2] \\ &= \frac{c}{2}t(1-t)\|x-y\|_2^2. \end{aligned}$$

Also ist f konvex ($c = 0$) bzw. gleichmäßig konvex ($c > 0$). □ □

Bemerkung: Konvexe Mengen und konvexe Funktionen machen natürlich nicht nur im bzw. auf dem $\mathbb{R}^n$ einen Sinn, sondern ganz allgemein in einem linearen Raum (der Einfachheit halber mit $\mathbb{R}$ als Skalarkörper). Man erhält dann z. B. die folgende Aussage:

- *Sei E ein linearer Raum, $D \subset E$ konvex und $f\colon D \longrightarrow \mathbb{R}$ eine Abbildung mit der Eigenschaft, dass die Gateaux-Variation in jedem $x \in D$ existiert und eine lineare Abbildung von E nach $\mathbb{R}$ ist. Ist dann $f'(x)(y-x) \leq f(y) - f(x)$ für alle $x, y \in D$, so ist f auf D konvex.*

Denn: Der Beweis kann offenbar vollständig analog zum zweiten Teil des obigen Beweises geführt werden. □

Beispiel: Als Anwendung der letzten Bemerkung wollen wir nachweisen:

- Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n \times n$-Matrizen und $\mathcal{S}_+^{n\times n}$ die (konvexe) Teilmenge der positiv definiten Matrizen. Man definiere $f\colon \mathcal{S}_+^{n\times n} \longrightarrow \mathbb{R}$ durch $f(A) := -\ln\det(A)$. Dann ist f auf $\mathcal{S}_+^{n\times n}$ konvex.

Denn: Wir zeigen zunächst, dass die Gateaux-Variation von f auf $\mathcal{S}_+^{n\times n}$ existiert und linear ist. Für $A \in \mathcal{S}_+^{n\times n}$, also eine symmetrische, positiv definite $n \times n$-Matrix, und $P \in \mathcal{S}^{n\times n}$ ist $A + tP \in \mathcal{S}_+^{n\times n}$ für alle hinreichend kleinen $t > 0$. Für diese t ist

$$\begin{aligned}
\frac{f(A+tP) - f(A)}{t} &= -\frac{\ln\det(A+tP) - \ln\det(A)}{t} \\
&= -\frac{\ln\det(I + tA^{-1/2}PA^{-1/2})}{t} \\
&= -\frac{1}{t}\ln\prod_{i=1}^{n}\lambda_i(I + tA^{-1/2}PA^{-1/2}) \\
&= -\frac{1}{t}\sum_{i=1}^{n}\ln(1 + t\lambda_i(A^{-1/2}PA^{-1/2})).
\end{aligned}$$

Folglich ist

$$f'(A)P = -\sum_{i=1}^{n}\lambda_i(A^{-1/2}PA^{-1/2}) = -\mathrm{tr}(A^{-1/2}PA^{-1/2}),$$

die Gateaux-Variation existiert also auf $\mathcal{S}_+^{n\times n}$ und ist linear. Für beliebige $A, B \in \mathcal{S}_+^{n\times n}$ ist ferner

$$\begin{aligned}
f'(A)(B-A) &= -\mathrm{tr}(A^{-1/2}(B-A)A^{-1/2}) \\
&= -\mathrm{tr}(A^{-1/2}BA^{-1/2}) + n \\
&= -n\Big(\frac{1}{n}\sum_{i=1}^{n}\lambda_i(A^{-1/2}BA^{-1/2})\Big) + n \\
&\leq -n\Big(\prod_{i=1}^{n}\lambda_i(A^{-1/2}BA^{-1/2})\Big)^{1/n} + n \\
&\qquad \text{(Ungleichung vom geometrisch-arithmetischen Mittel)} \\
&= -n\det(A^{-1/2}BA^{-1/2})^{1/n} + n \\
&= -\frac{n}{\det(A)^{1/n}}[\det(B)^{1/n} - \det(A)^{1/n}]
\end{aligned}$$

$$\begin{aligned} &\leq -n[\ln\det(B)^{1/n} - \ln\det(A)^{1/n}] \\ &\qquad \text{(Konkavität des Logarithmus auf } \mathbb{R}_+) \\ &= f(B) - f(A). \end{aligned}$$

Daher ist f konvex auf $\mathcal{S}_+^{n\times n}$. □

Satz 2.3 *Sei $D \subset \mathbb{R}^n$ konvex und $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ auf einer offenen Obermenge von D zweimal stetig differenzierbar. Dann gilt:*

1. *Ist $\nabla^2 f(x)$ positiv semidefinit für alle $x \in D$, so ist f auf D konvex.*
2. *Existiert eine Konstante $c > 0$ mit*

$$c\|p\|_2^2 \leq p^T \nabla^2 f(x) p \qquad \text{für alle } x \in D \text{ und alle } p \in \mathbb{R}^n,$$

   so ist f auf D gleichmäßig konvex (mit der Konstanten c).
3. *Ist D auch offen, so gelten in 1. und 2. auch die Umkehrungen.*

Beweis: Die ersten beiden Aussagen werden zusammen bewiesen, indem wir annehmen, es existiere eine Konstante $c \geq 0$ mit $c\|p\|_2^2 \leq p^T \nabla^2 f(x) p$ für alle $x \in D$ und alle $p \in \mathbb{R}^n$.

Seien $x, y \in D$. Definiert man $\phi\colon [0,1] \longrightarrow \mathbb{R}$ durch $\phi(t) := f(x + t(y-x))$, so ist

$$\phi(1) - \phi(0) - \phi'(0) = \frac{1}{2}\phi''(t_0) \qquad \text{mit } t_0 \in (0,1)$$

bzw.

$$f(y) - f(x) - \nabla f(x)^T (y-x) = \frac{1}{2}(y-x)^T \nabla^2 f(\underbrace{x + t_0(y-x)}_{\in D})(y-x) \geq \frac{c}{2}\|y-x\|_2^2.$$

Aus Satz 2.2 folgt, dass f konvex ($c = 0$) bzw. gleichmäßig konvex (mit der Konstanten $c > 0$) ist.

Nun sei D auch offen und f auf D konvex ($c = 0$) bzw. gleichmäßig konvex (mit der Konstanten $c > 0$). Seien $x \in D$ und $p \in \mathbb{R}^n$ beliebig. Wegen der Offenheit von D ist $x + tp \in D$ für alle hinreichend kleinen $|t|$. Nach Satz 2.2 gilt für diese t:

$$\begin{aligned} f(x+tp) - f(x) &\geq \nabla f(x)^T (tp) + \frac{c}{2} t^2 \|p\|_2^2, \\ f(x) - f(x+tp) &\geq -\nabla f(x+tp)^T (tp) + \frac{c}{2} t^2 \|p\|_2^2, \end{aligned}$$

so dass nach Addition $[\nabla f(x+tp) - \nabla f(x)]^T (tp) \geq ct^2 \|p\|_2^2$ für alle hinreichend kleinen $|t|$. Wegen

$$\begin{aligned} p^T \nabla^2 f(x) p &= \lim_{t\to 0} \frac{[\nabla f(x+tp) - \nabla f(x)]^T p}{t} \\ &= \lim_{t\to 0+} \frac{[\nabla f(x+tp) - \nabla f(x)]^T (tp)}{t^2} \\ &\geq c\|p\|_2^2 \end{aligned}$$

ist die Behauptung bewiesen. □ □

Mit Hilfe von Satz 2.3 kann die Konvexität gegebener Funktionen nachgewiesen werden, ähnlich wie im eindimensionalen Fall, bei dem man zum Nachweis der Konvexität die zweite Ableitung berechnet und zeigt, dass diese nichtnegativ ist. Wir geben gleich ein nichttriviales Beispiel an.

Beispiel: Für $k = 1, \ldots, m$ seien $c_k > 0$ und $a_k \in \mathbb{R}^n$. Hiermit seien auf dem $\mathbb{R}^n$ die reellwertigen Funktionen f und g durch

$$f(x) := \sum_{k=1}^{m} c_k \exp(a_k^T x), \qquad g(x) := \ln\Big(\sum_{k=1}^{m} c_k \exp(a_k^T x)\Big)$$

definiert. Wir wollen uns überlegen, dass f und g auf dem $\mathbb{R}^n$ konvex sind. Dass f konvex ist, ist ziemlich klar, da f eine positive Linearkombination konvexer Funktionen ist ($\exp(a_k^T \cdot)$ ist Kombination einer linearen und einer monoton wachsenden, konvexen Funktion, daher selbst konvex, siehe Aufgabe 1). Man erkennt es auch daran, dass

$$\nabla^2 f(x) = \sum_{k=1}^{m} c_k \exp(a_k^T x) a_k a_k^T,$$

also die Hessesche $\nabla^2 f(x)$ von f in x, eine positive Linearkombination der positiv semidefiniten Matrizen $a_k a_k^T$, $k = 1, \ldots, m$, ist, also selbst positiv semidefinit ist, was nach Satz 2.3 die Konvexität von f impliziert. Der Nachweis der Konvexität von g ist ein wenig komplizierter. Hier erhält man als Hessesche

$$\begin{aligned} \nabla^2 g(x) &= \frac{1}{f(x)^2}\Big\{\Big(\sum_{k=1}^{m} c_k \exp(a_k^T x)\Big)\Big(\sum_{k=1}^{m} c_k \exp(a_k^T x) a_k a_k^T\Big) \\ &\qquad - \Big(\sum_{k=1}^{m} c_k \exp(a_k^T x) a_k\Big)\Big(\sum_{k=1}^{m} c_k \exp(a_k^T x) a_k\Big)^T\Big\}. \end{aligned}$$

Für alle $p \in \mathbb{R}^n$ ist wegen der Cauchy-Schwarzschen Ungleichung

$$\begin{aligned} \Big(\sum_{k=1}^{m} c_k \exp(a_k^T x) a_k^T p\Big)^2 &= \Big(\sum_{k=1}^{m} \sqrt{c_k \exp(a_k^T x)}[\sqrt{c_k \exp(a_k^T x)} a_k^T p]\Big)^2 \\ &\le \Big(\sum_{k=1}^{m} c_k \exp(a_k^T x)\Big)\Big(\sum_{k=1}^{m} c_k \exp(a_k^T x)(a_k^T p)^2\Big) \end{aligned}$$

und daher

$$\begin{aligned} p^T \nabla^2 g(x) p &= \frac{1}{f(x)^2}\Big\{\Big(\sum_{k=1}^{m} c_k \exp(a_k^T x)\Big)\Big(\sum_{k=1}^{m} c_k \exp(a_k^T x)(a_k^T p)^2\Big) \\ &\qquad - \Big(\sum_{k=1}^{m} c_k \exp(a_k^T x) a_k^T p\Big)^2\Big\} \\ &\ge 0. \end{aligned}$$

Aus Satz 2.3 folgt die Konvexität von g. □

Bemerkung: Konvexe Funktionen spielen aus mehreren Gründen eine wichtige Rolle in der Optimierung, insbesondere auch bei unrestringierten Optimierungsaufgaben. Bei einer konvexen Zielfunktion stimmen lokale und globale Minima überein, daher wird man i. Allg. nur bei einer konvexen Zielfunktion eine globale Konvergenzaussage gegen eine globale Lösung für später zu untersuchende Verfahren erwarten können. Ist andererseits in einer lokalen Lösung x^* die hinreichende Bedingung zweiter Ordnung erfüllt, ist also f in einer Umgebung von x^* zweimal stetig differenzierbar und ist $\nabla^2 f(x^*)$ positiv definit, so existiert eine offene, konvexe Umgebung D von x^* und eine Konstante $c > 0$ derart, dass

$$c\|p\|_2^2 \leq p^T \nabla^2 f(x) p \qquad \text{für alle } x \in D \text{ und alle } p \in \mathbb{R}^n.$$

Denn ist $\lambda_{\min}^* > 0$ der kleinste Eigenwert von $\nabla^2 f(x^*)$, so ist bekanntlich

$$\lambda_{\min}^* \|p\|_2^2 \leq p^T \nabla^2 f(x^*) p \qquad \text{für alle } p \in \mathbb{R}^n.$$

Bestimmt man daher eine offene, konvexe Umgebung D von x^* (etwa eine offene Kugel um x^* mit einem hinreichend kleinen Radius) so, dass

$$\|\nabla^2 f(x) - \nabla^2 f(x^*)\|_2 \leq \frac{\lambda_{\min}^*}{2} \qquad \text{für alle } x \in D,$$

so ist für beliebiges $x \in D$ und alle $p \in \mathbb{R}^n$:

$$\begin{aligned} p^T \nabla^2 f(x) p &= p^T \nabla^2 f(x^*) p + p^T [\nabla^2 f(x) - \nabla^2 f(x^*)] p \\ &\geq \lambda_{\min}^* \|p\|_2^2 - \|\nabla^2 f(x) - \nabla^2 f(x^*)\|_2 \, \|p\|_2^2 \\ &\geq \frac{\lambda_{\min}^*}{2} \|p\|_2^2 \end{aligned}$$

und damit $(*)$ erfüllt. Man beachte, dass $(*)$ gleichwertig mit der Aussage ist, dass alle Eigenwerte von $\nabla^2 f(x)$ für jedes $x \in D$ größer oder gleich c sind. Wir haben uns damit überlegt: Ist f auf einer Umgebung von x^* zweimal stetig differenzierbar und ist $\nabla^2 f(x^*)$ positiv definit, so ist f auf einer hinreichend kleinen, offenen, konvexen Umgebung von x^* gleichmäßig konvex. □

Beispiel: Die Rosenbrock-Funktion

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

besitzt $x^* = (1, 1)$ als einziges lokales und damit globales Minimum. Als Hessesche von f in x berechnet man

$$\nabla^2 f(x) = \begin{pmatrix} 1200x_1^2 - 400x_2 + 2 & -400x_1 \\ -400x_1 & 200 \end{pmatrix},$$

so dass

$$\nabla^2 f(x^*) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix} \qquad \text{mit den Eigenwerten} \quad \begin{aligned} \lambda_1^* &\approx 1001.6, \\ \lambda_2^* &\approx 0.4. \end{aligned}$$

Also ist $\nabla^2 f(x^*)$ positiv definit, ein Eigenwert klein, der andere groß. Dies ist ein Indiz dagür, dass x^* niedrigster Punkt in einem "langgestreckten Tal" ist. Die Rosenbrock-Funktion ist nicht auf dem gesamten $\mathbb{R}^2$ konvex, da $\nabla^2 f(x)$ offenbar auch negative Eigenwerte besitzen kann. □

2.2.2 Nichtdifferenzierbare konvexe Funktionen

Konvexe Funktionen haben einige doch recht überraschende Glattheitseigenschaften. Wir fassen alle wesentlichen Eigenschaften auf dem ganzen $\mathbb{R}^n$ konvexer Funktionen in dem folgenden Satz zusammen, von dem die Teile 2 und 3 durch Lemma 1.6 schon bekannt sind.

Satz 2.4 *Sei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ konvex. Dann gilt:*

1. *f ist auf dem $\mathbb{R}^n$ stetig.*
2. *f besitzt in jedem $x \in \mathbb{R}^n$ in jede Richtung $p \in \mathbb{R}^n$ eine Richtungsableitung $f'(x;p)$, d. h.*
$$f'(x;p) := \lim_{t \to 0+} \frac{f(x+tp) - f(x)}{t}$$
existiert für alle $x, p \in \mathbb{R}^n$.
3. *Die sogenannte Gateaux-Variation $f'(x;\cdot)\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ hat bei festem $x \in \mathbb{R}^n$ die folgenden Eigenschaften:*
   (a) *$f'(x;p) \le f(x+p) - f(x)$ für alle $p \in \mathbb{R}^n$.*
   (b) *$f'(x;\cdot)$ ist nichtnegativ homogen, d. h. $f'(x;\alpha p) = \alpha f'(x;p)$ für alle $\alpha \ge 0$, $p \in \mathbb{R}^n$.*
   (c) *$f'(x;\cdot)$ ist subadditiv, d. h. $f'(x;p+q) \le f'(x;p) + f'(x;q)$ für alle $p, q \in \mathbb{R}^n$.*
   (d) *$f'(x;\cdot)$ ist konvex.*
4. *Sei $K \subset \mathbb{R}^n$ kompakt. Dann ist f auf K lipschitzstetig, d. h. es existiert eine Konstante $L > 0$ mit*
$$|f(x) - f(y)| \le L \, \|x - y\| \qquad \text{für alle } x, y \in \mathbb{R}^n.$$
5. *Bei vorgegebenem $x \in \mathbb{R}^n$ heißt*
$$\partial f(x) := \{z \in \mathbb{R}^n : z^T(y-x) \le f(y) - f(x) \ \text{ für alle } y \in \mathbb{R}^n\}$$
das Subdifferential von f in x, Elemente von $\partial f(x)$ heißen Subgradienten. Dann gilt:
   (a) *Für jedes $x \in \mathbb{R}^n$ ist das Subdifferential $\partial f(x)$ nichtleer, konvex und kompakt.*
   (b) *Es ist*
$$f'(x;p) = \max_{z \in \partial f(x)} z^T p \qquad \text{für alle } p \in \mathbb{R}^n.$$

Beweis: Alle Aussagen und jeweils einen Beweis findet man (eventuell mit einiger Mühe) in dem Standardwerk von R. T. Rockafellar (1970)[8]. Wir wollen auf Beweise verzichten. □ □

[8] R. T. Rockafellar (1970) *Convex Analysis*. Princeton University Press, Princeton.

2.2.3 Aufgaben

1. Sei $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ affin linear und $f\colon \mathbb{R}^m \longrightarrow \mathbb{R}$ konvex. Dann ist auch $h := f \circ g$ konvex.

2. Sei $D \subset \mathbb{R}^n$ konvex, $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ auf D konvex und $f\colon \mathbb{R} \longrightarrow \mathbb{R}$ konvex und monoton nicht fallend. Dann ist $h := f \circ g$ auf D konvex.

3. Seien $F_i\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$, konvex auf dem $\mathbb{R}^n$. Dann ist auch die durch
$$f(x) := \max_{i=1,\ldots,m} F_i(x)$$
definierte Funktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ konvex (auf dem $\mathbb{R}^n$).

4. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit. Die Abbildung $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ sei definiert durch
$$f(x) := \frac{1}{2}(x^T A x)^2.$$
Man berechne den Gradienten und die Hessesche von f und weise nach, dass f auf dem $\mathbb{R}^n$ konvex ist.

5. Man zeige: Ist $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ gleichmäßig konvex mit der Konstanten $c > 0$, so ist
$$\frac{c}{2}\|p\|_2^2 + f'(x;p) \le f(x+p) - f(x) \qquad \text{für alle } x, p \in \mathbb{R}^n.$$

6. Sei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ gleichmäßig konstant mit der Konstanten $c > 0$. Man zeige:

   (a) Es ist
   $$\frac{c}{2}\|p\|_2^2 + v^T p \le f(x+p) - f(x)$$
   für beliebige $x \in \mathbb{R}^n$, $v \in \partial f(x)$ und $p \in \mathbb{R}^n$.

   (b) Für beliebiges $x_0 \in \mathbb{R}^n$ ist die Niveaumenge
   $$L_0 := \{x \in \mathbb{R}^n : f(x) \le f(x_0)\}$$
   kompakt.

7. Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n \times n$-Matrizen und $\mathcal{S}_+^{n\times n}$ die konvexe, offene Teilmenge aller positiv definiten $n \times n$-Matrizen. Man definiere $f\colon \mathcal{S}_+^{n\times n} \subset \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ durch
$$f(A) := \frac{\operatorname{tr}(A)/n}{\det(A)^{1/n}}.$$
Man zeige:

   (a) Es ist $f(A) \ge 1$ für alle $A \in \mathcal{S}_+^{n\times n}$, ferner ist $f(A) = 1$ genau dann, wenn A ein positives Vielfaches der Identität ist.

   (b) Für jedes $A \in \mathcal{S}_+^{n\times n}$ und jedes $P \in \mathcal{S}^{n\times n}$ existiert
   $$f'(A;P) := \lim_{t\to 0+} \frac{f(A+tp) - f(A)}{t}.$$
   Man berechne $f'(A;P)$ und zeige, dass die Abbildung $f'(A;\cdot)\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ linear ist.

(c) Sind $A, B \in \mathcal{S}_+^{n\times n}$ und $0 \leq f'(A; B - A)$, so ist $f(A) \leq f(B)$.

8. Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n \times n$-Matrizen. Wir definieren die Abbildung $\lambda_{\max}\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ dadurch, dass $\lambda_{\max}(A)$ den maximalen Eigenwert von $A \in \mathcal{S}^{n\times n}$ bedeutet. Man zeige:

   (a) Die Abbildung $\lambda_{\max}\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ ist konvex.

   (b) Die Abbildung $\lambda_{\max}\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ besitzt auf $\mathcal{S}^{n\times n}$ eine Gateaux-Variation, d. h. für alle $A, P \in \mathcal{S}^{n\times n}$ existiert

$$\lambda'_{\max}(A; P) := \lim_{t\to 0+} \frac{\lambda_{\max}(A + tP) - \lambda_{\max}(A)}{t}.$$

   (c) Die Gateaux-Variation von $\lambda_{\max}$ ist für $A, P \in \mathcal{S}^{n\times n}$ gegeben durch

$$f'(A; P) = \max_{q\in Q_{\max}(A)\setminus\{0\}} \frac{q^T P q}{q^T q}.$$

   Hierbei bezeichne $Q_{\max}(A)$ den Eigenraum zu $\lambda_{\max}(A)$, also die lineare Hülle aller Eigenvektoren zu $\lambda_{\max}(A)$.

Kapitel 3

Schrittweitenverfahren

Wir betrachten in diesem Kapitel eine Klasse von Verfahren zur Lösung unrestringierter Optimierungsaufgaben, der sich bis auf die später zu besprechenden Trust-Region-Verfahren fast alle gebräuchlichen Verfahren unterordnen lassen. Dem deutschen Wort "Schrittverfahren" entspricht der englische Begriff "Line Search Methods".

3.1 Ein Modellalgorithmus

Wir betrachten die unrestringierte Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n,$$

und nehmen an, die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ besitze in jedem $x \in \mathbb{R}^n$ eine Gateaux-Variation $f'(x;\cdot)$, sei also z. B. aus $C^1(\mathbb{R}^n)$. Die meisten der später zu untersuchenden konkreten Verfahren lassen sich dem folgenden *Modellalgorithmus* unterordnen:

- Sei $x_0 \in \mathbb{R}^n$ gegeben.
- Für $k = 0, 1, \ldots$:
  - Test auf Abbruch: Falls $f'(x_k; p) \geq 0$ für alle $p \in \mathbb{R}^n$, dann: STOP, da x_k stationäre Lösung von (P) ist.
  - Wahl einer (Abstiegs-) Richtung: Bestimme $p_k \in \mathbb{R}^n$ mit $f'(x_k; p_k) < 0$.
  - Wahl einer Schrittweite: Bestimme $t_k > 0$ mit $f(x_k + t_k p_k) < f(x_k)$.
  - Bestimme neue Näherung: Setze $x_{k+1} := x_k + t_k p_k$.

Uns kommt es darauf an, deutlich zu machen, dass dieser Algorithmus sich aus einer *Richtungs-* und einer *Schrittweitenstrategie* zusammensetzt. Durch eine Spezifikation dieser Strategien wird aus dem Modellalgorithmus ein auf die gegebene Aufgabe anwendbares, konkretes Verfahren. In diesem Kapitel wird noch keine Festlegung der Richtungsstrategie erfolgen, dies wird erst in den folgenden Kapiteln geschehen.

3.1.1 Schrittweitenstrategien bei glatter Zielfunktion

Wir betrachten die unrestringierte Optimierungsaufgabe (P) und nehmen an, dass die folgenden Voraussetzungen für die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ erfüllt sind:

(V) (a) Mit einem gegebenen $x_0 \in \mathbb{R}^n$ (gewöhnlich Startwert eines Iterationsverfahrens) ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ kompakt.

(b) Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar.

(c) Der Gradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig, d. h. es existiert eine Konstante $\gamma > 0$ mit

$$\|\nabla f(x) - \nabla f(y)\| \leq \gamma \|x - y\| \qquad \text{für alle } x, y \in L_0.$$

Auf die Wahl von Richtungsstrategien im Modellalgorithmus werden wir (bis auf einige Beispiele) erst in späteren Abschnitten eingehen. In diesem Unterabschnitt stellen wir uns auf den Standpunkt, es sei eine aktuelle Näherung $x_c = x_k \in L_0$ gegeben (der Index c links bedeutet **c**urrent), es sei x_c keine stationäre Lösung von (P), also $\nabla f(x_c) \neq 0$, und $p = p_k$ eine *Abstiegsrichtung* für f in x_c bzw. $\nabla f(x_c)^T p < 0$. Z. B. ist $p := -\nabla f(x_c)$ (diese Richtungswahl führt auf das schon 1827 von Cauchy angegebene *Gradientenverfahren*, manchmal auch *Verfahren des steilsten Abstiegs* genannt) oder allgemeiner $p := -H_c \nabla f(x_c)$ mit einer symmetrischen, positiv definiten Matrix $H_c \in \mathbb{R}^{n\times n}$ eine Abstiegsrichtung.

Ziel dieses Unterabschnittes ist es, Strategien zur Berechnung einer Schrittweite $t > 0$ anzugeben, für welche die Verminderung $f(x_c) - f(x_c + tp)$ der Zielfunktion einerseits positiv und andererseits so groß ist, dass (einfache) Konvergenzaussagen für das entstehende, von einer speziellen Richtungsstrategie weitgehend unabhängige Verfahren gemacht werden können. Hierbei wird i. Allg. von der Schrittweite gefordert, dass

$$f(x_c + tp) - f(x_c) \geq -\alpha t \nabla f(x_c)^T p$$

mit einer von (x_c, p) unabhängigen Konstanten $\alpha \in (0, 1)$, wobei typischerweise $\alpha := 0.0001$ gesetzt ist. Diese Bedingung wird bei C. T. Kelley (1999, S. 41) und J. Nocedal, S. J. Wright (1999, S. 37) eine *Bedingung für hinreichenden Abstieg* bzw. *sufficient decrease condition* genannt.

Das folgende Lemma wird sich als nützlich erweisen, wenn $f(x_c) - f(x_c + tp)$ nach unten abgeschätzt werden soll.

Lemma 1.1 *Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Sei $x_c \in L_0$ keine stationäre Lösung von (P) und $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x_c, d. h. es sei $\nabla f(x_c)^T p < 0$. Ist dann $\hat{t} = \hat{t}(x_c, p)$ die erste positive Nullstelle der durch $\psi(t) := f(x_c) - f(x_c + tp)$ definierten Abbildung $\psi\colon [0, \infty) \longrightarrow \mathbb{R}$, so ist*

$$f(x_c + tp) \leq f(x_c) + t\nabla f(x_c)^T p + t^2 \frac{\gamma}{2} \|p\|_2^2 \qquad \text{für alle } t \in [0, \hat{t}]$$

und

$$-\frac{2\nabla f(x_c)^T p}{\gamma \|p\|_2^2} \leq \hat{t}.$$

Beweis: Wegen Voraussetzung (V) existiert $\hat{t}$ und es ist $x_c + tp \in L_0$ für alle $t \in [0, \hat{t}]$. Für $t \in [0, \hat{t}]$ erhält man

$$\begin{aligned} f(x_c + tp) &= f(x_c) + t\nabla f(x_c)^T p + \int_0^t [\nabla f(\underbrace{x_c + sp}_{\in L_0}) - \nabla f(x_c)]^T p \, ds \\ &\leq f(x_c) + t\nabla f(x_c)^T p + \int_0^t \gamma \|p\|_2^2 \, ds \\ &\qquad \text{(Voraussetzung (V) (c) und Cauchy-Schwarzsche Ungleichung)} \\ &= f(x_c) + t\nabla f(x_c)^T p + t^2 \frac{\gamma}{2} \|p\|_2^2, \end{aligned}$$

und damit die erste Behauptung. Die zweite folgt, indem man in der gerade eben bewiesenen Ungleichung $t = \hat{t}$ setzt. □ □

Eine naheliegende Schrittweitenstrategie besteht darin, als Schrittweite $t^* > 0$ eine globale oder die erste stationäre Lösung der eindimensionalen Minimierungsaufgabe

$$\text{Minimiere} \quad f(x_c + tp), \quad t \in [0, \infty),$$

zu wählen. In diesem Fall wird $t^* > 0$ also so bestimmt, dass

$$f(x_c + t^* p) = \min_{t \geq 0} f(x_c + tp)$$

bzw.

$$\nabla f(x_c + t^* p)^T p = 0 \quad \text{und} \quad \nabla f(x_c + tp)^T p < 0 \quad \text{für alle } t \in [0, t^*).$$

Unter der Voraussetzung (V) ist die Existenz dieser Schrittweite gesichert. Man spricht von einer *exakten Schrittweite*, da eine eindimensionale Minimierungsaufgabe bzw. eine eindimensionale Nullstellenaufgabe zur Bestimmung der Schrittweite exakt zu lösen ist. Es istklar, dass nur in Ausnahmefällen die exakte Schrittweite (in endlich vielen Schritten) berechnet werden kann, i. Allg. muss man sich mit einer Näherung begnügen.

Im folgenden Satz werden wir die durch die exakte Schrittweite erreichbare Verminderung abschätzen.

Satz 1.2 *Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Sei $x_c \in L_0$ keine stationäre Lösung von (P) und $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x_c. Zur Abkürzung sei $\phi(t) := f(x_c + tp)$. Ist t^* die erste positive Nullstelle von $\phi'(\cdot)$ auf $[0, \infty)$, so ist*

$$(*) \qquad -\frac{\nabla f(x_c)^T p}{\gamma \|p\|_2^2} \leq t^*, \qquad f(x_c) - f(x_c + t^* p) \geq \frac{1}{2\gamma} \Big(\frac{\nabla f(x_c)^T p}{\|p\|_2} \Big)^2.$$

Beweis: Wegen der Voraussetzung (V) ist die Existenz der exakten Schrittweite t^* gesichert. Die Funktion ϕ ist monoton fallend auf $[0, t^*]$, ferner ist $t^* \leq \hat{t}$, wobei $\hat{t}$ wie in Lemma 1.1 die erste positive Nullstelle von $\phi(0) - \phi(\cdot)$ ist. Der Beweis des Satzes erfolgt in zwei Schritten. Zunächst wird t^* nach unten abgeschätzt und dann mit Hilfe der Monotonie von ϕ und Lemma 1.1 die Behauptung bewiesen.

Wegen der Lipschitzstetigkeit von $\nabla f(\cdot)$ auf der Niveaumenge L_0 ist

$$0 = \nabla f(x_c + t^* p)^T p = \nabla f(x_c)^T p + [\nabla f(\underbrace{x_c + t^* p}_{\in L_0}) - \nabla f(x_c)]^T p \leq \nabla f(x_c)^T p + \gamma t^* \|p\|_2^2,$$

so dass

$$\tilde{t} := -\frac{\nabla f(x_c)^T p}{\gamma \|p\|_2^2} \leq t^*.$$

Hieraus erhält man

$$f(x_c + t^* p) \leq f(x_c + \tilde{t} p) \leq f(x_c) + \tilde{t} \nabla f(x_c)^T p + \tilde{t}^2 \frac{\gamma}{2} \|p\|_2^2 = f(x_c) - \frac{1}{2\gamma} \Big(\frac{\nabla f(x_c)^T p}{\|p\|_2} \Big)^2,$$

die Behauptung ist bewiesen. □ □

Da eine exakte Schrittweite i. Allg. nicht in endlich vielen Schritten realisiert werden kann, sind zunehmend *inexakte* Schrittweiten in Theorie und Praxis untersucht und angewandt worden. Bei der sogenannten *Wolfe-Schrittweite* wird bei vorgegebenen $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ ein $t > 0$ so bestimmt, dass

(a) $$f(x_c + tp) \leq f(x_c) + \alpha t \nabla f(x_c)^T p,$$

also die Bedingung für hinreichenden Abstieg erfüllt ist, und

(b) $$\nabla f(x_c + tp)^T p \geq \beta \nabla f(x_c)^T p$$

gilt. Die "sufficient decrease condition" (a) ist für alle hinreichend kleinen $t > 0$ erfüllt und besagt, dass die erwünschte Verminderung $f(x_c) - f(x_c + tp)$ der Zielfunktion nicht kleiner als ein kleines Vielfaches von $-t \nabla f(x_c)^T p$ ist. Die Forderung (b) soll sichern, dass nicht zu kleine Schrittweiten gewählt werden.

Beispiel: Wir betrachten ein einfaches Beispiel, und zwar sei $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ definiert durch

$$f(x) := c^T x + \frac{1}{2} x^T Q x$$

mit symmetrischem und positiv definitem $Q \in \mathbb{R}^{n \times n}$. Sei $(c + Qx_c)^T p < 0$, also p eine Abstiegsrichtung in x_c. Dann ist

$$\begin{aligned} f(x_c) - f(x_c + tp) + \alpha t \nabla f(x_c)^T p &= -(1 - \alpha) t \nabla f(x_c)^T p - \frac{1}{2} t^2 p^T \nabla^2 f(x_c) p \\ &= -t(1 - \alpha)(c + Qx_c)^T p - \frac{1}{2} t^2 p^T Q p. \end{aligned}$$

Daher ist (a) erfüllt, wenn

$$0 < t \leq -2(1 - \alpha) \frac{(c + Qx_c)^T p}{p^T Q p}.$$

Weiter ist

$$\begin{aligned} \nabla f(x_c + tp)^T p - \beta \nabla f(x_c)^T p &= (c + Q(x_c + tp))^T p - \beta (c + Qx_c)^T p \\ &= (1 - \beta)(c + Qx_c)^T p + t p^T Q p, \end{aligned}$$

so dass die Bedingung (b) für alle t mit

$$t \geq -(1-\beta)\frac{(c+Qx_c)^T p}{p^T Q p}$$

gilt. Daher sind (a) und (b) für alle t mit

$$-(1-\beta)\frac{(c+Qx_c)^T p}{p^T Q p} \leq t \leq -2(1-\alpha)\frac{(c+Qx_c)^T p}{p^T Q p}$$

erfüllt. Wegen $1-\beta < 1-\alpha < 2(1-\alpha)$ handelt es sich hier um ein ganzes Intervall von Punkten, für die (a) und (b) erfüllt sind. □

Im folgenden Satz wollen wir zeigen, dass stets eine Schrittweite $t > 0$ existiert, die den Forderungen (a) und (b) genügt, ferner soll die durch die Wolfe-Schrittweite erreichbare Verminderung der Zielfunktion nach unten abgeschätzt werden.

Satz 1.3 *Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Sei $x_c \in L_0$ keine stationäre Lösung von (P) und p eine Abstiegsrichtung für f in x_c. Seien $\alpha \in (0, \frac{1}{2})$, $\beta \in (\alpha, 1)$ gegeben und*

$$T_W(x_c, p) := \left\{ t > 0 : \begin{array}{rcl} f(x_c + tp) & \leq & f(x_c) + \alpha t \nabla f(x_c)^T p, \\ \nabla f(x_c + tp)^T p & \geq & \beta \nabla f(x_c)^T p \end{array} \right\}$$

die Menge der Wolfe-Schrittweiten in x_c in Richtung p. Dann gilt:

1. *Es ist $T_W(x_c, p) \neq \emptyset$.*
2. *Es existiert eine Konstante $\theta > 0$, die nur von α, β und γ (der Lipschitzkonstanten von $\nabla f(\cdot)$ auf L_0) abhängt, nicht aber von x_c oder p, mit*

$$(*) \qquad f(x_c) - f(x_c + tp) \geq \theta \Big(\frac{\nabla f(x_c)^T p}{\|p\|_2} \Big)^2 \qquad \text{für alle } t \in T_W(x_c, p).$$

Beweis: Zur Abkürzung setze man

$$\Phi(t) := f(x_c) + \alpha t \nabla f(x_c)^T p - f(x_c + tp).$$

Ist t^* die erste positive Nullstelle $\nabla f(x_c + tp)^T p$, so ist

$$\Phi'(0) = -(1-\alpha)\nabla f(x_c)^T p > 0, \qquad \Phi'(t^*) = \alpha \nabla f(x_c)^T p < 0.$$

Wegen $\Phi(0) = 0$ existiert daher ein $t \in (0, t^*)$ mit $\Phi(t) > 0$ und $\Phi'(t) = 0$. Offenbar ist $t \in T_W(x_c, p)$, der erste Teil des Satzes ist bewiesen.

Sei $t \in T_W(x_c, p)$ gegeben, insbesondere ist $f(x_c + tp) < f(x_c)$ und $x_c + tp \in L_0$. Aus

$$-(1-\beta)\nabla f(x_c)^T p \leq [\nabla f(x_c + tp) - \nabla f(x_c)]^T p \leq \gamma t \, \|p\|_2^2$$

folgt

$$f(x_c) - f(x_c + tp) \geq -\alpha t \nabla f(x_c)^T p \geq \frac{\alpha(1-\beta)}{\gamma} \Big(\frac{\nabla f(x_c)^T p}{\|p\|_2} \Big)^2,$$

die Behauptung ist mit $\theta := \alpha(1-\beta)/\gamma$ bewiesen. □ □

Bemerkung: Etwas genauer wollen wir auf die Frage eingehen, wie man in endlich vielen Schritten ein $t \in T_W(x_c, p)$ bestimmen kann. Wir orientieren uns an J. E. Dennis, R. B. Schnabel (1983, S. 328), siehe aber auch C. Geiger, C. Kanzow (1999, S. 45 ff.) und P. Spellucci (1993, S. 102 ff.). Die Aufgabenstellung ist also die folgende:

- Gegeben: Konstanten $\alpha \in (0, \frac{1}{2})$, $\beta \in (\alpha, 1)$ sowie $x_c \in L_0$ und $p \in \mathbb{R}^n$ mit $\nabla f(x_c)^T p < 0$.
- Gesucht: Schrittweite $t > 0$ mit

  (a) $$f(x_c + tp) \le f(x_c) + \alpha t \nabla f(x_c)^T p$$

  und

  (b) $$\nabla f(x_c + tp)^T p \ge \beta \nabla f(x_c)^T p.$$

Die Schrittweite $t = 1$ spielt bei Quasi-Newton-Verfahren eine besondere Rolle, so dass es sehr oft vernünftig ist, am Anfang zu testen, ob diese Schrittweite den beiden geforderten Ungleichungen genügt. Der Algorithmus wird aus zwei Phasen bestehen. In der ersten Phase wird ein Intervall $[t_{\min}, t_{\max}]$ bestimmt mit:

- Es ist $t_{\min} < t_{\max}$, weiter ist

  $$f(x_c + t_{\min} p) \le f(x_c) + \alpha t_{\min} \nabla f(x_c)^T p, \qquad f(x_c + t_{\max} p) > f(x_c) + \alpha t_{\max} \nabla f(x_c)^T p$$

  und

  $$\nabla f(x_c + t_{\min} p)^T p < \beta \nabla f(x_c)^T p.$$

Dann ist es nämlich nicht schwierig zu zeigen:

- Es existiert ein $t \in [t_{\min}, t_{\max}]$, für welches (a) und (b) erfüllt sind.

Denn: Man definiere

$$\Phi(t) := f(x_c + tp) - f(x_c) - \alpha t \nabla f(x_c)^T p.$$

Nach Konstruktion ist $\Phi(t_{\min}) \le 0$, $\Phi(t_{\max}) > 0$ und

$$\Phi'(t_{\min}) = \nabla f(x_c + t_{\min} p)^T p - \alpha \nabla f(x_c)^T p < \underbrace{(\beta - \alpha)}_{>0} \underbrace{\nabla f(x_c)^T p}_{<0} < 0.$$

Ist $t^* > t_{\min}$ die erste Nullstelle von $\Phi(\cdot)$ in $(t_{\min}, t_{\max})$, so ist (a) für alle $t \in [t_{\min}, t^*]$ erfüllt. In $[t_{\min}, t^*]$ existiert ein t^{**} mit $\Phi'(t^{**}) = 0$ bzw. $\nabla f(x_c + t^{**} p)^T p = \alpha \nabla f(x_c)^T p$. Offensichtlich erfüllt t^{**} auch (b).

Jetzt geben wir den Algorithmus zur Bestimmung des Intervalles $[t_{\min}, t_{\max}]$ an:

(0) Setze $t := 1$.

(1) Gelten (a) und (b), dann: STOP, Ziel erreicht, t ist gesuchte Schrittweite.

(2) Gilt (a) (und nicht (b)), dann:

Setze $t_{\min} := t$.

Solange ((a) gilt): Setze $t := 2t$.

Setze $t_{\max} := t$.

(3) Gilt (a) nicht, dann:

Setze $t_{\max} := t$.

Solange ((a) gilt nicht) oder ((b) gilt): Setze $t := \frac{1}{2}t$.

Setze $t_{\min} := t$.

Aus den Schleifen in (2) und (3) kommt man ganz offensichtlich nach endlich vielen Durchläufen heraus. Denn einerseits gilt (a) für alle hinreichend großen t nicht (andernfalls hätte man einen Widerspruch zur Kompaktheit der Niveaumenge) andererseits gilt (a) für alle hinreichend kleinen t, während (b) für alle hinreichend kleinen t nicht gilt. Man beachte, dass man in Schritt (3) daher auch einfach $t_{\min} := 0$ setzen könnte.

Nun kommt es darauf an, in endlich vielen Schritten einen Punkt $t \in [t_{\min}, t_{\max}]$ zu bestimmen, für welchen (a) und (b) erfüllt sind. Die primitivste Methode besteht darin, ein Bisektionsverfahren anzuwenden:

(i) Setze $t := \frac{1}{2}(t_{\min} + t_{\max})$.

(ii) Falls (a) (durch t) nicht erfüllt: $t_{\max} := t$, gehe nach (i).

Andernfalls:

> Falls (b) (durch t) erfüllt, dann: STOP, Ziel erreicht.
> Andernfalls: Setze $t_{\min} := t$ und gehe nach (i).

Wir müssen uns jetzt überlegen, dass dieses Intervallhalbierungsverfahren nach endlich vielen Schritten abbricht. Angenommen, dies wäre nicht der Fall. Dann existieren Folgen $\{t^k_{\min}\}$ und $\{t^k_{\max}\}$ mit $t^k_{\min} \nearrow t$ und $t^k_{\max} \searrow t$ sowie

$$f(x_c + t^k_{\min} p) \le f(x_c) + \alpha t^k_{\min} \nabla f(x_c)^T p, \qquad f(x_c + t^k_{\max} p) > f(x_c) + \alpha t^k_{\max} \nabla f(x_c)^T p$$

und

$$\nabla f(x_c + t^k_{\min} p)^T p < \beta \nabla f(x_c)^T p.$$

Wegen des Mittelwertsatzes existiert $\theta_k \in (0,1)$ mit

$$\begin{aligned} \alpha \nabla f(x_c)^T p &\le \frac{f(x_c + t^k_{\max} p) - f(x_c + t^k_{\min} p)}{t^k_{\max} - t^k_{\min}} \\ &= \nabla f(x_c + t^k_{\min} p + \underbrace{\theta_k (t^k_{\max} - t^k_{\min}) p}_{\to 0})^T p \\ &\to \nabla f(x_c + tp)^T p. \end{aligned}$$

Also ist $\alpha\nabla f(x_c)^T p \leq \nabla f(x_c + tp)^T p$. Andererseits folgt aus $\nabla f(x_c + t_{\min}^k p)^T p < \beta\nabla f(x_c)^T p$ mit $k \to \infty$, dass $\nabla f(x_c + tp)^T p \leq \beta\nabla f(x_c)^T p$. Wegen $\nabla f(x_c)^T p < 0$ und $\alpha < \beta$ hat man einen Widerspruch erreicht. Das obige Verfahren bricht also nach endlich vielen Schritten ab.

Sei $\phi(t) := f(x_c + tp)$. Im obigen Intervallhalbierungsverfahren hat man für das linke Intervallende $t_{\min}$ die Werte $\phi(t_{\min})$ und $\phi'(t_{\min})$, für das rechte Intervallende $t_{\max}$ den Wert $\phi(t_{\max})$ zur Verfügung. Statt des Mittelpunktes des Intervalls $[t_{\min}, t_{\max}]$ kann man nun die Minimalstelle der durch diese drei Werte bestimmten Parabel berechnen und diese, wenn sie hinreichend im Innern von $[t_{\min}, t_{\max}]$ liegt, als neuen Testwert t akzeptieren. Die gesuchte Parabel ist gegeben durch

$$q(s) = \phi(t_{\min}) + \phi'(t_{\min})(s - t_{\min}) + a_2(s - t_{\min})^2,$$

wobei

$$a_2 := \frac{\phi(t_{\max}) - (\phi(t_{\min}) + \phi'(t_{\min})(t_{\max} - t_{\min}))}{(t_{\max} - t_{\min})^2}.$$

Nun berücksichtige man, dass

$$\begin{aligned}
\phi(t_{\max}) &> f(x_c) + \alpha t_{\max}\nabla f(x_c)^T p \\
&= f(x_c) + \alpha t_{\min}\nabla f(x_c)^T p + \alpha(t_{\max} - t_{\min})\nabla f(x_c)^T p \\
&\geq \phi(t_{\min}) + \alpha(t_{\max} - t_{\min})\nabla f(x_c)^T p \\
&> \phi(t_{\min}) + \beta(t_{\max} - t_{\min})\nabla f(x_c)^T p \\
&> \phi(t_{\min}) + \phi'(t_{\min})(t_{\max} - t_{\min}).
\end{aligned}$$

Also ist $a_2 > 0$ und $q(\cdot)$ besitzt ein eindeutiges Minimum bei

$$t^* = t_{\min} - \frac{\phi'(t_{\min})}{2a_2} = t_{\min} + \Delta t$$

mit

$$\Delta t := -\frac{\phi'(t_{\min})(t_{\max} - t_{\min})^2}{2[\phi(t_{\max}) - (\phi(t_{\min}) + \phi'(t_{\min})(t_{\max} - t_{\min}))]} > 0.$$

Als neuen Testwert kann man dann z. B. setzen

$$t := \begin{cases} t^*, & t^* \in [t_{\min} + \tau(t_{\max} - t_{\min}), t_{\max} - \tau(t_{\max} - t_{\min})], \\ \frac{1}{2}(t_{\min} + t_{\max}), & t^* \notin [t_{\min} + \tau(t_{\max} - t_{\min}), t_{\max} - \tau(t_{\max} - t_{\min})]. \end{cases}$$

Hierbei ist $\tau \in (0, \frac{1}{2})$ gegeben, etwa $\tau := 0.1$. Damit ist eine mögliche Realisierung der Wolfe-Schrittweite beschrieben.

Jetzt wollen wir noch eine Matlab-Implementation der Wolfe-Schrittweite angeben. Wir folgen hierbei, auch bei den Bezeichnungen, eng der obigen Vorgehensweise. Insbesondere erfolgt die Berechnung der Schrittweite in zwei Phasen.

```
function [t,anz]=Wolfe(x_c,p,fun)
%Input-Parameter:
%           x_c     current iterate
%           p       search direction
```

```
%           fun    function to be minimized, accepts a vector
%                  x as argument, returns f=fun(x) objective
%                  function value resp [f,g]=fun(x) objective
%                  function and gradient at x.
%Output-Parameter:
%           t      Wolfe-stepsize
%           anz    number of calls to fun (if line search succeeds)
%**************************************************************************
alpha=0.0001; beta=0.9; tau=0.1;                %line search parameters
[f_c,g_c]=feval(fun,x_c);init_slope=g_c'*p;t=1.0;x_plus=x_c+t*p;anz=0;
[f_plus,g_plus]=feval(fun,x_plus);anz=anz+1;
new_slope=g_plus'*p;
if (f_plus<=f_c+alpha*t*init_slope)&(new_slope>=beta*init_slope)
   return
end;
if (f_plus<=f_c+alpha*t*init_slope)
   t_min=t;f_min=f_plus;g_min=g_plus;
   while (f_plus<=f_c+alpha*t*init_slope)
      t=2*t;x_plus=x_c+t*p;f_plus=feval(fun,x_plus);anz=anz+1;
   end;
   t_max=t;f_max=f_plus;
else
   t_max=t;f_max=f_plus;
   while (f_plus>f_c+alpha*t*init_slope)|(new_slope>=beta*init_slope)
      t=0.5*t;x_plus=x_c+t*p;[f_plus,g_plus]=feval(fun,x_plus);
      anz=anz+1;new_slope=g_plus'*p;
   end;
   t_min=t;f_min=f_plus;g_min=g_plus;
end; %end of phase I
success=0;
while success==0
   slope=g_min'*p;
   Delta_t=-slope*(t_max-t_min)^2/(2*(f_max-(f_min+slope*(t_max-t_min))));
   t_stern=t_min+Delta_t;
   if (t_min+tau*(t_max-t_min)<=t_stern)&(t_stern<=t_max-tau*(t_max-t_min))
      t=t_stern;
   else
      t=0.5*(t_min+t_max);
   end;
   x_plus=x_c+t*p;[f_plus,g_plus]=feval(fun,x_plus);anz=anz+1;
   if (f_plus>f_c+alpha*t*init_slope)
      t_max=t;f_max=f_plus;
   else
      new_slope=g_plus'*p;
      if (new_slope>=beta*init_slope)
          success=1;
      else
          t_min=t;f_min=f_plus;g_min=g_plus;
      end;
   end;
end;
```

So könnte eine Realisierung der Wolfe-Schrittweite in Matlab aussehen, wobei es uns mehr auf Durchsichtigkeit als auf Effizienz ankam. □

Beispiel: Wir wollen ein Beispiel bei P. SPELLUCCI (1993, S. 105) reproduzieren. Hier ist $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$ gegeben durch

$$f(x) := (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2,$$

ferner ist

$$x_c := (-4, -4)^T, \qquad p := (8, 48/7)$$

(von Spellucci werden die etwas ungewöhnlichen Werte nicht erläutert). Wir schreiben ein File `Myfun.m` mit dem Inhalt:

```
function [f,g]=Myfun(x);
f=(x(1)^2+x(2)-11)^2+(x(1)+x(2)^2-7)^2;
if nargout>1
g=[4*x(1)*(x(1)^2+x(2)-11)+2*(x(1)+x(2)^2-7);
    2*(x(1)^2+x(2)-11)+4*x(2)*(x(1)+x(2)^2-7)];
end;
```

Nach

```
>>x_c=[-4;-4];p=[8;48/7];[f_c,g_c]=Myfun(x_c);
>>[t,anz]=Wolfe(x_c,p,'Myfun');
```

jeweils nach dem Matlab-Prompt » erhalten wir $t = 0.0637$ als Wolfe-Schrittweite. Mit 9 Funktionsaufrufen wird das Intervall $[0.0039, 1]$ gefunden, danach sind in der Phase 2 noch 3 Funktionsaufrufe nötig, wobei jeweils die Minimalstellen der konstruierten Parabeln als Testwerte genommen werden. In Abbildung 3.1 geben wir $\phi(t) := f(x_c{+}tp)$

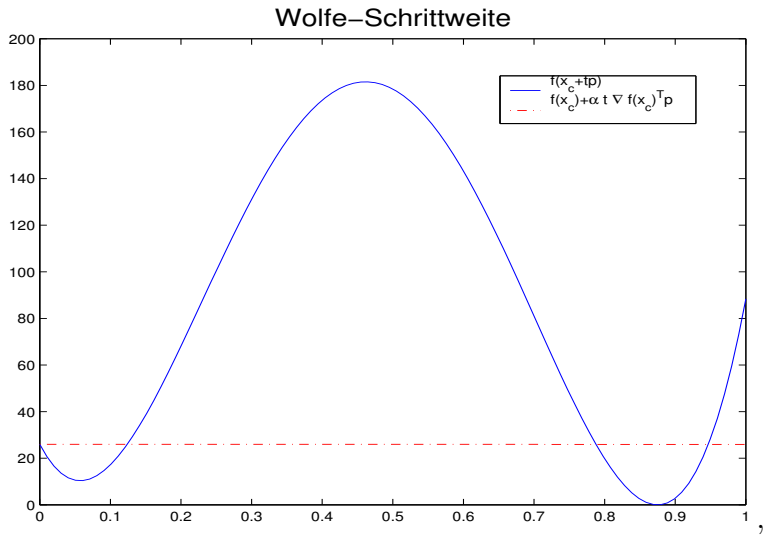

Abbildung 3.1: $f(x_c + tp)$ (blau) und $f(x_c) + \alpha t \nabla f(x_c)^T p$ (rot)

und die Gerade $f(x_c) + \alpha t \nabla f(x_c)^T p$ auf dem Intervall $[0, 1]$ an. Wir sehen, dass die von uns gefundene Wolfe-Schrittweite in der Nähe des ersten lokalen Minimums von f liegt. □

Wie schon früher erwähnt spielt die Schrittweite $t = 1$ oft eine besondere Rolle. Dies ist insbesondere dann der Fall, wenn $p \approx -\nabla^2 f(x_c)^{-1} \nabla f(x_c)$, die Abstiegsrichtung also nahe bei der sogenannten *Newton-Richtung* liegt. Denn dann wird man hoffen, in der Nähe einer lokalen Lösung x^* von einem gedämpften Verfahren (die neue Näherung ist $x_+ = x_c + tp$ mit einer geeigneten Schrittweite $t > 0$) zu einem ungedämpften Verfahren (hier ist $x_+ = x_c + p$, die Schrittweite also $t = 1$) übergehen zu können. Bei der *Armijo-Schrittweite* testet man zunächst, ob bei einem vorgegebenen $\alpha \in (0, \frac{1}{2})$ die Ungleichung

$$f(x_c + tp) \leq f(x_c) + \alpha t \nabla f(x_c)^T p,$$

also die Bedingung für hinreichenden Abstieg, für $t := 1$ erfüllt ist. Ist dies der Fall, so wird $t = 1$ als Schrittweite akzeptiert. Andernfalls wird t "kontrolliert verkleinert" und die Bedingung für hinreichenden Abstieg mit der neuen Schrittweite erneut getestet. Sobald diese erfüllt ist (wir wissen, dass sie für alle hinreichend kleinen $t > 0$ erfüllt ist) wird die entsprechende Schrittweite akzeptiert.

Etwas genauer sieht die Berechnung der Armijo-Schrittweite folgendermaßen aus.

- Seien $\alpha \in (0, \frac{1}{2})$ und $0 < l \leq u < 1$ gegeben.
- Setze $\rho_0 := 1$.
- Für $j = 0, 1, \ldots$:
  - Falls $f(x_c + \rho_j p) \leq f(x_c) + \alpha \rho_j \nabla f(x_c)^T p$, dann: $t := \rho_j$, STOP.
  - Andernfalls: Wähle $\rho_{j+1} \in [l\rho_j, u\rho_j]$.

Bemerkung: Auf eine Matlab-Implementation der Armijo-Schrittweite werden wir später ausführlich eingehen. Hier sollen nur einige wenige Bemerkungen zu einer möglichen Implementation gemacht werden.

Ist z. B. $l = u =: \rho$, so ist die Armijo-Schrittweite durch $t = \rho^j$ gegeben, wobei j die kleinste nichtnegative ganze Zahl mit

$$f(x_c + \rho^j p) \leq f(x_c) + \alpha \rho^j \nabla f(x_c)^T p$$

ist. Man spricht dann auch von einem "backtracking line search". Von S. P. Han (1981)[1] wird

$$\rho_{j+1} := \max(0.1\rho_j, \rho_j^*) \qquad \text{mit} \quad \rho_j^* := -\frac{\rho_j^2 \nabla f(x_c)^T p}{2[f(x_c + \rho_j p) - (f(x_c) + \rho_j \nabla f(x_c)^T p)]}$$

gesetzt. Ist $f(x_c + \rho_j p) > f(x_c) + \alpha \rho_j \nabla f(x_c)^T p$, die zu testende Ungleichung im j-ten Schritt also nicht erfüllt, so zeigt eine einfache Rechnung, dass

$$0.1\rho_j \leq \rho_{j+1} \leq \underbrace{\frac{1}{2(1-\alpha)}}_{<1} \rho_j.$$

[1] S. P. Han (1981) "Variable metric methods for minimizing a class of nondifferentiable functions." Mathematical Programming 20, 1–13.

Die von Han benutzte Schrittweite fällt also mit

$$l := 0.1, \qquad u := \frac{1}{2(1-\alpha)}$$

unter obiges Konzept. Man rechnet leicht nach, dass bei ρ_j^* gerade das Minimum des quadratischen Polynoms q liegt, das den Interpolationsbedingungen $q(0) = f(x_c)$, $q'(0) = \nabla f(x_c)^T p$ und $q(\rho_j) = f(x_c + \rho_j p)$ genügt. Auf eine etwas raffiniertere Version gehen wir später ein. □

Im folgenden Satz wird die durch eine Armijo-schrittweite erreichte Verminderung der Zielfunktion nach unten abgeschätzt.

Satz 1.4 *Die Zielfunktion f von (P) genüge den Voraussetzungen (V) (a)–(c). Sei $x_c \in L_0$ keine stationäre Lösung von (P) und p eine Abstiegsrichtung für f in x. Seie $\alpha \in (0, \frac{1}{2})$, $0 < l \le u < 1$ und hiermit eine Schrittweite $t = \rho_j$ gegeben. Dann existiert eine Konstante $\theta > 0$, die nur von α, γ sowie l und u, nicht aber von x_c oder p abhängt, mit*

$$(**) \qquad f(x_c) - f(x_c + tp) \ge \theta \min\Big[-\nabla f(x_c)^T p, \Big(\frac{\nabla f(x_c)^T p}{\|p\|_2}\Big)^2\Big].$$

Beweis: Offenbar muss die zu testende Ungleichung nach endlich vielen Schritten erfüllt sein. Ist $j = 0$ bzw. $t = \rho_0 = 1$, so ist $f(x_c + tp) \le f(x_c) + \alpha \nabla f(x_c)^T p$. Ist dagegen $j > 0$, so gelten mit $s := \rho_{j-1}$ zwei Ungleichungen, nämlich:

$$f(x_c + tp) \le f(x_c) + \alpha t \nabla f(x_c)^T p, \qquad f(x_c + sp) > f(x_c) + \alpha s \nabla f(x_c)^T p.$$

Ferner ist $ls \le t$. Mit $\hat{t}$ wie in Lemma 1.1 machen wir eine Fallunterscheidung. Für $s \le \hat{t}$ ist

$$f(x_c) + \alpha s \nabla f(x_c)^T p < f(x_c + sp) \le f(x_c) + s \nabla f(x_c)^T p + s^2 \frac{\gamma}{2} \|p\|_2^2,$$

daher

$$-\frac{2l(1-\alpha)}{\gamma} \frac{\nabla f(x_c)^T p}{\|p\|_2^2} \le ls \le t$$

und folglich

$$f(x_c) - f(x_c + tp) \ge -\alpha t \nabla f(x_c)^T p \ge \frac{2\alpha l(1-\alpha)}{\gamma} \Big(\frac{\nabla f(x_c)^T p}{\|p\|_2}\Big)^2.$$

Ist dagegen $s > \hat{t}$, so ist wiederum wegen Lemma 1.1

$$-\frac{2l \nabla f(x_c)^T p}{\gamma \|p\|_2^2} \le l\hat{t} < ls \le t$$

und daher

$$f(x_c) - f(x_c + tp) \ge -\alpha t \nabla f(x_c)^T p \ge \frac{2\alpha l}{\gamma} \Big(\frac{\nabla f(x_c)^T p}{\|p\|_2}\Big)^2.$$

Mit

$$\theta := \alpha \min(1, 2l(1-\alpha)/\gamma)$$

ist die Aussage des Satzes bewiesen. □ □

Bemerkungen: Die Voraussetzung $\alpha \in (0, \frac{1}{2})$ bei der Definition der Wolfe- und der Armijo-Schrittweite könnte durch $\alpha \in (0, 1)$ ersetzt werden und die Sätze 1.3 und 1.4 würden immer noch gelten. Bei dem Nachweis der superlinearen Konvergenz von Newton- und Quasi-Newton-Verfahren wird klar werden, weshalb $\alpha \in (0, \frac{1}{2})$ vorausgesetzt wird.

Schrittweiten t, für die eine Aussage wie bei der exakten Schrittweite oder der Wolfe-Schrittweite gemacht werden kann, für die also unter den Voraussetzungen (V) (a)–(c) eine von x_c und p unabhängige Konstante $\theta > 0$ mit

$$(*) \qquad f(x_c) - f(x_c + tp) \geq \theta \Big(\frac{\nabla f(x_c)^T p}{\|p\|_2} \Big)^2$$

existiert, wurden von W. Warth, J. Werner (1977)[2] *effizient* genannt. Entsprechend werden Schrittweiten t, wie z. B. die Armijo-Schrittweite, zu denen es unter den Voraussetzungen (V) (a)–(c) eine von x_c und p unabhängige Konstante $\theta > 0$ mit

$$(**) \qquad f(x_c) - f(x_c + tp) \geq \theta \min \Big[-\nabla f(x_c)^T p, \Big(\frac{\nabla f(x_c)^T p}{\|p\|_2} \Big)^2 \Big]$$

gibt, von P. Kosmol (1989, S. 92)[3] *semi-effizient* genannt. Die Beziehungen $(*)$ und $(**)$ stellen sich als fundamental bei der Konvergenzanalyse heraus. Etwas vereinfacht gesagt: Hat man für eine Schrittweitenstrategie $(*)$ bzw. $(**)$ bewiesen, so kann man für die Konvergenzanalyse vergessen, wodurch die Schrittweitenstrategie spezifiziert ist, alleine die Richtungsstrategie spielt danach noch eine Rolle. □

Bemerkung: Wir wollen nun auf eine Matlab-Implementation der Armijo-Schrittweite eingehen. Wir orientieren uns hierbei an J. E. Dennis, R. B. Schnabel (1983, S. 325 ff.) und C. T. Kelley (1999, S. 43 ff.). Hierbei wird eine quadratische (im ersten Schritt) bzw. kubische Interpolation (in allen weiteren Schritten) benutzt. Mit $\phi(t) := f(x_c + tp)$ sind am Anfang (dann wird die Schrittweite $t := 1$ getestet) $\phi(0) = f(x_c)$, $\phi(1) = f(x_c + p)$ und $\phi'(0) = \nabla f(x_c)^T p$ bekannt. Dann ist

$$q(s) = \phi(0) + \phi'(0)s + (\phi(1) - \phi(0) - \phi'(0))s^2$$

das durch diese Werte bestimmte quadratische Polynom. Da wir davon ausgehen, dass die Schrittweite $t = 1$ keinen hinreichenden Abstieg garantiert, ist

$$\phi(1) > \phi(0) + \alpha\phi'(0) = \phi(0) + \phi'(0) + \underbrace{(\alpha - 1)\phi'(0)}_{>0} > \phi(0) + \phi'(0).$$

[2] W. Warth, J. Werner (1977) "Effiziente Schrittweitenfunktionen bei unrestringierten Optimierungsaufgaben." Computing 19, 59–72.

[3] P. Kosmol (1989) *Methoden numerischer Behandlung nichtlinearer Gleichungen und Optimierungsaufgaben.* B. G. Teubner, Stuttgart.

Folglich hat q bei

$$t_{\text{temp}} := -\frac{\phi'(0)}{2[\phi(1) - \phi(0) - \phi'(0)]} > 0$$

ein Minimum. Nach dem ersten Schritt sind neben $\phi(0)$ und $\phi'(0)$ noch $\phi(t_{\text{prev}})$ und $\phi(t)$ bekannt, wobei $0 < t < t_{\text{prev}}$ die beiden letzten getesteten Schrittweiten sind. Das entsprechende kubische Interpolationspolynom ist

$$q(s) = \phi(0) + \phi'(0)s + a_1 s^2 + a_2 s^3,$$

wobei (a_1, a_2) so zu bestimmen sind, dass $q(t) = \phi(t)$ und $q(t_{\text{prev}}) = \phi(t_{\text{prev}})$, was auf das lineare Gleichungssystem

$$\begin{pmatrix} t^2 & t^3 \\ t_{\text{prev}}^2 & t_{\text{prev}}^3 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \phi(t) - \phi(0) - \phi'(0)t \\ \phi(t_{\text{prev}}) - \phi(0) - \phi'(0)t_{\text{prev}} \end{pmatrix}$$

(mit einer nichtsingulären Koeffizientenmatrix) führt. Ist hier $a_2 = 0$ (dann ist $a_1 > 0$), so ist q eine Parabel mit dem eindeutigen Minimum

$$t_{\text{temp}} := -\frac{\phi'(0)}{2a_1}.$$

Andernfalls besitzt $q'(\cdot)$ die beiden Nullstellen

$$s_{\pm} = \frac{1}{3a_2}\left[-a_1 \pm \sqrt{a_1^2 - 3a_2\phi'(0)}\right].$$

Für $a_2 \neq 0$ setzen wir daher (eine Diskussion, ob es sich hierbei wirklich um ein lokales Minimum handelt und ob der so berechnete Wert überhaupt reell ist, wollen wir uns ersparen[4])

$$t_{\text{temp}} := \frac{1}{3a_2}\left[-a_1 + \sqrt{a_1^2 - 3a_2\phi'(0)}\right].$$

Die neue Schrittweite t_plus wird dann bestimmt als

$$t_+ := \begin{cases} 0.5t, & t_{\text{temp}} > 0.5t, \\ 0.1t, & t_{\text{temp}} < 0.1t, \\ t_{\text{temp}}, & \text{sonst.} \end{cases}$$

Eine auf diesen Überlegungen basierende Matlab-Implementation könnte dann folgendermaßen aussehen:

```
function [t,anz]=Armijo(x_c,p,fun)
%Input-Parameter:
%           x_c     current iterate
%           p       search direction
%           fun     function to be minimized, accepts a vector
%                   x as argument, returns f=fun(x) objective
%                   function value resp [f,g]=fun(x) objective
```

[4] Auch J. E. Dennis, R. B. Schnabel (1983, S. 128) sind hier sehr kurz.

```
%                  function and gradient at x.
%Output-Parameter:
%           t      Armijo-stepsize
%           anz    number of calls to fun
%*********************************************************************
alpha=0.0001;        %line search parameter
[f_c,g_c]=feval(fun,x_c);
init_slope=g_c'*p;
t=1;x_plus=x_c+t*p;f_plus=feval(fun,x_plus);anz=1;
while (f_plus>f_c+alpha*t*init_slope)
    if (t==1)
        t_temp=-init_slope/(2*(f_plus-f_c-init_slope));
    else
        A=[t^2 t^3;t_prev^2 t_prev^3];
        b=[f_plus-f_c-t*init_slope;f_prev-f_c-t_prev*init_slope];
        a=A\b;
        if (a(2)==0)
             t_temp=-init_slope/(2*a(1));
        else
             diskr=a(1)^2-3*a(2)*init_slope;
             t_temp=(-a(1)+sqrt(diskr))/(3*a(2));
        end;
    end;
    t_prev=t;f_prev=f_plus;
    t=max(0.1*t,min(0.5*t,t_temp));
    x_plus=x_c+t*p;f_plus=feval(fun,x_plus);anz=anz+1;
end;
```

□

Beispiel: Wir betrachten dasselbe Beispiel wie zur Wolfe-Schrittweite, d. h. $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$ sei definiert durch

$$f(x) := (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2,$$

ferner sei

$$x_c := (-4, -4)^T, \qquad p := (8, 48/7).$$

Mit der oben definierten Funktion `Myfun` erhalten wir nach

```
>>x_c=[-4;-4];p=[8;48/7];[f_c,g_c]=Myfun(x_c);
>>[t,anz]=Armijo(x_c,p,'Myfun');
```

jeweils nach dem Matlab-Prompt » mit 3 Funktionsauswertungen die Schrittweite $t = 0.1036$. □

3.1.2 Konvergenz des Modellalgorithmus bei glatter Zielfunktion

Der folgende Satz gibt unter verhältnismäßig schwachen Voraussetzungen an die Zielfunktion f sowie an die benutzten Abstiegsrichtungen ein, wie man nicht anders erwarten kann, schwaches Konvergenzergebnis an.

Satz 1.5 *Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Voraussetzungen (V) (a)–(c). Als Schrittweite im Modellalgorithmus verwende man $t_k := t^*(x_k, p_k)$ (exakte Schrittweite), $t_k := t_W(x_k, p_k)$ (Wolfe-Schrittweite) oder $t_k := t_A(x_k, p_k)$ (Armijo-Schrittweite). Zur Abkürzung setze man $g_k := \nabla f(x_k)$. Ferner wird vorausgesetzt:*

1. *Es existiert eine Konstante $\sigma > 0$ mit*

$$-\frac{g_k^T p_k}{\|g_k\|_2 \, \|p_k\|_2} \geq \sigma, \qquad k = 0, 1, \ldots .$$

2. *Es existiert eine Konstante $\tau > 0$ mit*

$$\|p_k\|_2 \geq \tau \|g_k\|_2, \qquad k = 0, 1, \ldots .$$

Dann gilt: Jeder Häufungspunkt der durch den Modellalgorithmus mit Abstiegsrichtungen p_k erzeugten Folge $\{x_k\}$ ist eine stationäre Lösung von (P). Besitzt (P) genau eine stationäre Lösung x^* in der Niveaumenge L_0, so konvergiert die gesamte Folge $\{x_k\}$ gegen x^*.

Beweis: Wegen der Sätze 1.2, 1.3, 1.4 existiert eine Konstante $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \geq \theta \min\Big[-g_k^T p_k, \Big(\frac{g_k^T p_k}{\|p_k\|_2}\Big)^2\Big], \qquad k = 0, 1, \ldots .$$

Wegen der Voraussetzungen 1. und 2. ist daher

$$f(x_k) - f(x_{k+1}) \geq \theta\sigma \min(\tau, \sigma) \, \|g_k\|_2^2, \qquad k = 0, 1, \ldots .$$

Da $\{f(x_k)\}$ eine monoton fallende, nach unten beschränkte Folge ist, konvergiert damit $\{g_k\}$ gegen den Nullvektor.

Ist $x^* \in L_0$ ein Häufungspunkt von $\{x_k\}$, so ist x^* Limes einer konvergenten Teilfolge $\{x_{k(j)}\} \subset \{x_k\}$. Da $\{g_{k(j)}\}$ einerseits gegen $\nabla f(x^*)$ und andererseits gegen 0 konvergiert, ist $\nabla f(x^*) = 0$, also x^* eine stationäre Lösung von (P).

Nun besitze (P) genau eine stationäre Lösung x^* in L_0. Angenommen, $\{x_k\}$ konvergiert nicht gegen x^*. Dann existiert ein $\epsilon > 0$ und eine Teilfolge $\{x_{k(j)}\} \subset \{x_k\}$ mit $\|x_{k(j)} - x^*\|_2 \geq \epsilon$, $j = 1, 2, \ldots$. Da L_0 kompakt ist, kann aus $\{x_{k(j)}\}$ eine gegen ein $\hat{x} \in L_0$ konvergente Teilfolge ausgewählt werden. Als Häufungspunkt der Folge $\{x_k\}$ ist $\hat{x}$ wegen des gerade eben bewiesenen ersten Teils eine stationäre Lösung von (P). Wegen $\|\hat{x} - x^*\|_2 \geq \epsilon$ ist $\hat{x} \neq x^*$. Dies ist ein Widerspruch dazu, dass x^* die einzige stationäre Lösung von (P) ist. □ □

Bemerkungen: Man erkennt, dass in den Beweis des Konvergenzsatzes nicht die Definition der jeweiligen Schrittweiten eingeht, sondern lediglich die Folgerungen $(*)$ bzw. $(**)$ aus den Sätzen 1.2, 1.3 und 1.4. Klar ist auch, dass man auf die zweite Voraussetzung in Satz 1.5, also die Existenz einer Konstanten $\tau > 0$ mit

$$\|p_k\|_2 \geq \tau \|g_k\|_2, \qquad k = 0, 1, \ldots ,$$

(hier und im folgenden wird die Abkürzung $g_k := \nabla f(x_k)$ benutzt), verzichten kann, wenn nur die exakte Schrittweite oder die Wolfe-Schrittweite (oder eine andere effiziente Schrittweite) verwendet wird.

Eine Folge von Abstiegsrichtungen $\{p_k\}$ wird *gradientenähnlich* genannt, wenn die erste Voraussetzung in Satz 1.5 erfüllt ist, wenn es also eine Konstante $\sigma > 0$ mit

$$-\frac{g_k^T p_k}{\|g_k\|_2 \, \|p_k\|_2} \geq \sigma, \qquad k = 0, 1, \ldots,$$

gibt. Diese Voraussetzung besagt, dass der Winkel zwischen $-g_k$ und p_k gleichmäßig kleiner als der rechte Winkel sein muss. Für $p_k = -g_k$ (dann spricht man vom Gradientenverfahren) kann $\sigma = 1$ gewählt werden. □

Beispiel: Man betrachte den Modellalgorithmus mit einer Richtungsfolge $\{p_k\}$, wobei $p_k = -H_k g_k$, $k = 0, 1, \ldots$, mit einer symmetrischen und positiv definiten Matrix $H_k \in \mathbb{R}^{n \times n}$. Dann ist

$$-\frac{g_k^T p_k}{\|g_k\|_2 \, \|p_k\|_2} = \frac{g_k^T H_k g_k}{\|g_k\| \, \|H_k g_k\|_2} \geq \frac{1}{\sqrt{\|H_k\|_2 \, \|H_k^{-1}\|_2}} = \frac{1}{\sqrt{\kappa_2(H_k)}},$$

wobei $\kappa_2(\cdot)$ die Kondition einer Matrix bezüglich der Spektralnorm bedeutet. Die erste Voraussetzung in Satz 1.5 ist daher erfüllt, wenn $\{\kappa_2(H_k)\}$ beschränkt ist. Hierbei haben wir benutzt: Ist $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit, bezeichnet ferner $A^{1/2}$ die symmetrische und positiv definite Quadratwurzel aus A, so ist für beliebiges $x \in \mathbb{R}^n \setminus \{0\}$:

$$\frac{x^T A x}{\|x\|_2 \, \|Ax\|_2} = \frac{\|A^{1/2} x\|_2^2}{\|A^{-1/2} A^{1/2} x\|_2 \, \|A^{1/2} A^{1/2} x\|_2} \geq \frac{1}{\kappa_2(A^{1/2})} = \frac{1}{\sqrt{\kappa_2(A)}}.$$

Wegen

$$\|p_k\|_2 = \|H_k g_k\|_2 \geq \frac{1}{\|H_k^{-1}\|_2} \|g_k\|_2$$

ist die zweite Voraussetzung in Satz 1.5 erfüllt, wenn $\{\|H_k^{-1}\|_2\}$ beschränkt ist. Beide Voraussetzungen gelten, wenn $\{H_k\}$ eine Folge symmetrischer, gleichmäßig positiv definiter und beschränkter Matrizen ist, wenn es also Konstanten $0 < c \leq d$ mit

$$c \, \|z\|_2^2 \leq z^T H_k z \leq d \, \|z\|_2^2 \qquad \text{für alle } z \in \mathbb{R}^n, \; k = 0, 1, \ldots$$

gibt. Dies wiederum ist gleichbedeutend mit der Beschränktheit der Folgen $\{\|H_k\|_2\}$ und $\{\|H_k^{-1}\|_2\}$. Insbesondere ist dies natürlich für $H_k = I$, also das Gradientenverfahren der Fall. □

Beispiel: Das Gradientenverfahren, kombiniert mit einer der angegebenen Schrittweitenstrategien (wobei diese natürlich auch von Schritt zu Schritt geändert werden kann), ist global konvergent bei der Rosenbrock-Funktion

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2.$$

Denn diese besitzt genau einen stationären Punkt, nämlich $x^* = (1,1)$, ferner sind offensichtlich die Voraussetzungen (V) (a)–(c) erfüllt. Natürlich bedeutet "Konvergenz" nicht automatisch "gute Konvergenz". Dies wollen wir uns durch ein Zahlenbeispiel klar machen. Hierzu schreiben wir zunächst ein Function-M-File `Rosenbrock.m` mit dem Inhalt:

```
function [f,g]=Rosenbrock(x);
f=100*(x(2)-x(1)^2)^2+(1-x(1))^2;
if nargout>1
  g=[-400*x(1)*(x(2)-x(1)^2)-2*(1-x(1));200*(x(2)-x(1)^2)];
end;
```

Mit dem kleinen Programm

```
x=[1.2;1];x_stern=[1;1];
A=[];[f,g]=Rosenbrock(x);
for k=1:1001
   p=-g;
   t=Wolfe(x,p,'Rosenbrock');
   x=x+t*p;[f,g]=Rosenbrock(x);
   if (mod(k,100)==1)
       A=[A;k,norm(g),norm(x-x_stern),f];
   end;
end;
disp(A);
```

erhalten wir die folgenden Ergebnisse (wir benutzen `format short g`):

k	$\|\nabla f(x_k)\|_2$	$\|x_k - x^*\|_2$	$f(x_k)$
1	133.04	0.16376	10.491
101	0.02495	0.055729	0.00060819
201	0.022234	0.050273	0.00049586
301	0.019705	0.044999	0.00039801
401	0.017349	0.039944	0.00031416
501	0.015165	0.035149	0.00024367
601	0.013152	0.03065	0.00018557
701	0.011310	0.026478	0.0001387
801	0.0096205	0.022643	0.00010156
901	0.0080655	0.019089	$7.2275e-05$
1001	0.0066686	0.015852	$4.9895e-05$

Man erkennt, dass die Konvergenzgeschwindigkeit schlecht ist. □

Nun betrachten wir noch die Anwendung des Modellalgorithmus bei einer glatten, gleichmäßig konvexen Zielfunktion. Hierzu fassen wir zunächst einige Hilfsmittel in dem folgenden Lemma zusammen.

Lemma 1.6 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die folgenden Konvexitäts- und Glattheitsvoraussetzungen an die Zielfunktion f seien erfüllt:*

(K) (a) *Mit einem gegebenen $x_0 \in \mathbb{R}^n$ (Startwert eines Iterationsverfahrens) ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \le f(x_0)\}$ konvex.*

(b) *Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar und auf L_0 gleichmäßig konvex, d. h. es existiert eine Konstante $c > 0$ mit*

$$\frac{c}{2}\|y - x\|_2^2 + \nabla f(x)^T(y - x) \le f(y) - f(x) \qquad \text{für alle } x, y \in L_0.$$

.

(c) *Der Gradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig, d. h. es existiert eine Konstante $\gamma > 0$ mit*

$$\|\nabla f(x) - \nabla f(y)\| \le \gamma \|x - y\| \qquad \text{für alle } x, y \in L_0.$$

Dann ist die Niveaumenge L_0 kompakt, (P) besitzt daher eine globale Lösung x^*, diese liegt in L_0 und ist die einzige stationäre Lösung von (P) in L_0. Ferner gilt die Fehlerabschätzung

$$\frac{c}{2}\|x - x^*\|_2^2 \le f(x) - f(x^*) \le \frac{1}{2c}\|\nabla f(x)\|_2^2 \qquad \text{für alle } x \in L_0.$$

Beweis: Die Niveaumenge L_0 ist abgeschlossen. Für alle $x \in L_0$ ist wegen der gleichmäßigen Konvexität von f ferner

$$\frac{c}{2}\|x - x_0\|_2^2 + \nabla f(x_0)^T(x - x_0) \le f(x) - f(x_0) \le 0$$

und daher mit Hilfe der Cauchy-Schwarzschen Ungleichung

$$L_0 \subset \left\{x \in \mathbb{R}^n : \|x - x_0\|_2 \le \frac{2}{c}\|\nabla f(x_0)\|_2\right\}.$$

Insgesamt ist L_0 kompakt, die auf L_0 stetige Funktion f nimmt auf L_0 ihr (globales) Minimum an. Da eine globale Lösung von (P) nicht außerhalb von L_0 liegen kann, ist die Existenz einer globalen Lösung $x^* \in L_0$ bewiesen. Natürlich ist $\nabla f(x^*) = 0$, also x^* auch eine stationäre Lösung von (P). Wir zeigen nun noch die behaupteten Abschätzungen, aus denen insbesondere die Eindeutigkeit einer stationären Lösung von (P) in L_0 folgt.

Die erste Ungleichung folgt direkt aus der vorausgesetzten gleichmäßigen Konvexität, indem man $y = x$ und $x = x^*$ setzt. Bei festem $x \in L_0$ ist

$$-\frac{1}{2c}\|\nabla f(x)\|_2^2 \le \frac{c}{2}\|x^* - x\|_2^2 + \nabla f(x)^T(x^* - x) \le f(x^*) - f(x).$$

Dies erkennt man daran, dass die Aufgabe

$$\text{Minimiere} \quad f_x(p) := \frac{c}{2}\|p\|_2^2 + \nabla f(x)^T p, \quad p \in \mathbb{R}^n,$$

die eindeutige Lösung $p^* := -(1/c)\nabla f(x)$ besitzt. Insgesamt ist der Satz damit bewiesen. □

□

Im folgenden Satz wird bei gleichmäßig konvexer Zielfunktion eine hinreichende Konvergenzbedingung für den Modellalgorithmus angegeben.

Satz 1.7 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Voraussetzungen (K) (a)–(c) aus Lemma 1.6 seien erfüllt. Zur Lösung von (P) betrache man den Modellalgorithmus mit Abstiegsrichtungen p_k und Schrittweiten $t_k := t^*(x_k, p_k)$ (exakte Schrittweite), $t_k := t_W(x_k, p_k)$ (Wolfe-Schrittweite) oder $t_k := t_A(x_k, p_k)$ (Armijo-Schrittweite). Zur Abkürzung sei $g_k := \nabla f(x_k)$ gesetzt. Schließlich sei*

$$\delta_k := \begin{cases} \min\Big[-\dfrac{g_k^T p_k}{\|g_k\|_2^2}, \Big(\dfrac{g_k^T p_k}{\|g_k\|_2 \, \|p_k\|_2}\Big)^2\Big], & \text{falls} \quad t_k = t_A(x_k, p_k), \\ \Big(\dfrac{g_k^T p_k}{\|g_k\|_2 \, \|p_k\|_2}\Big)^2, & \text{falls} \quad t_k = t^*(x_k, p_k), t_k = t_W(x_k, p_k). \end{cases}$$

Dann gilt:

1. *Ist*
$$\sum_{j=0}^{\infty} \delta_j = \infty,$$
so konvergiert die durch den Modellalgorithmus erzeugte Folge $\{x_k\}$ gegen die eindeutige (globale) Lösung x^* von (P).

2. *Existiert ein $\delta > 0$ mit*
$$\delta \le \frac{1}{k+1} \sum_{j=0}^{k} \delta_j, \qquad k = 0, 1, \ldots,$$
so konvergiert die Folge $\{x_k\}$ R-linear gegen x^*, d. h. es existieren Konstanten $C > 0$ und $q \in (0,1)$ mit $\|x_k - x^*\|_2 \le Cq^k$, $k = 0, 1, \ldots$.

Beweis: Wegen der Sätze 1.2, 1.3 und 1.4 sowie der Definition der δ_k existiert eine von k unabhängige Konstante $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \ge \theta \delta_k \, \|g_k\|_2^2 \ge 2c\theta\delta_k \, [f(x_k) - f(x^*)], \qquad k = 0, 1, \ldots,$$

wobei auch noch die Fehlerabschätzung aus Lemma 1.6 benutzt wurde. Daher ist

$$\begin{aligned} 0 \;\le\; f(x_{k+1}) - f(x^*) \;&\le\; (1 - 2c\theta\delta_k)[f(x_k) - f(x^*)] \\ &\le\; \prod_{j=0}^{k} (1 - 2c\theta\delta_j)[f(x_0) - f(x^*)] \\ &\le\; \exp\Big(-2c\theta \sum_{j=0}^{k} \delta_j\Big)[f(x_0) - f(x^*)]. \end{aligned}$$

Wegen $\sum_{j=0}^{\infty} \delta_j = \infty$ konvergiert $\{f(x_k)\}$ gegen $f(x^*)$. Wiederum wegen der Fehlerabschätzung in Lemma 1.6 folgt die Konvergenz von $\{x_k\}$ gegen x^*.

Existiert ein $\delta > 0$ mit $\delta(k+1) \le \sum_{j=0}^{k} \delta_j$, $k = 0, 1, \ldots$, so ist

$$f(x_k) - f(x^*) \le \exp\Big(-2c\theta \sum_{j=0}^{k-1} \delta_j\Big)[f(x_0) - f(x^*)] \le \exp(-2c\theta\delta k)[f(x_0) - f(x^*)].$$

Mit Hilfe von $\|x_k - x^*\|_2 \leq \{2[f(x_k) - f(x^*)]/c\}^{1/2}$ (siehe Lemma 1.6) folgt daher

$$\|x_k - x^*\|_2 \leq \left\{\frac{2[f(x_0) - f(x^*)]}{c}\right\}^{1/2} \exp(-c\theta\delta)^k, \qquad k = 0, 1, \ldots.$$

Der Satz ist damit bewiesen. □ □

Bemerkung: Die Schrittweitenstrategien gingen wiederum nur dadurch ein, dass die Aussagen der Sätze 1.2–1.4 benutzt wurden.

Wird im Modellalgorithmus unter den Voraussetzungen von Satz 1.7 stets die exakte oder die Wolfe-Schrittweite (oder eine andere effiziente Schrittweite) gewählt, so ist

$$\delta_k = \left(\frac{g_k^T p_k}{\|g_k\|_2 \, \|p_k\|_2}\right)^2, \qquad k = 0, 1, \ldots.$$

Die Bedingung $\sum_{j=0}^{\infty} \delta_j = \infty$ besagt, dass der Winkel zwischen $-g_k$ und p_k sich zwar einem rechten Winkel annähern, dies aber nicht zu schnell geschehen darf. □

3.1.3 Aufgaben

1. Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genöge den Bedingungen (V) (a)–(c) in Unterabschnitt 3.1.1. Sei $x_c \in L_0$ keine stationäre Lösung von (P) und $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x_c, d. h. $\nabla f(x_c)^T p < 0$. Seien $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ gegeben. Hiermit definiere man

$$T_{SW}(x_c, p) := \left\{ t > 0 : \begin{array}{rcl} f(x_c + tp) & \leq & f(x_c) + \alpha t \nabla f(x_c)^T p, \\ |\nabla f(x_c + tp)^T p| & \leq & -\beta \nabla f(x_c)^T p \end{array} \right\},$$

die Menge der *strengen Wolfe-Schrittweiten*. Man zeige:

   (a) Es ist $T_{SW}(x_c, p) \neq \emptyset$.

   (b) Es existiert eine von x_c und p unabhängige Konstante $\theta > 0$ mit

$$f(x_c) - f(x_c + tp) \geq \theta \left(\frac{\nabla f(x_c)^T p}{\|p\|_2}\right)^2 \qquad \text{für alle } t \in T_{SW}(x_c, p).$$

2. Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Bedingungen (V) (a)–(c) in Unterabschnitt 3.1.1. Sei $x_c \in L_0$ keine stationäre Lösung von (P) und $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x_c, d. h. $\nabla f(x_c)^T p < 0$. Seien $\alpha \in (0, 1)$, $\sigma > 0$ und $\rho \in (0, 1)$ gegeben. Folgendermaßen bestimme man eine Schrittweite $t = t(x_c, p)$:

   - Wähle $\tau \geq -\sigma \nabla f(x_c)^T p / \|p\|_2^2$, bestimme die kleinste nichtnegative ganze Zahl j mit
$$f(x_c + \tau\rho^j p) \leq f(x_c) + \alpha\tau\rho^j \nabla f(x_c)^T p$$
   und setze $t := \tau\rho^j$.

   Man zeige, dass eine von (x_c, p) unabhängige Konstante $\theta > 0$ mit

$$f(x_c) - f(x_c + tp) \geq \theta \left(\frac{\nabla f(x_c)^T p}{\|p\|_2}\right)^2$$

   existiert.

3. Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Bedingungen (V) (a)–(c) in Unterabschnitt 3.1.1. Sei $x_c \in L_0$ keine stationäre Lösung von (P) und $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x_c, d. h. $\nabla f(x_c)^T p < 0$. Bei vorgegebenem $\alpha \in (0, \frac{1}{2})$ definiere man

$$T_G(x_c, p) := \{t > 0 : f(x_c) + (1-\alpha)t\nabla f(x_c)^T p \le f(x_c + tp) \le f(x_c) + \alpha t \nabla f(x_c)^T p\}$$

(Menge der *Goldstein-Schrittweiten*). Analog zu Satz 1.3 zeige man:

(a) Es ist $T_G(x_c, p) \neq \emptyset$.

(b) Es existiert eine von x_c und p unabhängige Konstante $\theta > 0$ mit

$$f(x_c) - f(x_c + tp) \ge \theta \left(\frac{\nabla f(x_c)^T p}{\|p\|_2} \right)^2 \qquad \text{für alle } t \in T_G(x_c, p).$$

4. Eine Funktion $\phi\colon [a,b] \longrightarrow \mathbb{R}$ heißt unimodal, wenn es genau ein $t^* \in (a,b)$ gibt mit $\phi(t^*) = \min_{t\in[a,b]} \phi(t)$, und wenn ϕ auf $[a,t^*]$ monoton fallend und auf $[t^*,b]$ monoton wachsend ist. Zur Lokalisierung des Minimums t^* der auf $[a,b]$ unimodularen Funktion ϕ betrachte man die Methode vom goldenen Schnitt:

- Sei $\epsilon > 0$ (gewünschte Genauigkeit) gegeben, setze $F := (\sqrt{5}-1)/2$.
- Berechne $\left\{ \begin{array}{lcllcl} s & := & a + (1-F)(b-a), & \phi_s & := & \phi(s), \\ t & := & a + F(b-a), & \phi_t & := & \phi(t). \end{array} \right.$
- Solange $b - a > \epsilon$:
  - Falls $\phi_s > \phi_t$, dann:

$$a := s, \quad s := t, \quad t := a + F(b-a), \quad \phi_s := \phi_t, \quad \phi_t := \phi(t)$$

  - Andernfalls:

$$b := t, \quad t := s, \quad s := a + (1-F)(b-a), \quad \phi_t := \phi_s, \quad \phi_s := \phi(s).$$

- $t^* \approx (a+b)/2$.

Man beweise, dass dieser Algorithmus nach endlich vielen Schritten mit einem Intervall $[a,b]$ abbricht, das t^* enthält.

5. Man gebe eine Matlab-Implementation der Methode des goldenen Schnittes an und erprobe sie an den Funktionen[5]

(a) $\phi(t) := -t/(t^2 + c)$ mit $c := 2$,

(b) $\phi(t) := (t+c)^5 - 2(t+c)^4$ mit $c := 0.004$.

Hierbei veranschauliche man sich die Funktionen durch einen Plot.

6. Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Bedingungen (V) (a)–(c) in Unterabschnitt 3.1.1. Zur Lösung von (P) wende man den Modellalgorithmus mit $p_k := -g_k$ (hierbei sei wieder $g_k := \nabla f(x_k)$) und der konstanten

[5]Siehe C. Geiger, C. Kanzow (1999, S. 52).

Schrittweite $t_k := 1/\gamma$ an. Dann ist jeder Häufungspunkt der durch das Verfahren erzeugten Folge $\{x_k\}$ eine stationäre Lösung von (P).

Hinweis: Man wende Lemma 1.1 an, um

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2\gamma}\|g_k\|^2, \qquad k = 0, 1, \ldots,$$

zu zeigen und beweise hiermit die Behauptung.

7. Die Voraussetzungen von Satz 1.5 seien erfüllt. Die Zielfunktion f besitze in der Niveaumenge L_0 nur endlich viele stationäre Punkte. Der Modellalgorithmus erzeuge eine Folge $\{x_k\}$ mit $\lim_{k\to\infty}(x_{k+1} - x_k) = 0$. Dann konvergiert die gesamte Folge $\{x_k\}$ gegen einen der stationären Punkte von f.

   Hinweis: Siehe J. M. Ortega, W. C. Rheinboldt (1970, S. 476)[6].

8. Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische, positiv definite Matrix mit kleinstem Eigenwert $\lambda_{\min}$ und größtem Eigenwert $\lambda_{\max}$ Dann gilt die *Ungleichung von Kantorowitsch*:

$$(x^T Ax)(x^T A^{-1}x) \leq \frac{(\lambda_{\min} + \lambda_{\max})^2}{4\lambda_{\min}\lambda_{\max}}(x^T x)^2 \qquad \text{für alle } x \in \mathbb{R}^n.$$

   Hinweis: Durch eine orthogonale Ähnlichkeitstranformation kann man erreichen, dass A o. B. d. A. eine Diagonalmatrix ist, siehe auch D. G. Luenberger (1973, S. 151)[7] und C. Geiger, C. Kanzow (1999, S. 71).

9. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit mit kleinstem Eigenwert $\lambda_{\min}$ und größtem Eigenwert $\lambda_{\max}$. Für alle $x \in \mathbb{R}^n \setminus \{0\}$ ist dann

$$\left(\frac{x^T Ax}{\|x\|_2\,\|Ax\|_2}\right)^2 \geq \frac{4\lambda_{\min}\lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2} = \frac{4\kappa_2(A)}{(1 + \kappa_2(A))^2},$$

   wobei $\kappa_2(A)$ natürlich die Kondition von A bezüglich der Spektralnorm bedeutet.

10. Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Voraussetzungen (K) (a)–(c) aus Lemma 1.6. Auf (P) wende man das Gradientenverfahren (d. h. $p_k := -g_k$, wobei $g_k := \nabla f(x_k)$) mit exakter Schrittweite (d. h. $t_k = t^*(x_k, p_k)$) an. Mit x^* werde die globale Lösung von (P) bezeichnet.

    (a) Mit Hilfe von Satz 1.2 und Lemma 1.6 zeige man, dass

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{c}{\gamma}\right)[f(x_k) - f(x^*)], \qquad k = 0, 1, \ldots.$$

    (b) Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische, positiv definite Matrix mit kleinstem Eigenwert $\lambda_{\min}$ und größtem Eigenwert $\lambda_{\max}$, ferner sei $f(x) := \frac{1}{2}x^T Ax - b^T x$. Dann sind

[6] J. M. Ortega, W. C. Rheinboldt (1970) *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York-London.

[7] D. G. Luenberger (1973) *Introduction to Linear and Nonlinear Programmin*. Addison-Wesley, Reading.

die Voraussetzungen (K) (a)–(c) mit $c := \lambda_{\min}$ und $\gamma := \lambda_{\max}$ erfüllt. Man zeige, dass sich die aus dem vorigen Teil der Aufgabe resultierende Abschätzung zu

$$f(x_{k+1}) - f(x^*) \leq \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right)^2 [f(x_k) - f(x^*)], \qquad k = 0, 1, \ldots$$

verbessern lässt.

Hinweis: Man zeige

$$\frac{f(x_k) - f(x^*)}{f(x_k) - f(x_{k+1})} = \frac{(g_k^T A g_k)(g_k^T A^{-1} g_k)}{\|g_k\|_2^4}$$

und wende die Ungleichung von Kantorowitsch an.

11. Sei[8] $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ stetig differenzierbar und gleichmäßig konvex sowie $\nabla f(\cdot)$ global lipschitzstetig auf dem gesamten $\mathbb{R}^n$. Es bezeichne $\gamma > 0$ die zugehörige Lipschitz-Konstante (bezüglich der euklidischen Norm) sowie $c > 0$ die Konstante aus der Definition der gleichmäßigen Konvexität. Dann konvergiert das Gradientenverfahren mit konstanter Schrittweite

$$x_{k+1} := x_k - \alpha \nabla f(x_k), \qquad k = 0, 1, \ldots,$$

für jeden Startvektor $x_0 \in \mathbb{R}^n$, sofern $\alpha \in (0, 2c/\gamma^2)$.

Hinweis: Man wende den Banachschen Fizpunktsatz an.

3.2 Quasi-Newton-Verfahren

3.2.1 Das Newton-Verfahren

Wir betrachten wieder die unrestringierte Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n,$$

wobei die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ als zweimal stetig differenzierbar vorausgesetzt wird. Stationäre Lösungen von (P) sind Lösungen des i. Allg. nichtlinearen Gleichungssystems $\nabla f(x) = 0$. Es liegt nahe, hierauf das Newton-Verfahren anzuwenden, was auf die Iterationsvorschrift

$$x_{k+1} := x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

führt. Der lokale Konvergenzsatz für das Newton-Verfahren sagt aus:

Satz 2.1 *Die Abbildung $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ sei auf einer Umgebung von $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbar. Es sei $\nabla f(x^*) = 0$ (also x^* ein stationärer Punkt von f bzw. eine stationäre Lösung von (P)) und $\nabla^2 f(x^*)$ nichtsingulär. $\|\cdot\|$ bezeichne eine beliebige Norm im $\mathbb{R}^n$ bzw. die zugeordnete Matrixnorm. Dann existiert ein $\delta > 0$ derart, dass für jedes*

$$x_0 \in B[x^*; \delta] := \{x \in \mathbb{R}^n : \|x - x^*\| \leq \delta\}$$

[8] Diese Aufgabe haben wir C. Geiger, C. Kanzow (1999, S. 80) entnommen.

die durch das Newton-Verfahren

$$x_{k+1} := x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k), \qquad k = 0, 1, \ldots,$$

gewonnene Folge $\{x_k\}$ *definiert ist (d. h.* $\nabla^2 f(x_k)$ *existiert und ist nichtsingulär,* $k = 0, 1, \ldots$*) und superlinear gegen* x^* *konvergiert*[9]*. Ist zusätzlich* $\nabla^2 f(\cdot)$ *auf einer (hinreichend kleinen) Kugel um* x^* *in* x^* *lipschitzstetig, d. h. existieren* $\eta > 0$ *und* $L > 0$ *mit*

$$\|\nabla^2 f(x) - \nabla^2 f(x^*)\| \leq L \|x - x^*\| \qquad \text{für alle } x \text{ mit } \|x - x^*\| \leq \eta,$$

so konvergiert $\{x_k\}$ *bei hinreichend kleinem* $\delta > 0$ *für jedes* $x_0 \in B[x^*; \delta]$ *sogar von mindestens zweiter Ordnung gegen* x^**, d. h. die Folge* $\{x_k\}$ *konvergiert gegen* x^* *und es existiert eine Konstante* $C > 0$ *mit* $\|x_{k+1} - x^*\| \leq C \|x_k - x^*\|^2$, $k = 0, 1, \ldots$.

Beispiel: Wir kommen auf ein Beispiel aus Kapitel 1 zurück. Es sei nämlich $f: \mathbb{R}^2 \longrightarrow \mathbb{R}$ definiert durch

$$f(x) := -x_1^2 x_2 + \frac{1}{4}(2x_1^2 - x_2^2) - \frac{1}{2}(2 - x_1^2 - x_2^2)^2.$$

In $(0, 0)$ liegt ein lokales Minimum. Wir schreiben ein function file `Geiger.m` (da die Funktion im Buch von Geiger-Kanzow vorkommt), in dem der Funktionswert, der Gradient und die Hessesche von f berechnet werden (es hätte für das jetzt kommende Beispiel genügt, nur den Gradienten und die Hessesche zu berechnen). Das könnte folgendermaßen aussehen:

```
function [f,g,B]=Geiger(x);
c=2-x(1)^2-x(2)^2;
f=-x(1)^2*x(2)+0.25*(2*x(1)^2-x(2)^2)-0.5*c^2;
if nargout>1
    g=[-2*x(1)*x(2)+x(1)+2*x(1)*c;-x(1)^2-0.5*x(2)+2*x(2)*c];
end;
if nargout>2
    B=[5-x(2)-6*x(1)^2-2*x(2)^2,-2*x(1)*(1+2*x(2));
       -2*x(1)*(1+2*x(2)),3.5-x(2)-2*x(1)^2-6*x(2)^2];
end;
```

Nun schreiben wir eine kleine, einfache Funktion `Newton` zum ungedämpften Newton-Verfahren. Diese könnte folgendermaßen aussehen:

```
function [x,iter]=Newton(fun,x_0,max_iter,tol);
%*******************************************************
%Undamped Newton method
%*******************************************************
%Input parameter:
%          fun        function to be minimized
%                     [f,g,B]=fun(x) gives function value,
%                     gradient and hessian at x
%          x_0        starting vector
```

[9] Dies bedeutet bekanntlich, dass $\lim_{k\to\infty} \|x_{k+1} - x^*\| / \|x_k - x^*\| = 0$, woraus natürlich auch die Konvergenz von $\{x_k\}$ gegen x^* folgt.

```
%          max_iter   maximal number of iterations
%          tol        method is stopped if norm of gradient
%                     is <=tol
%Output parameter:
%          x          approximate solution (if successful)
%          iter       number of iterations performed
%*************************************************
x_c=x_0; [f_c,g_c,B_c]=feval(fun,x_c);iter=0;
while (norm(g_c)>tol)&(iter<max_iter)
     p=-B_c\g_c; x_c=x_c+p;
     [f_c,g_c,B_c]=feval(fun,x_c);
     iter=iter+1;
end;
x=x_c;
```

Ein Aufruf

```
[x,iter]=Newton('Geiger',[5;4],100,1e-8);
```

ergibt (nach `format long`)

$$x = \begin{pmatrix} 0.70710678441217 \\ 0.99999999614887 \end{pmatrix}, \qquad \text{iter} = 18.$$

Hier wird also der Sattelpunkt $(1/\sqrt{2}, 1)$ approximiert. Mit dem Startwert $(2, -1)$ wird z. B. das (lokale) Maximum in $(\frac{1}{6}\sqrt{95}, -\sqrt{56})$ approximiert, entsprechendes gilt für den Startwert $(-2, -1)$. Experimente zeigen, dass man in einer kleinen Umgebung von $(0, 0)$ starten muss, um Konvergenz gegen dieses (lokale) Minimum zu erreichen. □

Durch die Einführung von Schrittweiten, also den Übergang zum gedämpften Verfahren

$$x_{k+1} := x_k - t_k \nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

kann man versuchen, zu einem global konvergenten Verfahren zu kommen. Unter geeigneten Konvexitätsvoraussetzungen, die u. a. sichern, dass $\nabla^2 f(x_k)$ positiv definit und damit die Newton-Richtung $p_k := -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$ eine Abstiegsrichtung ist, wird man die Konvergenz des gedämpften Newton-Verfahrens erwarten. Ferner wird man die Schrittweitenstrategie so gestalten wollen, dass nach endlich vielen Schritten, wenn die durch das gedämpfte Newton-Verfahren erzeugten Näherungen erst einmal hinreichend nahe bei einer Lösung liegen, automatisch vom gedämpften zum ungedämpften Newton-Verfahren übergegangen wird. Diese Erwartungen werden durch den folgenden globalen Konvergenzsatz für das Newton-Verfahren bestätigt.

Satz 2.2 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Über die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ wird vorausgesetzt:*

(a) Mit einem $x_0 \in \mathbb{R}^n$ ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ konvex.

(b) f ist auf einer offenen Obermenge von L_0 zweimal stetig differenzierbar und es existieren positive Konstanten $c \leq \gamma$ mit

$$(*) \qquad c\,\|p\|_2^2 \leq p^T \nabla^2 f(x) p \leq \gamma\,\|p\|_2^2 \qquad \text{für alle } x \in L_0,\ p \in \mathbb{R}^n.$$

Zur Bestimmung der unter diesen Voraussetzungen eindeutig existierenden (globalen) Lösung x^* von (P) betrachte man das gedämpfte Newton-Verfahren

$$x_{k+1} := x_k + t_k p_k \qquad \textit{mit} \quad p_k := -\nabla^2 f(x_k)^{-1} \nabla f(x_k),$$

wobei t_k in jedem Schritt die Wolfe- oder die Armijo-Schrittweite sei. Dann gilt: Bricht das Verfahren nicht vorzeitig mit der Lösung x^* von (P) ab, so erzeugt es eine gegen x^* konvergente Folge $\{x_k\}$. Ferner ist $t_k = 1$ für alle hinreichend größen k, nach endlich vielen Schritten geht das gedämpfte Newton-Verfahren also in das ungedämpfte über.

Beweis: Aus $c\,\|p\|_2^2 \le p^T \nabla^2 f(x) p$ für alle $x \in L_0$ und alle $p \in \mathbb{R}^n$ folgt wegen Satz 2.3, dass die Zielfunktion f auf der nach Voraussetzung konvexen Niveaumenge L_0 gleichmäßig konvex ist. Wegen $p^T \nabla^2 f(x) p \le \gamma\,\|p\|_2^2$ für alle $x \in L_0$ und alle $p \in \mathbb{R}^n$ ist $\|\nabla^2 f(x)\|_2 \le \gamma$. Hieraus folgt die Lipschitzstetigkeit von $\nabla f(\cdot)$ auf L_0 mit der Lipschitzkonstanten $\gamma > 0$ (bezüglich der euklidischen Norm). Denn für beliebige $x, y \in L_0$ ist

$$\begin{aligned}
\|\nabla f(x) - \nabla f(y)\| &= \Big\| \int_0^1 \nabla^2 f(\underbrace{x + s(y-x)}_{\in L_0})(x-y)\,ds \Big\| \\
&\le \int_0^1 \underbrace{\|\nabla^2 f(x + s(y-x))\|}_{\le \gamma}\,ds\,\|x-y\| \\
&\le \gamma\,\|x-y\|.
\end{aligned}$$

Zum Nachweis der Konvergenz der Folge $\{x_k\}$ wollen wir Satz 1.7 anwenden. Durch (a) und (b) sind die Voraussetzungen (K) (a)–(c) aus Lemma 1.6 erfüllt, wie wir uns gerade eben überlegt haben. Die Bedingung $(*)$ in Voraussetzung (b) impliziert

$$\delta_k := \min\Big[-\frac{\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|_2^2}, \Big(\frac{\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|_2\,\|p_k\|_2}\Big)^2\Big] \ge \min\Big[\frac{1}{\gamma}, \frac{c}{\gamma}\Big] =: \delta$$

wie man unschwer nachweist. Insbesondere ist

$$\delta \le \frac{1}{k+1} \sum_{j=0}^{k} \delta_j, \qquad k = 0, 1, \ldots.$$

Wegen Satz 1.7 konvergiert die Folge $\{x_k\}$ gegen die eindeutige (globale) Lösung x^* von (P). Um $t_k = 1$ für alle hinreichend großen k nachzuweisen, müssen wir (Wolfe-Schrittweite)

$$f(x_k + p_k) \le f(x_k) + \alpha \nabla f(x_k)^T p_k, \qquad \nabla f(x_k + p_k)^T p_k \ge \beta \nabla f(x_k)^T p_k$$

bzw. (Armijo-Schrittweite)

$$f(x_k + p_k) \le f(x_k) + \alpha \nabla f(x_k)^T p_k$$

für alle hinreichend großen k nachweisen. Es genügt, die Wolfe-Schrittweite zu betrachten und

$$\lim_{k\to\infty} \frac{f(x_k + p_k) - f(x_k)}{\nabla f(x_k)^T p_k} = \frac{1}{2}, \qquad \lim_{k\to\infty} \frac{\nabla f(x_k + p_k)^T p_k}{\nabla f(x_k)^T p_k} = 0$$

nachzuweisen.

Wegen $\|p_k\|_2 \leq \|\nabla f(x_k)\|_2/c$ konvergiert die Folge $\{p_k\}$ der Newton-Richtungen gegen den Nullvektor. Da außerdem o. B. d. A. x^* im Innern der Niveaumenge L_0 liegt und $\{x_k\}$ gegen x^* konvergiert, liegt die gesamte Verbindungsstrecke zwischen x_k und $x_k + p_k$ für alle hinreichend großen k in L_0. Mit einem $\theta_k \in (0,1)$ ist daher für diese k wegen des Mittelwertsatzes

$$\begin{aligned}\frac{f(x_k+p_k)-f(x_k)}{\nabla f(x_k)^T p_k} &= \frac{\nabla f(x_k)^T p_k + \frac{1}{2} p_k^T \nabla^2 f(x_k+\theta_k p_k) p_k}{\nabla f(x_k)^T p_k} \\ &= 1 - \frac{p_k^T \nabla^2 f(x_k+\theta_k p_k) p_k}{2 p_k^T \nabla^2 f(x_k) p_k} \\ &= \frac{1}{2} - \frac{p_k^T [\nabla^2 f(x_k+\theta_k p_k) - \nabla^2 f(x_k)] p_k}{2 p_k^T \nabla^2 f(x_k)}.\end{aligned}$$

Wegen $x_k + \theta_k p_k \to x^*$ und $x_k \to x^*$ ist nun

$$\frac{|p_k^T [\nabla^2 f(x_k+\theta_k p_k) - \nabla^2 f(x_k)] p_k|}{p_k^T \nabla^2 f(x_k) p_k} \leq \frac{1}{c} \|\nabla^2 f(x_k+\theta_k p_k) - \nabla^2 f(x_k)\|_2 \to 0,$$

womit

$$\lim_{k\to\infty} \frac{f(x_k+p_k)-f(x_k)}{\nabla f(x_k)^T p_k} = \frac{1}{2} > \alpha$$

und damit

$$f(x_k+p_k) \leq f(x_k) + \alpha \nabla f(x_k)^T p_k$$

für alle hinreichend großen k bewiesen ist. Zum Nachweis der zweiten Beziehung beachte man, dass wiederum wegen des Mittelwertsatzes ein $\eta_k \in (0,1)$ mit

$$\nabla f(x_k+p_k)^T p_k = \nabla f(x_k)^T p_k + p_k^T \nabla^2 f(x_k+\eta_k p_k) p_k$$

exisyiert. Folglich ist

$$\begin{aligned}\left|\frac{\nabla f(x_k+p_k)^T p_k}{\nabla f(x_k)^T p_k}\right| &= \frac{|p_k^T [\nabla^2 f(x_k+\eta_k p_k) - \nabla^2 f(x_k)] p_k}{p_k^T \nabla^2 f(x_k) p_k} \\ &\leq \frac{1}{c} \|\nabla^2 f(x_k+\eta_k p_k) - \nabla^2 f(x_k)\|_2 \\ &\to 0.\end{aligned}$$

Insgesamt ist der Satz damit bewiesen. □ □

Beispiel: Wir wollen einmal das Newton-Verfahren mit der Wolfe-Schrittweite auf die Rosenbrock-Funktion anwenden. Hierzu erweitern wir das schon früher erklärte Function-File `Rosenbrock.m` durch

```
function [f,g,B]=Rosenbrock(x);
f=100*(x(2)-x(1)^2)^2+(1-x(1))^2;
if nargout>1
  g=[-400*x(1)*(x(2)-x(1)^2)-2*(1-x(1));200*(x(2)-x(1)^2)];
```

```
end;
if nargout>2
  B=[1200*x(1)^2-400*x(2)+2 -400*x(1);-400*x(1) 200];
end;
```

Mit dem kleinen Programm

```
x=[-1.2;1];x_stern=[1;1];
[f,g,B]=Rosenbrock(x);
A=[];
for k=1:25
   p=-B\g;
   t=Wolfe(x,p,'Rosenbrock');
   x=x+t*p;[f,g,B]=Rosenbrock(x);
   A=[A;k,norm(g),norm(x-x_stern),f,t];
end;
disp(A)
```

erhalten wir Ergebnisse, die wir nicht angeben wollen, die aber zeigen, dass die Konvergenzgeschwindigkeit, wie nicht anders zu erwarten, zumindestens lokal gut ist. □

Es gibt einige Einwände gegen das Newton-Verfahren. Der erste ist theoretischer Art und besteht darin, dass die Anwendung des Newton-Verfahrens fern einer stationären Lösung von (P) keinen Sinn macht. Denn hier zieht die dem Newton-Verfahren zugrunde liegende Motivation nicht. Denn diese besteht ja darin, das zu lösende nichtlineare Gleichungssystem $\nabla f(x) = 0$ in einer aktuellen Näherung x_c zu linearisieren und die Lösung dieses nichtlinearen Gleichungssystems $\nabla f(x_c)+\nabla^2 f(x_c)(x-x_c) = 0$ als neue Näherung x_+ zu nehmen. Auch wenn in dem zu berechnenden Punkt x^* die hinreichende Optimalitätsbedingung zweiter Ordnung erfüllt ist, also $\nabla f(x^*) = 0$ gilt und $\nabla^2 f(x^*)$ positiv definit ist, wird die Hessesche $\nabla^2 f(x_c)$ für von x^* weit entferntes x_c i. Allg. nicht positiv definit sein und damit die Newton-Richtung $p = -\nabla^2 f(x_c)^{-1}\nabla f(x_c)$ nicht unbedingt eine Abstiegsrichtung sein. Der zweite Einwand ist praktischer Art. Das Newton-Verfahren verlangt die Berechnung der Hesseschen der Zielfunktion in der aktuellen Näherung, also der zweiten partiellen Ableitungen. Dies ist bei vielen Anwendungen, bei denen schon die Berechnung des Gradienten der Zielfunktion Mühe bereitet, unzumutbar. Weiter muss beim Newton-Verfahren in jedem Schritt ein lineares Gleichungssystem mit der (symmetrischen) Koeffizientenmatrix $\nabla^2 f(x_c)$ gelöst werden. Die Anzahl der hierzu nötigen arithmetischen Operationen ist im wesentlichen proportional zu n^3. Ferner wird man kaum hoffen können, Kenntnisse über eine Zerlegung (Cholesky-, LR- oder QR-Zerlegung) von $\nabla^2 f(x_c)$ nutzbringend auf die Berechnung einer entsprechenden Zerlegung von $\nabla^2 f(x_+)$ anwenden zu können.

3.2.2 Die Broyden-Klasse und das BFGS-Verfahren

Die *Quasi-Newton-Verfahren* versuchen, die Nachteile (Berechnung zweiter Ableitungen, kostspieliges Lösen linearer Gleichungssysteme) des Newton-Verfahrens zu vermeiden, ohne die Vorteile (globale Konvergenz durch Einführung von Schrittweiten und

automatischer Übergang zum ungedämpften Verfahren bei gleichmäßig konvexer Zielfunktion, lokal superlineare Konvergenz des ungedämpften Verfahrens) aufzugeben. Insbesondere das zu dieser Klasse gehörende *BFGS-Verfahren* (BFGS steht für **B**royden-**F**letcher-**G**oldfarb-**S**hanno, die dieses Verfahren unabhängig voneinander 1970 entdeckten) gilt für glatte, nicht zu hochdimensionale unrestringierte Optimierungsaufgaben, bei denen neben den Zielfunktionswerten auch der Gradient zur Verfügung steht, als das anerkanntermaßen beste Minimierungsverfahren. In den Grundzügen sehen die zu betrachtenden Quasi-Newton-Verfahren folgendermaßen aus (man vergleiche mit dem Modellalgorithmus zu Beginn von Abschnitt 3.1):

- Gegeben $x_0 \in \mathbb{R}^n$ und eine symmetrische, positiv definite Matrix $B_0 \in \mathbb{R}^{n\times n}$. Ferner sei $g_0 := \nabla f(x_0)$.
- Für $k = 0, 1, \ldots$:
  - Test auf Abbruch: Falls $g_k = 0$, dann: STOP.
  - Berechne Abstiegsrichtung $p_k := -B_k^{-1} g_k$.
  - Berechne Schrittweite t_k, etwa die exakte Schrittweite, die Wolfe- oder die Armijo-Schrittweite.
  - Berechne neue Näherung $x_{k+1} := x_k + t_k p_k$ und $g_{k+1} := \nabla f(x_{k+1})$.
  - Berechne symmetrische, positiv definite Matrix $B_{k+1} \in \mathbb{R}^{n\times n}$ durch eine sogenannte *Update-Formel*. In die Berechnung von B_{k+1} gehen i. Allg. B_k sowie $s_k := x_{k+1} - x_k$ und $y_k := g_{k+1} - g_k$ ein.

Ein Quasi-Newton-Verfahren ist also (neben der Wahl der Schrittweitenstrategie) durch die Update-Formel festgelegt. Um Schreibarbeit zu sparen, nehmen wir nun an, $x_c := x_k$ sei eine aktuelle Näherung mit $g_c := \nabla f(x_c) \neq 0$, $B_c := B_k$ sei symmetrisch und positiv definit und damit $p := -B_c^{-1} g_c$ eine Abstiegsrichtung, und mit einer geeigneten Schrittweite $t := t_k$ seien die neue Näherung $x_+ := x_c + tp$ und $g_+ := \nabla f(x_+)$ berechnet. Ferner sei $s := x_+ - x_c$ und $y := g_+ - g_c$. Wie sollte nun die neue symmetrische und positiv definite Matrix $B_+ := B_{k+1}$ berechnet werden? Hierauf gibt es keine eindeutige Antwort. Neben der Symmetrie und positiven Definitheit sollte B_+ der sogenannten *Quasi-Newton-Gleichung*

$$B_+ s = y$$

(gelegentlich auch *Sekantengleichung* genannt) genügen. Als Motivation für die Quasi-Newton-Gleichung geben wir an, dass für hinreichend glattes, etwa zweimal stetig differenzierbares f, die Beziehung

$$\nabla f(y) - \nabla f(x) = \int_0^1 \nabla^2 f(x + t(y-x))\, dt(y-x)$$

gilt. Es liegt daher nahe, $y = B_+ s$ zu fordern, um B_+ "in die Nähe der Hesseschen zu zwingen".

Eine *notwendige* Bedingung für die Existenz einer symmetrischen, positiv definiten Matrix mit $B_+ s = y$ ist offenbar $y^T s > 0$. Diese Bedingung ist z. B. unter den Bedingungen (K) (a)–(c) aus Lemma 1.6 erfüllt, wenn also insbesondere die Zielfunktion f

auf der konvexen Niveaumenge L_0 gleichmäßig konvex (mit einer Konstanten $c > 0$) ist. Denn dann ist

$$y^T s = [\nabla f(x_+) - \nabla f(x_c)]^T (x_+ - x_c) \geq c \, \|x_+ - x_c\|_2^2 = c \, \|s\|_2^2 > 0.$$

Aber auch ohne die gleichmäßige Konvexität von f ist i. Allg. $y^T s > 0$. Denn ist z. B. $t > 0$ eine Wolfe-Schrittweite, so ist

$$y^T s = t[\nabla f(x_+) - \nabla f(x_c)]^T p \geq t(\beta - 1)\nabla f(x_c)^T p > 0$$

mit vorgegebenem $\beta \in (0, 1)$. Ist ferner t die exakte Schrittweite, so ist $\nabla f(x_+)^T p = 0$ und daher $y^T s = -t \nabla f(x_c)^T p > 0$.

Eine Update-Formel der *Broyden-Klasse* ist bei gegebenem $\phi \in \mathbb{R}$ durch

$$B_\phi := B_c - \frac{(B_c s)(B_c s)^T}{s^T B_c s} + \frac{y y^T}{y^T s} + \phi(s^T B_c s) v v^T$$

mit

$$v := \frac{y}{y^T s} - \frac{B_c s}{s^T B_c s}$$

definiert, wobei wir in diesem Zusammenhang ausnahmsweise, um die Abhängigkeit von dem Parameter ϕ zu unterstreichen, B_ϕ statt B_+ schreiben. Die prominentesten Vertreter dieser Broyden-Klasse ergeben sich für $\phi = 0$ (BFGS-Update-Formel) und $\phi = 1$ (DFP-Update-Formel, wobei DFP für **D**avidon-**F**letcher-**P**owell steht, die 1959 bzw. 1963 diese Formel angaben und das zugehörige Verfahren untersuchten, das zu der Zeit einen wesentlichen Fortschritt gegenüber dem vorher noch ziemlich konkurrenzlosen Gradientenverfahren bedeutete). Ist $\phi \in [0, 1]$, so spricht man von einem Update aus der konvexen Broyden-Klasse.

In dem folgenden Satz wird unter der Voraussetzung $y^T s > 0$ u. a. gezeigt, dass mit B_c für alle $\phi \geq 0$ auch B_ϕ symmetrisch und positiv definit ist. Hierbei wird das Lemma von Sherman-Morrison benutzt[10], das wir jetzt zitieren.

Lemma 2.3 *Sei $A \in \mathbb{R}^{n \times n}$ nichtsingulär und $u, v \in \mathbb{R}^n$. Dann gilt:*

1. *Die Matrix $A + uv^T$ ist genau dann nichtsingulär, wenn $1 + v^T A^{-1} u \neq 0$.*

2. *Ist $1 + v^T A^{-1} u \neq 0$, so ist*

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}.$$

Es ist

$$\det(A + uv^T) = (1 + v^T A^{-1} u) \det(A).$$

[10] Siehe z. B.

J. Werner (1992) *Numerische Mathematik 1*. Vieweg, Braunschweig-Wiesbaden.

Satz 2.4 *Seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ sowie eine symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n\times n}$ gegeben. Durch*

$$B_\phi := B_{\mathrm{BFGS}} + \phi(s^T B_c s) v v^T$$

mit

$$B_{\mathrm{BFGS}} := B_c - \frac{(B_c s)(B_c s)^T}{s^T B_c s} + \frac{y y^T}{y^T s}, \qquad v := \frac{y}{y^T s} - \frac{B_c s}{s^T B_c s}$$

sei die zum Parameter $\phi \in \mathbb{R}$ gehörende Update-Formel des Broyden-Verfahrens definiert. Dann gilt:

1. *Für jedes $\phi \in \mathbb{R}$ ist $B_\phi s = y$, d. h. jedes Update der Broyden-Klasse genügt der Quasi-Newton-Gleichung.*
2. *Die Matrix B_ϕ ist für jedes $\phi \geq 0$ symmetrisch und positiv definit.*
3. *Es ist*

$$\det(B_{\mathrm{BFGS}}) = \frac{y^T s}{s^T B_c s} \det(B_c)$$

und

$$\begin{aligned} B_{\mathrm{BFGS}}^{-1} &= B_c^{-1} - \frac{(B_c^{-1}y)(B_c^{-1}y)^T}{y^T B_c^{-1} y} + \frac{s s^T}{y^T s} \\ &\quad + (y^T B_c^{-1} y)\Big[\frac{s}{y^T s} - \frac{B_c^{-1} y}{y^T B_c^{-1} y}\Big]\Big[\frac{s}{y^T s} - \frac{B_c^{-1} y}{y^T B_c^{-1} y}\Big]^T \\ &= B_c^{-1} + \Big(1 + \frac{y^T B_c^{-1} y}{y^T s}\Big)\frac{s s^T}{y^T s} - \frac{s(B_c^{-1}y)^T + (B_c^{-1}y)s^T}{y^T s} \\ &= \Big(I - \frac{s y^T}{y^T s}\Big) B_c^{-1} \Big(I - \frac{y s^T}{y^T s}\Big) + \frac{s s^T}{y^T s}. \end{aligned}$$

Beweis: Die Gültigkeit der Quasi-Newton-Gleichung $B_\phi s = y$ ist offensichtlich. Klar ist ferner, dass mit B_c auch B_ϕ symmetrisch ist. Für den zweiten Teil des Satzes genügt es offenbar zu zeigen, dass B_{BFGS} positiv definit ist.

Da $B_c \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, besitzt B_c eine Cholesky-Zerlegung $B_c = L L_c^T$ mit einer unteren Dreiecksmatrix L_c, deren Diagonalelemente positiv sind. Wir zeigen, dass $B_{\mathrm{BFGS}} = J_+ J_+^T$ mit einer nichtsingulären Matrix J_+, woraus die zweite Behauptung folgt. Hierzu definiere man

$$w := \Big(\frac{y^T s}{s^T B_c s}\Big)^{1/2} L_c^T s, \qquad J_+ := L_c + \frac{(y - L_c w) w^T}{w^T w}.$$

Da

$$\sigma := 1 + \frac{w^T L_c^{-1}(y - L_c w)}{w^T w} = \frac{w^T L_c^{-1} y}{w^T w} = \Big(\frac{y^T s}{s^T B_c s}\Big)^{1/2} \neq 0,$$

ist J_+ nach Lemma 2.3 nichtsingulär und

$$J_+^{-1} = L_c^{-1} - \frac{(L_c^{-1} y - w) w^T L_c^{-1}}{\sigma w^T w}, \qquad \det(J_+) = \sigma \det(L_c).$$

Ferner bestätigt man nach leichter Rechnung, dass $J_+ J_+^T = B_{\text{BFGS}}$, womit der zweite Teil des Satzes bewiesen ist.

Es ist

$$\det(B_{\text{BFGS}}) = \det(J_+)^2 = \sigma^2 \det(L_c)^2 = \frac{y^T s}{s^T B_c s} \det(B_c).$$

Aus $B_{\text{BFGS}}^{-1} = J_+^{-T} J_+^{-1}$ erhält man durch Einsetzen und Umformen leicht einen Beweis der restlichen Behauptungen. □ □

Bemerkung: Der Broyden-Update ist

$$B_\phi = B_{\text{BFGS}} + \phi(s^T B_c s) v v^T$$

so dass eine erneute Anwendung von Lemma 2.3 liefert, dass B_ϕ genau dann nichtsingulär ist, wenn

$$1 + \phi(s^T B_c s) v^T B_{\text{BFGS}}^{-1} v \neq 0.$$

Ist dies der Fall, so ist

$$B_\phi^{-1} = B_{\text{BFGS}}^{-1} - \frac{\phi(s^T B_c s)}{1 + \phi(s^T B_c s)(v^T B_{\text{BFGS}}^{-1} v)} (B_{\text{BFGS}}^{-1} v)(B_{\text{BFGS}}^{-1} v)^T.$$

Hierbei ist

$$B_{\text{BFGS}}^{-1} v = \frac{1}{y^T s}\Big(B_c^{-1} y - \frac{y^T B_c^{-1} y}{y^T s} s\Big), \qquad v^T B_{\text{BFGS}}^{-1} v = \frac{1}{y^T s}\Big(\frac{y^T B_c^{-1} y}{y^T s} - \frac{y^T s}{s^T B_c s}\Big).$$

Mit der im Beweis von Satz 2.4 vorkommenden Matrix J_+ ist ferner

$$B_\phi = J_+ [I + \phi(s^T B_c s)(J_+^{-1} v)(J_+^{-1} v)^T] J_+^T.$$

Daher ist B_ϕ genau dann symmetrisch und positiv definit, wenn es

$$J_\phi := I + \phi(s^T B_c s)(J_+^{-1} v)(J_+^{-1} v)^T$$

ist. Diese Matrix hat den $(n-1)$-fachen Eigenwert 1 und den weiteren Eigenwert $1 + \phi(s^T B_c s)(v^T B_{\text{BFGS}}^{-1} v)$. Mit den Abkürzungen

$$a := y^T B_c^{-1} y, \qquad b := y^T s, \qquad c := s^T B_c s$$

ist dies genau dann der Fall, wenn

$$\begin{aligned} \phi &> -\frac{1}{(s^T B_c s)(v^T B_{\text{BFGS}}^{-1} v)} \\ &= -\frac{b^2}{ac - b^2}. \end{aligned}$$

Hierbei beachte man, dass $b^2 \leq ac$ wegen der Cauchy-Schwarzschen Ungleichung und hier Gleichheit genau dann gilt, wenn y ein positives Vielfaches von $B_c s$ ist. Außerdem ist (wegen des zweiten Teils von Lemma 2.3)

$$\begin{aligned} \det(B_\phi) &= [1 + \phi(s^T B_c s)(v^T B_{\text{BFGS}}^{-1} v)] \det(B_{\text{BFGS}}) \\ &= \Big[1 + \phi c \frac{1}{b}\Big(\frac{a}{b} - \frac{b}{c}\Big)\Big] \frac{b}{c} \det(B_c) \\ &= \Big[(1-\phi)\frac{b}{c} + \phi\frac{a}{b}\Big] \det(B_c). \end{aligned}$$

Dieses Ergebnis hätten wir z. B. auch dadurch erhalten können, indem wir die Eigenwerte von $B_c^{-1/2} B_\phi B_c^{-1/2}$ berechnet hätten, siehe Aufgabe 3. □

Ausgehend von denselben Startelementen x_0 und B_0 erzeugen alle Verfahren der Broyden-Klasse identische Folgen $\{x_k\}$, wenn in jedem Schritt die exakte Schrittweite $t_k = t^*(x_k, p_k)$ gewählt wird. Dieses bemerkenswerte Ergebnis stammt von L. C. W. Dixon (1972)[11], für eine genaue Formulierung und einen Beweis vergleiche man auch J. Werner (1992, S. 198 ff.). Wendet man ein Quasi-Newton-Verfahren der Broyden-Klasse auf eine gleichmäßig konvexe quadratische Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ an und benutzt man in jedem Schritt die exakte Schrittweite, so bricht das Verfahren nach $m \leq n$ Schritten mit der Lösung ab. Wegen des Ergebnisses von Dixon genügt es, dies für das BFGS-Verfahren zu beweisen. Wir begnügen uns mit einer genauen Formulierung und verweisen für einen Beweis auf J. Werner (1992, S. 200).

Satz 2.5 *Auf die unrestringierte Optimierungsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) := \frac{1}{2} x^T A x - b^T x, \quad x \in \mathbb{R}^n,$$

mit symmetrischer und positiv definiter Matrix $A \in \mathbb{R}^{n\times n}$ wende man das BFGS-Verfahren mit exakter Schrittweite an:

- *Seien $x_0 \in \mathbb{R}^n$ und eine symmetrische, positiv definite Matrix $H_0 \in \mathbb{R}^{n\times n}$ vorgegeben. Berechne $g_0 := \nabla f(x_0)$.*
- *Für $k = 0, 1, \ldots$:*
  - *Falls $g_k = 0$, dann: $m := k$, STOP.*
  - *Berechne*

$$p_k := -H_k g_k, \quad t_k := -\frac{g_k^T p_k}{p_k^T A p_k}, \quad x_{k+1} := x_k + t_k p_k, \quad g_{k+1} := \nabla f(x_{k+1}).$$

  - *Mit $s_k := x_{k+1} - x_k$ und $y_k := g_{k+1} - g_k$ berechne*

$$H_{k+1} := H_k + \Big(1 + \frac{y_k^T H_k y}{y_k^T s_k}\Big) \frac{s_k s_k^T}{y_k^T s_k} - \frac{s_k (H_k y_k)^T + (H_k y_k) s_k^T}{y_k^T s_k}.$$

Dann gilt:

1. *Das Verfahren bricht nach $m \leq n$ Schritten mit der Lösung $x_m = x^*$ von (P) ab.*
2. *Es ist $p_i^T A p_k = 0$ für $0 \leq i < k \leq m-1$, d. h. die erzeugten Richtungen sind orthogonal bezüglich des durch $(x, y) := x^T A y$ definierten Skalarproduktes.*
3. *Es ist $p_i^T g_k = 0$ und $H_k y_i = s_i$ für $0 \leq i < k \leq m$.*

[11] L. C. W. Dixon (1972) "Variable metric algorithms: Necessary and sufficient conditions for identical behavior on nonquadratic functions." J. Opt. Theory Appl. 10, 34–40.

4. Ist $m = n$, so ist $H_n = A^{-1}$.

Beispiel: Wir wollen die Aussage des letzten Satzes an einem einfachen Beispiel (siehe P. Spelluci (1993, S. 136), dort allerdings ein kleiner Schreibfehler) nachprüfen. Es sei die Funktion

$$f(x) := x_1^2 - 4x_1x_2 + 8x_2^2 - 3x_1 - 4x_2$$

zu minimieren. Es ist also $f(x) = \frac{1}{2}x^TQx - b^Tx$ mit

$$b := \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \qquad A := \begin{pmatrix} 2 & -4 \\ -4 & 16 \end{pmatrix}.$$

Wir wenden das Verfahren aus Satz 2.5 mit

$$x_0 := \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad H_0 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

an. Wir legen ein Function-File `QuadBFGS.m` ab:

```
function [x,H,m]=QuadBFGS(b,A,x_0,H_0);
%Input-Parameter:
%               b         n-vector
%               A         n-by-n matrix, spd
%               x_0       initial iterate
%               H_0       starting spd matrix
%Output-Parameter:
%               x         solution of Ax-b=0
%               H         last BFGS-matrix, usually H=A^-1
%               m         number of steps
%************************************************************
m=0;
epsilon=1e-12*(norm(b)+norm(A));
x_c=x_0;H=H_0;
g_c=A*x_c-b;
while (norm(g_c)>epsilon)
     m=m+1;
     p=-H*g_c; q=A*p; t=-g_c'*p/(p'*q);
     x_plus=x_c+t*p; g_plus=g_c+t*q;
     s=x_plus-x_c;y=g_plus-g_c;
     b=y'*s;z=H*y;
     H=H+(1+(y'*z)/b)*(s*s')/b-(s*z'+z*s')/b;
     x_c=x_plus; g_c=g_plus;
end;
x=x_c;
```

Belegen wir die Eingangsparameter und rufen diese Funktion auf, so erhalten wir nach $m = 2$ Schritten

$$x = \begin{pmatrix} 4.0000 \\ 1.2500 \end{pmatrix}, \qquad H = \begin{pmatrix} 1.0000 & 0.2500 \\ 0.2500 & 0.1250 \end{pmatrix}.$$

Wir werden später zeigen, dass das auf eine gleichmäßig konvexe, quadratische Zielfunktion angewandte ungedämpfte BFGS-Verfahren global superlinear konvergent ist.

Ändert man obige Funktion indem man die Schrittweite $t = 1$ benutzt, so erhält man z. B. einen Ausstieg nach $m = 9$ Schritten. □

Es muss zugegeben werden, dass die Update-Formeln der Broyden-Klasse und damit insbesondere auch die BFGS-Update-Formel bisher vom Himmel gefallen sind. In den Aufgaben 4 und 5 gehen wir auf Variationsansätze ein, bei denen das Problem betrachtet wird, unter allen symmetrischen und positiv definiten Matrizen, welche der Quasi-Newton-Gleichung genügen, eine zu bestimmen, die bezüglich eines geeigneten "Abstandes" möglichst wenig von der aktuellen, symmetrischen und positiv definiten Matrix $B_c \in \mathbb{R}^{n\times n}$ entfernt ist.

3.2.3 Globale Konvergenz des BFGS-Verfahrens

In diesem Unterabschnitt gehen wir auf die globale Konvergenz des BFGS-Verfahrens ein, beweisen also Konvergenzergebnisse, bei denen *nicht* vorausgesetzt wird, dass der Startwert in der Nähe einer Lösung ist.

Wie stets in dieser Vorlesung ist auch hier die unrestringierte Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n,$$

gegeben. In der globalen und der später zu untersuchenden lokalen Konvergenzanalyse spielt eine Funktion $\psi(\cdot)$ eine Rolle (siehe auch Aufgabe 4), die wir im folgenden Lemma definieren und von der wir einige Eigenschaften beweisen werden.

Lemma 2.6 *Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n\times n$-Matrizen und $\mathcal{S}_+^{n\times n} \subset \mathcal{S}^{n\times n}$ die konvexe Teilmenge der positiv definiten Matrizen. Man definiere $\psi\colon \mathcal{S}_+^{n\times n} \longrightarrow \mathbb{R}$ durch*

$$\psi(A) := \operatorname{tr}(A) - \ln\det(A).$$

Dann gilt:

1. *Es ist $\psi(A) \geq n$ für alle $A \in \mathcal{S}_+^{n\times n}$. Es gilt Gleichheit genau dann, wenn $A = I$.*
2. *Ist $\{A_k\} \subset \mathcal{S}_+^{n\times n}$, so ist $\{\psi(A_k)\} \subset \mathbb{R}_+$ genau dann (nach oben) beschränkt, wenn $\{\|A_k\|_2\}$ und $\{\|A_k^{-1}\|\}$ beschränkt sind.*
3. *Die Abbildung $\psi(\cdot)$ ist auf $\mathcal{S}_+^{n\times n}$ konvex.*

Beweis: Sind $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$ die Eigenwerte der symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$, so ist

$$\psi(A) = \sum_{i=1}^{n} \underbrace{[\lambda_i(A) - \ln\lambda_i(A)]}_{\geq 1} \geq n.$$

Gilt hier Gleichheit, so ist $\lambda_i(A) = 1$, $i = 1, \ldots, n$, und folglich $A = I$. Für dieses Argument spielt die Funktion $t - \ln t$ auf $\mathbb{R}_+$ eine Rolle, wir veranschaulichen sie uns in Abbildung 3.2 Offenbar ist nämlich $1 \leq t - \ln t$ für alle $t > 0$ und es gilt Gleichheit genau dann, wenn $t = 1$. Ist $\{\psi(A_k)\}$ für eine Folge $\{A_k\} \subset \mathcal{S}_+^{n\times n}$ beschränkt, so

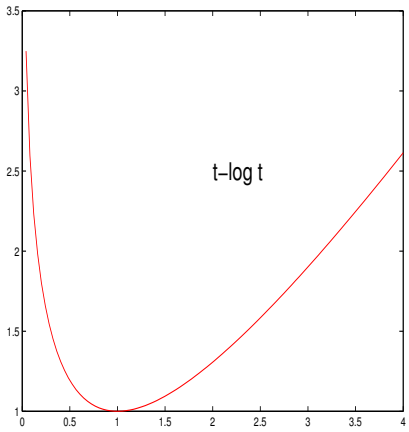

Abbildung 3.2: Die Funktion $t - \ln t$

existieren positive Konstanten c und C mit $c \leq \lambda_{\min}(A_k)$ und $\lambda_{\max}(A_k) \leq C$ für alle k, woraus auch die zweite Behauptung folgt. Die dritte Aussage hatten wir in einem Beispiel im Anschluss an Satz 2.2 in Abschnitt 2.2 bewiesen (leider wird diese Aussage im weiteren aber keine Rolle spielen). Damit ist das Lemma bewiesen. □ □

Es folgt nun ein globaler Konvergenzsatz für das BFGS-Verfahren, wobei die Zielfunktion als gleichmäßig konvex vorausgesetzt wird. Von den früheren Beweisansätzen sei nur der von M. J. D. Powell (1976)[12] genannt.

Satz 2.7 *Gegeben sei die unrestringierte Optimierungsaufgabe (P), die Voraussetzungen (K) (a)–(c) aus Lemma 1.6 seien erfüllt. Es gelte also:*

(K) (a) *Mit einem gegebenen $x_0 \in M$ (Startwert eines Iterationsverfahrens) ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ konvex.*

(b) *Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar und auf L_0 gleichmäßig konvex, d. h. es existiert eine Konstante $c > 0$ mit*

$$\frac{c}{2}\|y - x\|_2^2 + \nabla f(x)^T(y - x) \leq f(y) - f(x) \qquad \text{für alle } x, y \in L_0.$$

.

(c) *Der Gradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig, d. h. es existiert eine Konstante $\gamma > 0$ mit*

$$\|\nabla f(x) - \nabla f(y)\| \leq \gamma \|x - y\| \qquad \text{für alle } x, y \in L_0.$$

Man betrachte das durch die exakte, Wolfe- oder Armijo-Schrittweite gedämpfte BFGS-Verfahren:

[12] M. J. D. Powell (1976) "Some global convergence properties of a variable metric algorithm for minimization without exact line searches." SIAM-AMS Proceedings 9, 53–72.

- *Gegeben der Startwert $x_0 \in \mathbb{R}^n$, sei $g_0 := \nabla f(x_0)$. Ferner sei eine symmetrische, positiv definite Matrix $B_0 \in \mathbb{R}^{n\times n}$ gegeben.*
- *Für $k = 0, 1, \ldots$:*
  - *Falls $g_k = 0$, dann: STOP, x_k ist Lösung von (P).*
  - *Berechne $p_k := -B_k^{-1} g_k$.*
  - *Sei $t_k > 0$ die exakte Schrittweite, Wolfe-Schrittweite oder Armijo-Schrittweite in x_k in Richtung p_k.*
  - *Sei $x_{k+1} := x_k + t_k p_k$ und berechne $g_{k+1} := \nabla f(x_{k+1})$.*
  - *Mit $s_k := x_{k+1} - x_k$ und $y_k := g_{k+1} - g_k$ sei*

$$B_{k+1} := B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}.$$

Dann gilt: Das Verfahren bricht nach endlich vielen Schritten mit der Lösung x^* von (P) ab oder es liefert eine Folge $\{x_k\}$, die R-linear gegen x^* konvergiert, d. h. es existieren Konstanten $C > 0$ und $q \in (0, 1)$ mit $\|x_k - x^*\| \le Cq^k$ für alle k.

Beweis: Die Durchführbarkeit des Verfahrens ist gesichert. Denn ist $g_k \neq 0$ und B_k symmetrisch und positiv definit, so ist $p_k = -B_k^{-1} g_k$ eine Abstiegsrichtung und daher $s_k \neq 0$. Wegen der gleichmäßigen Konvexität der Zielfunktion f ist folglich $y_k^T s_k \ge c\,\|s_k\|^2 > 0$ und damit auch B_{k+1} positiv definit.

Es wird angenommen, das Verfahren breche nicht schon nach endlich vielen Schritten mit der Lösung ab. Wir wollen Satz 1.7 anwenden und zeigen hierzu die Existenz einer Konstanten $\delta > 0$ mit

$$\delta \le \frac{1}{k+1} \sum_{j=0}^{k} \delta_j, \qquad k = 0, 1, \ldots,$$

wobei

$$\delta_j := \min\Big[-\frac{g_j^T p_j}{\|g_j\|^2}, \Big(\frac{g_j^T p_j}{\|g_j\|\,\|p_j\|}\Big)^2\Big].$$

Hierzu benutzen wir die durch $\psi(A) := \operatorname{tr}(A) - \ln\det(A)$ auf der Menge $\mathcal{S}_+^{n\times n}$ der symmetrischen, positiv definiten $n\times n$-Matrizen in Lemma 2.6 definierte Abbildung. Wegen $\det(B_{k+1}) = (y_k^T s_k / s_k^T B_k s_k)\det(B_k)$ (siehe den dritten Teil von Satz 2.4) erhalten wir

$$\begin{aligned}
\psi(B_{k+1}) &= \psi(B_k) - \frac{\|B_k s_k\|^2}{s_k^T B_k s_k} + \frac{\|y_k\|^2}{y_k^T s_k} - \ln\frac{y_k^T s_k}{s_k^T B_k s_k} \\
&= \psi(B_k) + \ln\Big(\frac{s_k^T B_k s_k}{\|s_k\|\,\|B_k s_k\|}\Big)^2 + \Big[1 - \frac{\|B_k s_k\|^2}{s_k^T B_k s_k} + \ln\frac{\|B_k s_k\|^2}{s_k^T B_k s_k}\Big] \\
&\quad + \Big[\frac{\|y_k\|^2}{y_k^T s_k} - 1 - \ln\frac{y_k^T s_k}{\|s_k\|^2}\Big].
\end{aligned}$$

Man überzeugt sich leicht davon, dass der letzte Term beschränkt ist. Denn es ist

$$\frac{\|y_k\|^2}{y_k^T s_k} \le \frac{\gamma^2 \|s_k\|^2}{y_k^T s_k} \le \frac{\gamma^2}{c}$$

und daher

$$\Big[\frac{\|y_k\|^2}{y_k^T s_k} - 1 - \ln \frac{y_k^T s_k}{\|s_k\|^2}\Big] \le \frac{\gamma^2}{c} - 1 - \ln c =: \hat{C}.$$

Mit $C := \psi(B_0) + \hat{C}$ ist daher

$$\begin{aligned}
n &\le \psi(B_{k+1}) \\
&\le \psi(B_k) + \ln\Big(\frac{s_k^T B_k s_k}{\|s_k\| \, \|B_k s_k\|}\Big)^2 + \Big[1 - \frac{\|B_k s_k\|^2}{s_k^T B_k s_k} + \ln \frac{\|B_k s_k\|^2}{s_k^T B_k s_k}\Big] + \hat{C} \\
&\le \psi(B_0) + \sum_{j=0}^{k} \Big\{ \ln\Big(\frac{s_j^T B_j s_j}{\|s_j\| \, \|B_j s_j\|}\Big)^2 + \Big[1 - \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} + \ln \frac{\|B_j s_j\|^2}{s_j^T B_j s_j}\Big] \Big\} + \hat{C}(k+1) \\
&\le C(k+1) + \sum_{j=0}^{k} \Big\{ \underbrace{\ln\Big(\frac{s_j^T B_j s_j}{\|s_j\| \, \|B_j s_j\|}\Big)^2}_{\le 0} + \underbrace{\Big[1 - \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} + \ln \frac{\|B_j s_j\|^2}{s_j^T B_j s_j}\Big]}_{\le 0} \Big\}.
\end{aligned}$$

Folglich ist

$$\sum_{j=0}^{k} \Big\{ \underbrace{\ln\Big(\frac{\|s_j\| \, \|B_j s_j\|}{s_j^T B_j s_j}\Big)^2}_{\ge 0} + \underbrace{\Big[\frac{\|B_j s_j\|^2}{s_j^T B_j s_j} - 1 - \ln \frac{\|B_j s_j\|^2}{s_j^T B_j s_j}\Big]}_{\ge 0} \Big\} \le C(k+1)$$

für alle k. Nun überlegen wir uns:

- Sind $\alpha_0, \ldots, \alpha_k \ge 0$ und $a > 0$ eine Zahl mit $\sum_{j=0}^{k} \alpha_j \le a(k+1)$, so gibt es eine Indexmenge $J_k \subset \{0, \ldots, k\}$, die mindestens $\frac{1}{2}(k+1)$ Elemente enthält, und für die $\alpha_j \le 2a$ für alle $j \in J_k$.

Denn: Man definiere $I_k := \{i \in \{0, \ldots, k\} : \alpha_i > 2a\}$. Dann ist

$$(k+1)a \ge \sum_{j=0}^{k} \alpha_j \ge \sum_{i \in I_k} \alpha_i > \#(I_k) 2a,$$

so dass I_k weniger als $\frac{1}{2}(k+1)$ Elemente enthält. Dann ist $J_k := \{0, \ldots, k\} \setminus I_k$ die gesuchte Indexmenge.

Eine Anwendung dieser Zwischenbehauptung liefert für jedes k die Existenz einer Indexmenge $J_k \subset \{0, \ldots, k\}$ mit mindestens $\frac{1}{2}(k+1)$ Elementen und

$$\ln\Big(\frac{\|s_j\| \, \|B_j s_j\|}{s_j^T B_j s_j}\Big)^2 + \Big[\frac{\|B_j s_j\|^2}{s_j^T B_j s_j} - 1 - \ln \frac{\|B_j s_j\|^2}{s_j^T B_j s_j}\Big] \le 2C \qquad \text{für alle } j \in J_k.$$

Insbesondere erhält man hieraus die Existenz einer von k unabhängigen Konstanten $C_1 > 0$ mit

$$\Big(\frac{\|s_j\|\,\|B_j s_j\|}{s_j^T B_j s_j}\Big)^2 + \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} \le C_1 \qquad \text{für alle } j \in J_k.$$

Nun ist $s_j = t_j p_j = -t_j B_j^{-1} g_j$ und daher

$$\Big(\frac{\|s_j\|\,\|B_j s_j\|}{s_j^T B_j s_j}\Big)^2 = \Big(\frac{\|g_j\|\,\|p_j\|}{g_j^T p_j}\Big)^2, \qquad \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} = -\frac{\|g_j\|^2}{g_j^T p_j}.$$

Folglich ist

$$\delta_j := \min\Big[-\frac{g_j^T p_j}{\|g_j\|^2}, \Big(\frac{g_j^T p_j}{\|g_j\|\,\|p_j\|}\Big)^2\Big] \ge \hat\delta := \frac{1}{C_1} \qquad \text{für alle } j \in J_k.$$

Für alle k ist daher

$$\sum_{j=0}^{k} \delta_j \ge \sum_{j\in J_k} \delta_j \ge \#(J_k)\hat\delta \ge \frac{\hat\delta}{2}(k+1),$$

so dass die Behauptung des Satzes aus dem allgemeinen Konvergenzsatz 1.7 folgt. □□

Nun folgt ein globaler Konvergenzsatz, in welchem die Zielfunktion nur noch als konvex vorausgesetzt wird, siehe M. J. D. Powell (1976). Ferner wird vorausgesetzt, dass die Wolfe-Schrittweite benutzt wird.

Satz 2.8 *Die Zielfunktion f sei konvex und auf einer offenen Obermenge der Niveaumenge L_0 zweimal stetig differenzierbar, ferner existiere eine Konstante $M > 0$ mit $\|\nabla^2 f(x)\| \le M$ für alle $x \in L_0$. Auf die unrestringierte Optimierungsaufgabe (P) wende man das BFGS-Verfahren mit der Wolfe-Schrittweite an. Mit vorgegebenen $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ bestimme man die Schrittweite $t_k > 0$ also so, dass*

$$f(x_k + t_k p_k) \le f(x_k) + \alpha t_k g_k^T p_k, \qquad g_{k+1}^T p_k \ge \beta g_k^T p_k.$$

Ist dann f nach unten beschränkt auf der Niveaumenge L_0, so gilt

$$\liminf_{k\to\infty} \|g_k\| = 0.$$

Beweis: Die Durchführbarkeit des Verfahrens ist auch ohne die gleichmäßige Konvexität der Zielfunktion gesichert, da die Wolfe-Schrittweite verwendet wird.

Zunächst zeigen wir, dass

$$\frac{\|y_k\|^2}{y_k^T s_k} \le M, \qquad k = 0, 1, \ldots.$$

Denn mit der symmetrischen, positiv semidefiniten Matrix

$$G_k := \int_0^1 \nabla^2 f(x_k + \theta s_k)\, d\theta$$

ist

$$\frac{\|y_k\|^2}{y_k^T s_k} = \frac{\|G_k s_k\|^2}{s_k^T G_k s_k} = \frac{(G_k^{1/2} s_k)^T G_k (G_k^{1/2} s_k)}{\|G_k^{1/2} s_k\|^2} \leq \|G_k\| \leq M.$$

Es ist

$$\begin{aligned} \operatorname{tr}(B_{k+1}) &= \operatorname{tr}(B_k) - \frac{\|B_k s_k\|^2}{s_k^T B_k s_k} + \frac{\|y_k\|^2}{y_k^T s_k} \\ &= \operatorname{tr}(B_0) - \sum_{j=0}^{k} \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} + \sum_{j=0}^{k} \frac{\|y_k\|^2}{y_k^T s_k} \\ &\leq c_1(k+1) - \sum_{j=0}^{k} \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} \end{aligned}$$

mit

$$c_1 := \operatorname{tr}(B_0) + M.$$

Wegen $\operatorname{tr}(B_{k+1}) > 0$ und der Ungleichung vom geometrisch arithmetischen Mittel ist

$$\sum_{j=0}^{k} \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} \leq c_1(k+1), \qquad \prod_{j=0}^{k} \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} \leq c_1^{k+1}.$$

Aus $\operatorname{tr}(B_{k+1}) \leq c_1(k+1)$, $\det(B_{k+1}) = (y_k^T s_k / s_k^T B_k s_k) \det(B_k)$ und der Ungleichung vom geometrisch-arithmetischen Mittel folgt ferner, dass $\det(B_{k+1}) \leq (c_1(k+1)/n)^n$ und folglich

$$\prod_{j=0}^{k} \frac{y_j^T s_j}{s_j^T B_j s_j} = \frac{\det(B_{k+1})}{\det(B_0)} \leq c_2^{k+1}$$

für alle $k \in \mathbb{N}$ mit einer hinreichend großen Konstanten $c_2 > 0$. Folglich ist

$$\prod_{j=0}^{k} \frac{\|B_j s_j\|^2 \, y_j^T s_j}{(s_j^T B_j s_j)^2} \leq (c_1 c_2)^{k+1}$$

für alle $k \in \mathbb{N}$. Unter Benutzung der zweiten Bedingung an die Wolfe-Schrittweite erhalten wir, dass

$$\begin{aligned} (c_1 c_2)^{k+1} &\geq \prod_{j=0}^{k} \frac{\|B_j s_j\|^2 \, y_j^T s_j}{(s_j^T B_j s_j)^2} \\ &= \prod_{j=0}^{k} \frac{\|g_j\|^2 \, y_j^T s_j}{(-g_j^T s_j)^2} \\ &\geq (1-\beta)^{k+1} \prod_{j=0}^{k} \frac{\|g_j\|^2}{-g_j^T s_j}. \end{aligned}$$

Im Widerspruch zur Behauptung nehmen wir nun an, es existiere ein $\eta > 0$ mit $\|g_j\| \geq \eta$ für alle j. Dann ist, wieder unter Benutzung der Ungleichung vom geometrisch-arithmetischen Mittel,

$$\frac{(1-\beta)\eta}{c_1 c_2} \leq \Big(\prod_{j=0}^{k}(-g_j^T s_j)\Big)^{1/(k+1)} \leq \frac{1}{k+1}\sum_{j=0}^{k}(-g_j^T s_j).$$

Insbesondere ist $\sum_{j=0}^{k}(-g_j^T s_j) = \infty$. Wegen der ersten Bedingung an die Wolfe-Schrittweite ist andererseits

$$\sum_{j=0}^{k}(-g_j^T s_j) \leq \frac{1}{\alpha}\sum_{j=0}^{k}[f(x_j) - f(x_{j+1})] = \frac{1}{\alpha}[f(x_0) - f(x_{k+1})] \leq \frac{f(x_0) - f_u}{\alpha},$$

wobei f_u eine untere Schranke von f auf L_0 ist. Also ist $\sum_{j=0}^{k}(-g_j^T s_j) < \infty$, der gewünschte Widerspruch ist erreicht. □ □

Bemerkung: Die Konvexität von f wurde nur benutzt, um aus der Beschränktheit von $\|\nabla^2 f(\cdot)\|$ auf der Niveaumenge L_0 auf die Beschränktheit von $\{\|y_k\|^2/y_k^T s_k\}$ zu schließen. Setzt man diese voraus, so ist die Konvexitätsvoraussetzung unnötig. □

Es ist nach wie vor ungeklärt, ob das BFGS-Verfahren auch ohne Konvexitätsvoraussetzung global konvergent ist bzw. die folgende Frage mit "ja" beantwortet werden kann.

- Sei die Zielfunktion f auf der Niveaumenge L_0 nach unten beschränkt. Die Folge $\{x_k\}$ sei durch das BFGS-Verfahren mit Wolfe-Schrittweite gewonnen. Ist dann $\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0$?

Von D.-H. Li, M. Fukushima (2001)[13] ist in diese Richtung ein kleiner Fortschritt erreicht worden. Genauer ist ihr Hauptergebnis der folgende Satz.

Satz 2.9 *Über die Zielfunktion f wird vorausgesetzt, dass die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ beschränkt ist, f auf einer offenen Obermenge von L_0 stetig differenzierbar ist und eine Konstante $\gamma > 0$ mit $\|\nabla f(x) - \nabla f(y)\| \leq \gamma\|x - y\|$ für alle $x, y \in L_0$ existiert. Seien $\epsilon > 0$ und $\alpha > 0$ gegeben. Man betrachte das folgende Verfahren:*

- *Gegeben der Startwert $x_0 \in \mathbb{R}^n$, sei $g_0 := \nabla f(x_0)$. Ferner sei eine symmetrische, positiv definite Matrix $B_0 \in \mathbb{R}^{n\times n}$ gegeben.*
- *Für $k = 0, 1, \ldots$:*
  - *Falls $g_k = 0$, dann: STOP, x_k ist stationäre Lösung von (P).*
  - *Berechne $p_k := -B_k^{-1} g_k$.*

[13] D.-H. Li, M. Fukushima (2001) "On the global convergence of the BFGS method for nonconvex unconstrained optimization problems." SIAM Journal on Optimization 11, 1054–1064.

- *Sei $t_k > 0$ die exakte Schrittweite, Wolfe-Schrittweite oder Armijo-Schrittweite in x_k in Richtung p_k.*
- *Sei $x_{k+1} := x_k + t_k p_k$ und berechne $g_{k+1} := \nabla f(x_{k+1})$.*
- *Mit $s_k := x_{k+1} - x_k$ und $y_k := g_{k+1} - g_k$ sei*

$$B_{k+1} := \begin{cases} B_k - \dfrac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \dfrac{y_k y_k^T}{y_k^T s_k}, & \dfrac{y_k^T s_k}{\|s_k\|^2} \geq \epsilon \|g_k\|^\alpha \\ B_k, & \text{sonst.} \end{cases}$$

Dann gilt: Das Verfahren bricht nach endlich vielen Schritten mit einer stationären Lösung ab oder es liefert eine Folge $\{x_k\}$ mit $\liminf_{k\to\infty} \|g_k\| = 0$.

Beweis: Die Durchführbarkeit des Verfahrens ist offenbar gesichert, es handelt sich um ein Abstiegsverfahren. Es wird angenommen, das Verfahren breche nicht nach endlich vielen Schritten mit einer stationären Lösung ab. Wegen der Sätze 1.2, 1.3 und 1.4 existiert eine Konstante $\theta > 0$ mit

$$f(x_j) - f(x_{j+1}) \geq \theta \delta_j \|g_j\|^2, \qquad j = 0, 1, \ldots,$$

wobei

$$\delta_j := \min\Big[-\frac{g_j^T p_j}{\|g_j\|^2}, \Big(\frac{g_j^T p_j}{\|g_j\| \, \|p_j\|}\Big)^2\Big].$$

Im Widerspruch zur Behauptung nehmen wir an, dass es ein $\eta > 0$ mit $\|g_j\| \geq \eta$ für alle j gibt. Da dann $f(x_j) - f(x_{j+1}) \geq \theta\eta\delta_j$ für alle j, genügt es, die Existenz einer Konstanten $\delta > 0$ mit $\delta_j \geq \delta$ für unendlich viele j zu zeigen. Denn dann erhalten wir einen Widerspruch dazu, dass f auf der Niveaumenge L_0 nach unten beschränkt ist.

Wir definieren die Indexmenge

$$J := \Big\{j : \frac{y_j^T s_j}{\|s_j\|^2} \geq \epsilon \|g_j\|^\alpha\Big\}.$$

Ist J endlich, so ist $B_j = B$ mit einer symmetrischen, positiv definiten Matrix B für fast alle j. In diesem Falle ist daher

$$\delta_j = \min\Big[\frac{s_j^T B s_j}{\|B s_j\|^2}, \Big(\frac{s_j^T B s_j}{\|s_j\| \, \|B s_j\|}\Big)^2\Big] \geq \frac{1}{\lambda_{\max}(B)} \min(1, \lambda_{\min}(B))$$

für fast alle j und die Behauptung richtig. Wir können daher im folgenden annehmen, dass J nicht endlich ist. Zunächst definieren wir die Indexmengen

$$J_k := J \cap \{0, 1, \ldots, k\}, \qquad k = 0, 1, \ldots,$$

und beachten, dass $\lim_{k\to\infty} \#(J_k) = \infty$. Dann ist (wir benutzen wieder die Funktion $\psi(\cdot)$ aus Lemma 2.6, siehe auch den Beweis zu Satz 2.7)

$$\psi(B_{k+1}) \;=\; \psi(B_0) + \sum_{j\in J_k} \Big\{\ln\Big(\frac{s_j^T B_j s_j}{\|s_j\| \, \|B_j s_j\|}\Big)^2 + \Big[1 - \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} + \ln \frac{\|B_j s_j\|^2}{s_j^T B_j s_j}\Big]\Big\}$$

$$
\begin{aligned}
&\quad + \sum_{j\in J_k}\Big(\frac{\|y_j\|^2}{y_j^T s_j} - 1 - \ln\frac{y_j^T s_j}{\|s_j\|^2}\Big) \\
&\le \psi(B_0) + \sum_{j\in J_k}\Big\{\ln\Big(\frac{s_j^T B_j s_j}{\|s_j\|\,\|B_j s_j\|}\Big)^2 + \Big[1 - \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} + \ln\frac{\|B_j s_j\|^2}{s_j^T B_j s_j}\Big]\Big\} \\
&\quad + \sum_{j\in J_k}\Big(\frac{\gamma^2}{\epsilon\eta^\alpha} - 1 - \ln(\epsilon\eta^\alpha)\Big) \\
&\le C\#(J_k) + \sum_{j\in J_k}\Big\{\ln\Big(\frac{s_j^T B_j s_j}{\|s_j\|\,\|B_j s_j\|}\Big)^2 + \Big[1 - \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} + \ln\frac{\|B_j s_j\|^2}{s_j^T B_j s_j}\Big]\Big\}
\end{aligned}
$$

mit einer von k unabhängigen Konstanten $C > 0$. Hieraus erhalten wir die Existenz einer Indexmenge $J_k^* \subset J_k$ mit $\#(J_k^*) \ge \frac{1}{2}\#(J_k)$ und einer Konstanten $C_1 > 0$ mit

$$
\max\Big[\Big(\frac{\|s_j\|\,\|B_j s_j\|}{s_j^T B_j s_j}\Big)^2, \frac{\|B_j s_j\|^2}{s_j^T B_j s_j}\Big] \le C_1 \qquad \text{für alle } j \in J_k^*.
$$

Folglich ist $\delta_j \ge \delta := 1/C_1$ für alle $j \in J_k^*$. Für alle $k \in \mathbb{N}$ ist also

$$
\begin{aligned}
f(x_0) - f(x_{k+1}) &= \sum_{j=0}^{k}[f(x_j) - f(x_{j+1})] \\
&\ge \frac{\theta\eta^2}{C_1}\#(J_k^*) \\
&\ge \frac{\theta\eta^2}{2C_1}\#(J_k).
\end{aligned}
$$

Da $\lim_{k\to\infty}\#(J_k) = \infty$, erhalten wir einen Widerspruch dazu, dass f auf L_0 nach unten beschränkt ist. Der Satz ist bewiesen. □ □

Im letzten Satz dieses Unterabschnitts beweisen wir die globale und superlineare Konvergenz des ungedämpften BFGS-Verfahrens bei gleichmäßig konvexer, quadratischer Zielfunktion.

Satz 2.10 *Gegeben sei die unrestringierte Optimierungsaufgabe*

$$
\text{(P)} \qquad \textit{Minimiere} \quad f(x) := c^T x + \frac{1}{2}x^T Q x, \quad x \in \mathbb{R}^n,
$$

wobei $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Dann ist das auf (P) angewandte ungedämpfte BFGS-Verfahren global und superlinear konvergent.

Beweis: Mit $s_k := x_{k+1} - x_k$ lautet die Update-Formel des BFGS-Verfahrens in diesem Falle

$$
B_{k+1} = B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{(Q s_k)(Q s_k)^T}{s_k^T Q s_k}.
$$

Mit

$$
\tilde{B}_k := Q^{-1/2} B_k Q^{-1/2}, \qquad \tilde{s}_k := Q^{1/2} s_k
$$

erhalten wir

$$\tilde{B}_{k+1} = \tilde{B}_k - \frac{(\tilde{B}_k\tilde{s}_k)(\tilde{B}_k\tilde{s}_k)^T}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k} + \frac{\tilde{s}_k\tilde{s}_k^T}{\|\tilde{s}_k\|^2}.$$

Mit der in Lemma 2.6 eingeführten Funktion $\psi(A) := \mathrm{tr}(A) - \ln\det(A)$ auf der Menge der symmetrischen, positiv definiten $n \times n$-Matrizen ist

$$\begin{aligned}
\psi(\tilde{B}_{k+1}) &= \psi(\tilde{B}_k) - \frac{\|\tilde{B}_k\tilde{s}_k\|^2}{\tilde{s}_k\tilde{B}_k\tilde{s}_k} + 1 - \ln\frac{\|\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k} \\
&= \psi(\tilde{B}_k) + \underbrace{\ln\Big(\frac{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k}{\|\tilde{s}_k\|\,\|\tilde{B}_k\tilde{s}_k\|}\Big)^2}_{\le 0} + \underbrace{\Big[1 - \frac{\|\tilde{B}_k\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k} + \ln\frac{\|\tilde{B}_k\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k}\Big]}_{\le 0} \\
&\le \psi(\tilde{B}_k).
\end{aligned}$$

Daher ist einerseits die nach unten beschränkte Folge $\{\psi(\tilde{B}_k)\}$ monoton nicht wachsend und damit konvergent und insbesondere beschränkt, damit $\{\tilde{B}_k\}$ und $\{\tilde{B}_k^{-1}\}$ sowie $\{B_k\}$ und $\{B_k^{-1}\}$ beschränkt, andererseits $\{\psi(\tilde{B}_k) - \psi(\tilde{B}_{k+1})\}$ eine Nullfolge, daher

$$\lim_{k\to\infty} \ln\Big(\frac{\|\tilde{s}_k\|\,\|\tilde{B}_k\tilde{s}_k\|}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k}\Big)^2 = 0, \qquad \lim_{k\to\infty}\Big[\frac{\|\tilde{B}_k\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k} - 1 - \ln\frac{\|\tilde{B}_k\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k}\Big] = 0,$$

dann

$$\lim_{k\to\infty} \frac{\|\tilde{s}_k\|\,\|\tilde{B}_k\tilde{s}_k\|}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k} = 1, \qquad \lim_{k\to\infty} \frac{\|\tilde{B}_k\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k} = 1$$

und folglich

$$\lim_{k\to\infty} \frac{\|\tilde{B}_k\tilde{s}_k\|}{\|\tilde{s}_k\|} = 1, \qquad \lim_{k\to\infty} \frac{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k}{\|\tilde{s}_k\|^2} = 1.$$

Also ist

$$\Big(\frac{\|(\tilde{B}_k - I)\tilde{s}_k\|}{\|\tilde{s}_k\|}\Big)^2 = \underbrace{\Big(\frac{\|\tilde{B}_k\tilde{s}_k\|}{\|\tilde{s}_k\|}\Big)^2}_{\to 1} - 2\underbrace{\frac{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k}{\|\tilde{s}_k\|^2}}_{\to 1} + 1 \to 0,$$

also

$$\lim_{k\to\infty} \frac{\|(\tilde{B}_k - I)\tilde{s}_k\|}{\|\tilde{s}_k\|} = 0.$$

Nun ist

$$\tilde{s}_k = Q^{1/2}s_k = -Q^{1/2}B_k^{-1}(c + Qx_k) = -Q^{1/2}B_k^{-1}Q(x_k - x^*) = -\tilde{B}_k^{-1}Q^{1/2}(x_k - x^*)$$

und

$$Q^{1/2}(x_{k+1} - x^*) = \tilde{s}_k + Q^{1/2}(x_k - x^*) = (I - \tilde{B}_k)\tilde{s}_k.$$

Folglich ist

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \frac{\|Q^{-1/2}(\tilde{B}_k - I)\tilde{s}_k\|}{\|Q^{-1/2}\tilde{B}_k\tilde{s}_k\|} \le \|Q^{-1/2}\|\,\|Q^{1/2}\|\,\|\tilde{B}_k^{-1}\|\,\frac{\|(\tilde{B}_k - I)\tilde{s}_k\|}{\|\tilde{s}_k\|} \to 0,$$

womit die Behauptung bewiesen ist. □ □

3.2.4 Lokale superlineare Konvergenz des BFGS-Verfahrens

Gegeben sei weiter die unrestringierte Optimierungsaufgabe

(P) $$\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n.$$

Wir setzen über (P) in diesem Abschnitt generell voraus:

(V) Die Zielfunktion f ist auf einer konvexen Umgebung U^* der (isolierten) lokalen Lösung $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbar, die Hessesche $\nabla^2 f$ dort in x^* lipschitzstetig und $B^* := \nabla^2 f(x^*)$ positiv definit.

Grundlage lokaler Konvergenzaussagen für Quasi-Newton-Verfahren ist häufig das auf C. G. Broyden, J. E. Dennis, J. J. Moré (1973)[14] zurückgehende sogenannte Bounded-Deterioration-Theorem. Wir ziehen gegenüber dem üblichen Bounded-Deterioration-Theorem (einen recht übersichtlichen Beweis findet man bei J. E. Dennis, H. J. Martinez, R. A. Tapia (1989)[15]) die folgende einfacher zu beweisende Version vor. Hierbei benutzen wir wieder die auf der Menge der symmetrischen und positiv definiten Matrizen durch $\psi(A) := \operatorname{tr}(A) - \ln\det(A)$ definierte Funktion ψ.

Satz 2.11 *Gegeben sei die unrestringierte Optimierungsaufgabe (P), die Voraussetzung (V) sei erfüllt. Dann existieren positive Konstanten ϵ und α mit der folgenden Eigenschaft: Sind $x_c, x_+ \in B[x^*;\epsilon] := \{x \in \mathbb{R}^n : \|x - x^*\| \le \epsilon\}$ mit $x_c \neq x_+$ und $B_c \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit gegeben, so ist*

$$B_+ := B_c - \frac{(B_c s)(B_c s)^T}{s^T B_c s} + \frac{yy^T}{y^T s}$$

mit

$$s := x_+ - x_c, \qquad y := \nabla f(x_+) - \nabla f(x_c)$$

symmetrisch und positiv definit und

$$\psi(B^{*-1/2} B_+ B^{*-1/2}) \le \psi(B^{*-1/2} B_c B^{*-1/2}) + \alpha\sigma(x_c, x_+),$$

wobei

$$\sigma(u,v) := \max(\|u - x^*\|, \|v - x^*\|).$$

Beweis: Wegen Voraussetzung (V) existiert ein $\epsilon_1 > 0$ derart, dass f auf $B[x^*;\epsilon_1]$ zweimal stetig partiell differenzierbar und $\nabla^2 f$ auf $B[x^*;\epsilon_1]$ in x^* mit einer Lipschitzkonstanten $L > 0$ lipschitzstetig ist. Der kleinste Eigenwert von $B^* = \nabla^2 f(x^*)$ werde mit $\lambda^*_{\min}$ bezeichnet. Seien $x_c, x_+ \in B[x^*;\epsilon_1]$ gegeben. Mit $\tilde{s} := B^{*1/2} s$ und $\tilde{y} := B^{*-1/2} y$

[14] C. G. Broyden, J. E. Dennis, J. J. Moré (1973) "On the local and superlinear convergence of quasi-Newton methods." IMA Journal of Applied Mathematics 12, 223–246.

[15] J. E. Dennis, H. J. Martinez, R. A. Tapia (1989) "Convergence theory for the structured BFGS secant method with an application to nonlinear least squares." JOTA 61, 161–178.

ist dann

$$\begin{aligned}
\|\tilde{y}-\tilde{s}\| &= \|B^{*-1/2}(y-B^*s)\| \\
&\le \frac{1}{\sqrt{\lambda^*_{\min}}}\|y-B^*s\| \\
&= \frac{1}{\sqrt{\lambda^*_{\min}}}\|\nabla f(x_+)-\nabla f(x_c)-\nabla^2 f(x^*)(x_+-x_c)\| \\
&= \frac{1}{\sqrt{\lambda^*_{\min}}}\Big\|\int_0^1 [\nabla^2 f(x_+ + t(x_c-x_+))-\nabla^2 f(x^*)](x_+-x_c)\,dt\Big\| \\
&\le \frac{L}{\sqrt{\lambda^*_{\min}}}\,\|x_+-x_c\| \int_0^1 \|x_+ + t(x_c-x_+)-x^*\|\,dt \\
&= \frac{L}{\sqrt{\lambda^*_{\min}}}\,\|x_+-x_c\| \int_0^1 \|(1-t)(x_+-x^*)+t(x_c-x^*)\|\,dt \\
&\le \frac{L}{2\sqrt{\lambda^*_{\min}}}\,\|x_+-x_c\|\,(\|x_+-x^*\|+\|x_c-x^*\|) \\
&\le \frac{L}{\sqrt{\lambda^*_{\min}}}\,\|x_+-x_c\|\,\max(\|x_c-x^*\|,\|x_+-x^*\|) \\
&\le \frac{L}{\lambda^*_{\min}}\,\|\tilde{s}\|\,\sigma(x_c,x_+).
\end{aligned}$$

Für $x_c, x_+ \in B[x^*;\epsilon_1]$ ist daher

$$\tilde{y}^T\tilde{s} = (\tilde{y}-\tilde{s})^T\tilde{s}+\|\tilde{s}\|^2 \ge \Big[1-\frac{L}{\lambda^*_{\min}}\sigma(x_c,x_+)\Big]\,\|\tilde{s}\|^2.$$

Definiert man daher

$$\epsilon := \min\Big(\epsilon_1, \frac{\lambda^*_{\min}}{2L}\Big),$$

so ist

$$\tilde{y}^T\tilde{s} \ge \frac{1}{2}\,\|\tilde{s}\|^2 \qquad \text{für alle } x_c, x_+ \in B[x^*;\epsilon].$$

Seien nun $x_c, x_+ \in B[x^*;\epsilon]$ mit $x_c \neq x_+$ und eine symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n\times n}$ gegeben. Zur Abkürzung setzen wir

$$\tilde{B}_c := B^{*-1/2}B_cB^{*-1/2}, \qquad \tilde{B}_+ := B^{*-1/2}B_+B^{*-1/2}.$$

Dann ist auch

$$\tilde{B}_+ := \tilde{B}_c - \frac{(\tilde{B}_c\tilde{s})(\tilde{B}_c\tilde{s})^T}{\tilde{s}^T\tilde{B}_c\tilde{s}} + \frac{\tilde{y}\tilde{y}^T}{\tilde{y}^T\tilde{s}}$$

symmetrisch und positiv definit. Hieraus leitet man (siehe den Beweis zu Satz 2.7) ab, dass

$$\psi(\tilde{B}_+) = \psi(\tilde{B}_c) - \frac{\|\tilde{B}_c\tilde{s}\|^2}{\tilde{s}\tilde{B}_c\tilde{s}} + \frac{\|\tilde{y}\|^2}{\tilde{y}^T\tilde{s}} - \ln\frac{\tilde{y}^T\tilde{s}}{\tilde{s}^T\tilde{B}_c\tilde{s}}$$

$$
\begin{aligned}
&= \psi(\tilde{B}_c) + \underbrace{\ln\Big(\frac{\tilde{s}\tilde{B}_c\tilde{s}}{\|\tilde{s}\|\,\|\tilde{B}_c\tilde{s}\|}\Big)^2}_{\leq 0} + \underbrace{\Big[1 - \frac{\|\tilde{B}_c\tilde{s}\|^2}{\tilde{s}\tilde{B}_c\tilde{s}} + \ln\frac{\|\tilde{B}_c\tilde{s}\|^2}{\tilde{s}\tilde{B}_c\tilde{s}}\Big]}_{\leq 0} \\
&\quad + \Big[\frac{\tilde{y}\|^2}{\tilde{y}^T\tilde{s}} - 1 - \ln\frac{\tilde{y}^T\tilde{s}}{\|\tilde{s}\|^2}\Big] \\
&\leq \psi(\tilde{B}_c) + \Big[\frac{\tilde{y}\|^2}{\tilde{y}^T\tilde{s}} - 1 - \ln\frac{\tilde{y}^T\tilde{s}}{\|\tilde{s}\|^2}\Big].
\end{aligned}
$$

Es kommt darauf an, den letzten Term durch ein Vielfaches von $\sigma(x_c, x_+)$ abzuschätzen. Nun ist

$$
\begin{aligned}
\frac{\|\tilde{y}\|^2}{\tilde{y}^T\tilde{s}} - 1 &= \frac{\tilde{y}^T(\tilde{y}-\tilde{s})}{\tilde{y}^T\tilde{s}} \\
&\leq \frac{\|\tilde{y}\|\,\|\tilde{y}-\tilde{s}\|}{\tilde{y}^T\tilde{s}} \\
&\leq \frac{L}{\lambda^*_{\min}}\,\frac{\|\tilde{y}\|\,\|\tilde{s}\|}{\tilde{y}^T\tilde{s}}\sigma(x_c, x_+) \\
&\leq \frac{2L}{\lambda^*_{\min}}\,\frac{\|\tilde{y}\|}{\|\tilde{s}\|}\sigma(x_c, x_+) \\
&\leq \frac{2L}{\lambda^*_{\min}}\,\frac{\|\tilde{y}-\tilde{s}\| + \|\tilde{s}\|}{\|\tilde{s}\|}\sigma(x_c, x_+) \\
&\leq \frac{2L}{\lambda^*_{\min}}\Big(\frac{L}{\lambda^*_{\min}}\sigma(x_c, x_+) + 1\Big)\sigma(x_c, x_+) \\
&\leq \frac{2L}{\lambda^*_{\min}}\Big(\frac{L}{\lambda^*_{\min}}\epsilon + 1\Big)\sigma(x_c, x_+).
\end{aligned}
$$

Weiter ist

$$
\frac{\tilde{y}^T\tilde{s}}{\|\tilde{s}\|^2} = 1 + \frac{(\tilde{y}-\tilde{s})^T\tilde{s}}{\|\tilde{s}\|^2} \geq 1 - \frac{L}{\lambda^*_{\min}}\sigma(x_c, x_+).
$$

Wegen

$$
0 \leq \frac{L}{\lambda^*_{\min}}\sigma(x_c, x_+) \leq \frac{L}{\lambda^*_{\min}}\epsilon \leq \frac{1}{2}
$$

und $-\ln(1-t) \leq 2t$ für alle $t \in [0, 1/2]$ ist

$$
-\ln\frac{\tilde{y}^T\tilde{s}}{\|\tilde{s}\|^2} \leq -\ln\Big(1 - \frac{L}{\lambda^*_{\min}}\sigma(x_c, x_+)\Big) \leq \frac{2L}{\lambda^*_{\min}}\sigma(x_c, x_+).
$$

Mit

$$
\alpha := \frac{2L}{\lambda^*_{\min}}\Big(\frac{L}{\lambda^*_{min}}\epsilon + 2\Big)
$$

ist dann

$$
\psi(B^{*-1/2}B_+B^{*-1/2}) \leq \psi(B^{*-1/2}B_cB^{*-1/2}) + \alpha\sigma(x_c, x_+),
$$

die Behauptung ist bewiesen. □ □

Satz 2.11 ist Haupthilfsmittel des folgenden Satzes, in dem die lokale Q-lineare Konvergenz des ungedämpften BFGS-Verfahrens ausgesagt wird. Genau wie in Satz 2.11 arbeiten wir nicht mit gewichteten Frobenius-Normen sondern mit der ψ-Funktion.

Satz 2.12 *Gegeben sei die unrestringierte Optimierungsaufgabe (P), die Voraussetzung (V) sei erfüllt. Dann gibt es zu jedem $r \in (0,1)$ positive Zahlen $\epsilon(r)$ und $\delta(r)$ mit der folgenden Eigenschaft: Ist $\|x_0 - x^*\| \leq \epsilon(r)$ und $B_0 \in \mathbb{R}^{n\times n}$ eine symmetrische, positiv definite Matrix mit $\psi(B^{*-1/2}B_0B^{*-1/2}) - n \leq \delta(r)$, so ist das (ungedämpfte) BFGS-Verfahren $x_{k+1} := x_k - B_k^{-1}\nabla f(x_k)$ mit*

$$B_{k+1} := B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}$$

und

$$s_k := x_{k+1} - x_k, \qquad y_k := \nabla f(x_{k+1}) - \nabla f(x_k)$$

durchführbar und liefert (o. B. d. A. findet kein vorzeitiger Abbruch statt) eine Folge $\{x_k\}$ mit $\|x_{k+1} - x^*\| \leq r\,\|x_k - x^*\|$ für alle k, die also Q-linear gegen x^* konvergiert. Ferner sind die Folgen $\{\|B_k\|\}$ und $\{\|B_k^{-1}\|\}$ beschränkt.

Beweis: Auch im folgenden wird die Abkürzung $B^* := \nabla^2 f(x^*)$ benutzt. Für eine symmetrische, positiv definite Matrix $B \in \mathbb{R}^{n\times n}$ ist

$$0 \leq \psi(B^{*-1/2}BB^{*-1/2}) - n = \sum_{i=1}^{n} [\underbrace{\lambda_i(B^{*-1/2}BB^{*-1/2}) - 1 - \ln \lambda_i(B^{*-1/2}BB^{*-1/2})}_{\geq 0}],$$

wobei

$$\lambda_1(B^{*-1/2}BB^{*-1/2}) \geq \cdots \geq \lambda_n(B^{*-1/2}BB^{*-1/2})$$

die (positiven) Eigenwerte von $B^{*-1/2}BB^{*-1/2}$ bezeichnen. Wir definieren für positives δ die Menge

$$\mathcal{B}_\delta := \{B \in \mathbb{R}^{n\times n} : B \text{ symmetrisch, positiv definit, } \psi(B^{*-1/2}BB^{*-1/2}) - n \leq \delta\}$$

und anschließend die Funktion $g\colon (0,\infty) \longrightarrow \mathbb{R}$ durch

$$g(\delta) := \sup_{B\in\mathcal{B}_\delta} \|B - B^*\|.$$

Wir überlegen uns, dass die Funktion g die folgenden Eigenschaften besitzt:

- Für alle positiven δ ist $g(\delta)$ positiv. Weiter ist $g\colon (0,\infty) \longrightarrow (0,\infty)$ monoton nicht fallend und $\lim_{\delta\to 0+} g(\delta) = 0$.

Denn: Sei $\delta > 0$ gegeben. Dann ist $B^* + \alpha I \in \mathcal{B}_\delta$ für alle hinreichend kleinen $|\alpha| > 0$ und folglich $g(\delta) > 0$. Sei nun $0 < \delta_1 \leq \delta_2$. Dann ist $\mathcal{B}_{\delta_1} \subset \mathcal{B}_{\delta_2}$ und folglich $g(\delta_1) \leq g(\delta_2)$. Nun zeigen wir, dass $\lim_{\delta\to\delta+} g(\delta) = 0$. Zur Abkürzung setzen wir $h(t) := t - 1 - \ln t$ für $t > 0$. Zu vorgegebenem $\epsilon \in (0,1)$ gibt es ein $\delta(\epsilon) > 0$ mit $|t-1| \leq \epsilon/\|B^*\|$, falls $t > 0$ und $h(t) \leq \delta(\epsilon)$. Ist nun $0 < \delta \leq \delta(\epsilon)$ und $B \in \mathcal{B}_\delta$, so ist insbesondere

$h(\lambda_i(B^{*-1/2}BB^{*-1/2})) \le \delta$, $i = 1, \dots, n$, und folglich $|\lambda_i(B^{*-1/2}BB^{*-1/2})-1| \le \epsilon/\|B^*\|$. Daher ist

$$\begin{aligned}
\|B - B^*\| &= \|B^{*1/2}(I - B^{*-1/2}BB^{*-1/2})B^{*1/2}\| \\
&\le \|B^*\| \, \|I - B^{*-1/2}BB^{*-1/2}\| \\
&= \|B^*\| \max_{i=1,\dots,n} |\lambda_i(I - B^{*-1/2}BB^{*-1/2})| \\
&= \|B^*\| \max_{i=1,\dots,n} |\lambda_i(B^{*-1/2}BB^{*-1/2}) - 1| \\
&\le \|B^*\| \, \epsilon/\|B^*\| \\
&= \epsilon.
\end{aligned}$$

Insbesondere ist $g(\delta) \le \epsilon$ für alle positiven δ mit $\delta \le \delta(\epsilon)$. Damit ist $\lim_{\delta \to 0+} g(\delta) = 0$ und die ganze Hilfsbehauptung bewiesen.

Nun beginnen wir mit dem eigentlichen Beweis. Wegen Satz 2.11 existieren positive Konstanten α_1, ϵ_1 derart, dass für jede symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n\times n}$ und alle $x_c, x_+ \in B[x^*; \epsilon_1]$ mit $x_c \ne x_+$ der BFGS-Update B_+ symmetrisch und positiv definit ist und die Ungleichung

$$\psi(B^{*-1/2}B_+B^{*-1/2}) \le \psi(B^{*-1/2}B_cB^{*-1/2}) + \alpha_1\sigma(x_c, x_+)$$

gilt. Nun sei $r \in (0, 1)$ beliebig vorgegeben. Man bestimme positive $\epsilon = \epsilon(r) \le \epsilon_1$ und $\delta = \delta(r)$ mit

$$\frac{1+r}{\lambda^*_{\min}}[L\epsilon + g(2\delta)] \le r, \qquad \frac{\alpha_1\epsilon}{1-r} \le \delta.$$

Nun sei $(x_0, B_0) \in B[x^*; \epsilon] \times \mathcal{B}_\delta$ beliebig. Durch vollständige Induktion zeigen wir, dass

1. Die Matrix B_k ist symmetrisch und positiv definit, $\psi(B^{*-1/2}B_kB^{*-1/2}) - n \le 2\delta$,
2. $\|B_k^{-1}\| \le (1+r)/\lambda^*_{\min}$,
3. $\|x_{k+1} - x^*\| \le r \, \|x_k - x^*\|$

für $k = 0, 1, \dots$. Ist dies gelungen, so ist der Satz bewiesen. Der Induktionsanfang liegt bei $k = 0$. Die erste Aussage ist nach Wahl von B_0 trivialerweise richtig. Zum Beweis der zweiten beachten wir, dass

$$\|B^{*-1}\| \, \|B_0 - B^*\| \le \frac{1}{\lambda^*_{\min}} g(\delta) \le \frac{1}{\lambda^*_{\min}} g(2\delta) \le \frac{r}{1+r} < 1.$$

Aus dem Störungslemma[16] folgt

$$\|B_0^{-1}\| \le \frac{\|B^{*-1}\|}{1 - r/(1+r)} = \frac{1+r}{\lambda^*_{\min}}.$$

[16] Dieses besagt bekanntlich:
Ist $A \in \mathbb{R}^{n\times n}$ nichtsingulär und $S \in \mathbb{R}^{n\times n}$ mit $\|A^{-1}\| \, \|S\| < 1$, so ist $A + S$ nichtsingulär und

$$\|(A+S)^{-1}\| \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \, \|S\|}.$$

Siehe z. B. J. Werner (1992, S. 25).

Weiter ist

$$\begin{aligned}
\|x_1 - x^*\| &= \|x_0 - B_0^{-1}\nabla f(x_0) - x^*\| \\
&\leq \|B_0^{-1}[\nabla f(x_0) - \underbrace{\nabla f(x^*)}_{=0} - \nabla^2 f(x^*)(x_0 - x^*)]\| \\
&\qquad + \|B_0^{-1}(B_0 - B^*)(x_0 - x^*)\| \\
&\leq \|B_0^{-1}\| \left(L\epsilon + \|B_0 - B^*\|\right) \|x_0 - x^*\| \\
&\leq \frac{1+r}{\lambda_{\min^*}}[L\epsilon + g(\delta)] \|x_0 - x^*\| \\
&\leq \frac{1+r}{\lambda_{\min^*}}[L\epsilon + g(2\delta)] \|x_0 - x^*\| \\
&\leq r \|x_0 - x^*\|.
\end{aligned}$$

Damit ist der Induktionsanfang gelegt. Nun nehmen wir an, die Aussagen 1.–3. seien für $k = 0, \ldots, m-1$ richtig, was nach dem gerade eben bewiesenen jedenfalls für $m = 1$ der Fall ist. Nach Induktionsvoraussetzung ist $x_{m-1}, x_m \in B[x^*; \epsilon]$ und B_{m-1} symmetrisch und positiv definit. Da $x_{m-1} \neq x_m$ folgt aus Satz 2.11, dass $B_m \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit ist. Für $k = 0, \ldots, m-1$ ist weiter

$$\psi(B^{*-1/2} B_{k+1} B^{*-1/2}) - \psi(B^{*-1/2} B_k B^{*-1/2}) \leq \alpha_1 \max(\|x_k - x^*\|, \|x_{k+1} - x^*\|) \leq \alpha_1 \epsilon r^k.$$

Aufsummieren liefert

$$\psi(B^{*-1/2} B_m B^{*-1/2}) - n \leq \underbrace{\psi(B^{*-1/2} B_0 B^{*-1/2}) - n}_{\leq \delta} + \underbrace{\frac{\alpha_1 \epsilon}{1-r}}_{\leq \delta} \leq 2\delta.$$

Damit ist die erste Aussage für $k = m$ bewiesen. Folglich ist $B_m \in \mathcal{B}_{2\delta}$ und daher $\|B_m - B^*\| \leq 2\delta$. Jetzt kann wieder wie beim Induktionsanfang geschlossen werden. Denn wegen

$$\|B^{*-1}\| \|B_m - B^*\| \leq \frac{1}{\lambda^*_{\min}} g(2\delta) \leq \frac{r}{r+1} < 1$$

ist

$$\|B_m^{-1}\| \leq \frac{\|B^{*-1}\|}{1 - r/(1+r)} = \frac{1+r}{\lambda^*_{\min}}.$$

Damit ist auch die zweite Aussage für $k = m$ bewiesen. Auch die dritte kann wie beim Induktionsanfang gezeigt werden. Es ist nämlich

$$\begin{aligned}
\|x_{m+1} - x^*\| &= \|x_m - B_m^{-1}\nabla f(x_m) - x^*\| \\
&\leq \|B_m^{-1}[\nabla f(x_m) - \underbrace{\nabla f(x^*)}_{=0} - \nabla^2 f(x^*)(x_m - x^*)]\| \\
&\qquad + \|B_m^{-1}(B_m - B^*)(x_m - x^*)\| \\
&\leq \|B_m^{-1}\| \left(L\epsilon + \|B_m - B^*\|\right) \|x_m - x^*\| \\
&\leq \frac{1+r}{\lambda_{\min^*}}[L\epsilon + g(2\delta)] \|x_m - x^*\| \\
&\leq r \|x_m - x^*\|.
\end{aligned}$$

Damit ist auch die dritte Aussage für $k = m$ gezeigt und der Satz bewiesen. □ □

In dem folgenden Satz von Dennis-Moré wird eine hinreichende (und wie man zeigen kann auch notwendige) Bedingung für die superlineare Konvergenz eines Quasi-Newton-Verfahrens angegeben.

Satz 2.13 *Gegeben sei die unrestringierte Optimierungsaufgabe (P), die Voraussetzung (V) sei erfüllt. Für eine Folge $\{B_k\} \subset \mathbb{R}^{n\times n}$ nichtsingulärer Matrizen konvergiere die Folge $\{x_k\}$ mit*

$$x_{k+1} := x_k - B_k^{-1}\nabla f(x_k), \qquad k = 0, 1, \ldots,$$

gegen x^*, es sei $\nabla f(x_k) \neq 0$ für alle k. Ist dann

$$(*) \qquad \lim_{k\to\infty} \frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} = 0,$$

so konvergiert die Folge $\{x_k\}$ superlinear gegen x^*.

Beweis: Aus

$$\begin{aligned}
[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k) &= -\nabla f(x_k) - \nabla^2 f(x^*)(x_{k+1} - x_k) \\
&= \nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x^*)(x_{k+1} - x_k) \\
&\qquad - \nabla f(x_{k+1}) \\
&= \int_0^1 [\nabla^2 f(x_k + t(x_{k+1} - x_k)) - \nabla^2 f(x^*)](x_{k+1} - x_k)\, dt \\
&\qquad - \nabla f(x_{k+1})
\end{aligned}$$

folgt

$$\frac{\|\nabla f(x_{k+1})\|}{\|x_{k+1} - x_k\|} \leq \underbrace{\frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|}}_{\to 0} + \underbrace{\int_0^1 [\nabla^2 f(x_k + t(x_{k+1} - x_k)) - \nabla^2 f(x^*)]\, dt}_{\to 0}$$

und damit

$$\lim_{k\to\infty} \frac{\|\nabla f(x_{k+1})\|}{\|x_{k+1} - x_k\|} = 0.$$

Hieraus wollen wir auf die superlineare Konvergenz von $\{x_k\}$ gegen x^* schließen. Wegen $\nabla f(x^*) = 0$ ist

$$\nabla f(x_{k+1}) = \nabla f(x_{k+1}) - \nabla f(x^*) = \underbrace{\int_0^1 \nabla^2 f(x^* + t(x_{k+1} - x^*))\, dt}_{=:G_{k+1}} = G_{k+1}(x_{k+1} - x^*).$$

Wegen $G_k \to \nabla^2 f(x^*)$ ist

$$\|\nabla^2 f(x^*)^{-1}\| \, \|G_{k+1} - \nabla^2 f(x^*)\| \le \frac{1}{2}$$

für alle hinreichend großen k. Das oben schon erwähnte Störungslemma impliziert, dass G_{k+1} für alle hinreichend großen k nichtsingulär ist und $\|G_{k+1}^{-1}\| \le 2\,\|\nabla^2 f(x^*)^{-1}\|$ gilt. Setzt man zur Abkürzung $\beta := 1/(2\,\|\nabla^2 f(x^*)^{-1}\|)$, so ist daher

$$\|\nabla f(x_{k+1})\| = \|G_{k+1}(x_{k+1} - x^*)\| \ge \frac{1}{\|G_{k+1}^{-1}\|}\,\|x_{k+1} - x^*\| \ge \beta\,\|x_{k+1} - x^*\|$$

für alle hinreichend großsen k. Für diese k ist

$$\underbrace{\frac{\|\nabla f(x_{k+1})\|}{\|x_{k+1} - x_k\|}}_{\to 0} \ge \frac{\beta\,\|x_{k+1} - x^*\|}{\|x_{k+1} - x^*\| + \|x_k - x^*\|} = \beta \frac{\|x_{k+1} - x^*\|/\|x_k - x^*\|}{1 + \|x_{k+1} - x^*\|/\|x_k - x^*\|},$$

woraus die superlineare Konvergenz der Folge $\{x_k\}$ gegen x^* folgt. □ □

Zum Nachweis der entscheidenden Voraussetzung $(*)$ in Satz 2.13 geben wir einen Beweis an, der auf R. H. Byrd, J. Nocedal (1989)[17] zurückgeht. Wieder wird die durch $\psi(A) := \mathrm{tr}(A) - \ln\det(A)$ auf der Menge der symmetrischen, positiv definiten $n \times n$-Matrizen A definierte Funktion ψ benutzt und zur Abkürzung $B^* := \nabla^2 f(x^*)$ gesetzt.

Satz 2.14 *Gegeben sei die unrestringierte Optimierungsaufgabe (P), die Voraussetzung (V) sei erfüllt. Das (gedämpfte oder ungedämpfte) BFGS-Verfahren erzeuge eine (o. B. d. A. nicht abbrechende) Folge $\{x_k\}$ mit $\sum_{k=0}^{\infty} \|x_k - x^*\| < \infty$ und eine Folge symmetrischer, positiv definiter Matrizen $\{B_k\}$. Dann sind die Folgen $\{B_k\}$ und $\{B_k^{-1}\}$ beschränkt und es gilt*

$$(*) \qquad \lim_{k\to\infty} \frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} = 0.$$

Beweis: Zur Abkürzung definieren wir

$$\tilde{B}_k := B^{*-1/2} B_k B^{*-1/2}, \qquad \tilde{s}_k := B^{*1/2} s_k, \qquad \tilde{y}_k := B^{*-1/2} y_k,$$

ähnlich wie im Beweis von Satz 2.11. Mit den positiven Konstanten α, ϵ aus Satz 2.11 ist $x_k \in B[x^*; \epsilon]$ und

$$\psi(\tilde{B}_{k+1}) \le \psi(\tilde{B}_k) + \alpha \max(\|x_k - x^*\|, \|x_{k+1} - x^*\|)$$

für alle hinreichend großen k. Wegen $\sum_{k=0}^{\infty} \max(\|x_k - x^*\|, \|x_{k+1} - x^*\|) < \infty$ folgt die Beschränktheit der Folge $\{\psi(\tilde{B}_k)\}$ und hieraus die Beschränktheit von $\{\tilde{B}_k\}$ und

[17] R. H. Byrd, J. Nocedal (1989) "A tool for the analysis of quasi-Newton methods on convex problems." SIAM J. Numer. Anal. 26, 727–739.

$\{\tilde{B}_k^{-1}\}$ (siehe Lemma 2.6) und folglich von $\{B_k\}$ und $\{B_k^{-1}\}$. Im Beweis von Satz 2.11 wurde genauer nachgewiesen, dass

$$\psi(\tilde{B}_{k+1}) \;\leq\; \psi(\tilde{B}_k) + \underbrace{\ln\Big(\frac{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k}{\|\tilde{s}_k\|\,\|\tilde{B}_k\tilde{s}_k\|}\Big)^2}_{\leq 0} + \underbrace{\Big[1 - \frac{\|\tilde{B}_k\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k} + \ln\frac{\|\tilde{B}_k\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k}\Big]}_{\leq 0}$$
$$+ \alpha \max(\|x_k - x^*\|, \|x_{k+1} - x^*\|).$$

Wiederum wegen der Voraussetzung $\sum_{k=0}^{\infty} \|x_k - x^*\| < \infty$ ist

$$\sum_{k=0}^{\infty} \Big\{ \underbrace{\ln\Big(\frac{\|\tilde{B}_k\tilde{s}_k\|\,\|\tilde{s}_k\|}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k}\Big)^2}_{\geq 0} + \underbrace{\Big[\frac{\|\tilde{B}_k\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k} - 1 - \ln\frac{\|\tilde{B}_k\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k}\Big]}_{\geq 0} \Big\} < \infty.$$

Hieraus folgt

$$\lim_{k\to\infty} \ln\Big(\frac{\|\tilde{B}_k\tilde{s}_k\|\,\|\tilde{s}_k\|}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k}\Big)^2 = 0, \qquad \lim_{k\to\infty}\Big[\frac{\|\tilde{B}_k\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k} - 1 - \ln\frac{\|\tilde{B}_k\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k}\Big] = 0,$$

danach

$$\lim_{k\to\infty} \frac{\|\tilde{B}_k\tilde{s}_k\|\,\|\tilde{s}_k\|}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k} = 1, \qquad \lim_{k\to\infty} \frac{\|\tilde{B}_k\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k} = 1.$$

Damit wird

$$\begin{aligned}
\frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|^2}{\|x_{k+1} - x_k\|^2} &= \frac{\|B^{*1/2}(\tilde{B}_k - I)\tilde{s}_k\|^2}{\|B^{*-1/2}\tilde{s}_k\|^2} \\
&\leq \|B^*\|^2 \frac{\|(\tilde{B}_k - I)\tilde{s}_k\|^2}{\|\tilde{s}_k\|^2} \\
&= \|B^*\|^2 \Big(\underbrace{\frac{\|\tilde{B}_k\tilde{s}_k\|^2}{\|\tilde{s}_k\|^2}}_{\to 1} - 2\underbrace{\frac{\tilde{s}_k^T\tilde{B}_k\tilde{s}_k}{\|\tilde{s}_k\|^2}}_{\to 1} + 1 \Big) \\
&\to 0 \qquad \text{für } k \to \infty,
\end{aligned}$$

womit schließlich die Behauptung des Satzes bewiesen ist. □ □

Die Voraussetzung $\sum_{k=0}^{\infty} \|x_k - x^*\| < \infty$ in Satz 2.14 ist offensichtlich erfüllt, wenn die Folge $\{x_k\}$ zumindestens R-linear (siehe Satz 2.7) oder Q-linear (siehe Satz 2.12) gegen x^* konvergiert. Aus Satz 2.12 (lokale Q-lineare Konvergenz des ungedämpften BFGS-Verfahrens), Satz 2.13 (Satz von Dennis-Moré) und Satz 2.14 folgt die lokale superlineare Konvergenz des ungedämpften BFGS-Verfahrens. In der angegebenen Version des Satzes von Dennis-Moré gingen wir von einem ungedämpften Quasi-Newton-Verfahren aus. Es ist aber leicht einzusehen, dass die Bedingung

$$(*) \qquad \lim_{k\to\infty} \frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} = 0$$

auch die superlineare Konvergenz eines gedämpften Quasi-Newton-Verfahrens impliziert, wenn nur die Folge $\{t_k\}$ der Schrittweiten gegen 1 konvergiert, also wenigstens asymptotisch zu einem ungedämpften Verfahren übergegangen wird. Ist t_k die Armijo- oder die Wolfe-Schrittweite, so ist sogar $t_k = 1$ für alle hinreichend großen k. (Hierbei wird bei der Wolfe-Schrittweite natürlich davon ausgegangen, dass die Schrittweite 1 als erste getestet wird.) Die entsprechenden Aussagen werden im nächsten Satz zusammengefasst.

Satz 2.15 *Gegeben sei die unrestringierte Optimierungsaufgabe (P), die Voraussetzung (V) sei erfüllt. Sei $\{B_k\} \subset \mathbb{R}^{n\times n}$ eine Folge symmetrischer, gleichmäßig positiv definiter Matrizen (d. h. $\{B_k^{-1}\}$ ist beschränkt). Die Folge $\{x_k\}$ mit*

$$x_{k+1} := x_k - t_k B_k^{-1} \nabla f(x_k), \qquad k = 0, 1, \ldots,$$

konvergiere gegen x^*, es sei $\nabla f(x_k) \neq 0$ für alle k. Hierbei sei $t_k > 0$ die Armijo- oder die Wolfe-Schrittweite. Ist dann

$$(*) \qquad \lim_{k\to\infty} \frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} = 0,$$

so gilt:

1. *Es ist $t_k = 1$ für alle hinreichend großen k.*
2. *Die Folge $\{x_k\}$ konvergiert superlinear gegen x^*.*

Beweis: Zunächst[18] gehen wir auf die Armijo-Schrittweite ein. Bei vorgegebenem $\alpha \in (0, \frac{1}{2})$ muss gezeigt werden, dass

$$f(x_k + p_k) \leq f(x_k) + \alpha \nabla f(x_k)^T p_k$$

für alle hinreichend großen k. Hierzu überlegen wir uns, dass

$$\lim_{k\to\infty} \frac{f(x_k + p_k) - f(x_k)}{\nabla f(x_k)^T p_k} = \frac{1}{2} > \alpha,$$

woraus die Behauptung folgt. Wegen $x_k \to x^*$ und $p_k \to 0$ (man beachte, dass $\{B_k^{-1}\}$ beschränkt ist) ist die gesamte Verbindungsstrecke zwischen x_k und $x_k + p_k$ für alle hinreichend großen k in der Umgebung U^* von x^* enthalten, auf der f zweimal stetig differenzierbar und $\nabla^2 f(\cdot)$ in x^* noch lipschitzstetig ist. Aus dem Mittelwertsatz folgt die Existenz von $\theta_k \in (0, 1)$ mit

$$f(x_k + p_k) = f(x_k) + \nabla f(x_k)^T p_k + \frac{1}{2} p_k^T \nabla^2 f(x_k + \theta_k p_k) p_k$$

bzw.

$$\frac{f(x_k + p_k) - f(x_k)}{\nabla f(x_k)^T p_k} = \frac{1}{2} - \frac{p_k^T [\nabla^2 f(x_k + \theta_k p_k) - B_k] p_k}{2 p_k^T B_k p_k}.$$

[18] Siehe auch den Beweis von Satz 2.2.

Nun ist

$$\frac{|p_k^T[\nabla^2 f(x_k+\theta_k p_k)-B_k]p_k|}{p_k^T B_k p_k} \leq \|B_k^{-1}\| \Big[\underbrace{\|\nabla^2 f(x_k+\theta_k p_k)-\nabla^2 f(x^*)\|}_{\to 0} + \underbrace{\frac{\|[B_k-\nabla^2 f(x^*)]p_k\|}{\|p_k\|}}_{\to 0}\Big]$$

und hieraus folgt wegen der Beschränktheit von $\{B_k^{-1}\}$ die Behauptung.

Um die entsprechende Aussage auch für die Wolfe-Schrittweite mit vorgegebenen $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ machen zu können, muss zusätzlich nachgewiesen werden, dass $\nabla f(x_k+p_k)^T p_k \geq \beta \nabla f(x_k)^T p_k$ für alle hinreichend großen k. Hierzu überlegen wir uns, dass

$$\lim_{k\to\infty} \frac{\nabla f(x_k+p_k)^T p_k}{\nabla f(x_k)^T p_k} = 0 < \beta.$$

Aus dem Mittelwertsatz folgt diesmal die Existenz von $\eta_k \in (0,1)$ mit

$$\nabla f(x_k+p_k)^T p_k = \nabla f(x_k)^T p_k + p_k^T \nabla^2 f(x_k+\eta_k p_k) p_k.$$

Dann ist

$$\begin{aligned}\left|\frac{\nabla f(x_k+p_k)^T p_k}{\nabla f(x_k)^T p_k}\right| &= \frac{|p_k^T[B_k-\nabla^2 f(x_k+\eta_k p_k)]p_k|}{p_k^T B_k p_k} \\ &\leq \|B_k^{-1}\| \Big[\underbrace{\frac{\|(B_k-\nabla^2 f(x^*))p_k\|}{\|p_k\|}}_{\to 0} + \underbrace{\|\nabla^2 f(x^*)-\nabla^2 f(x_k+\eta_k p_k)\|}_{\to 0}\Big] \\ &\to 0,\end{aligned}$$

womit die Behauptung bewiesen ist.

Damit ist gezeigt, dass bei Verwendung der Armijo- oder Wolfe-Schrittweite das gedämpfte Verfahren nach endlich vielen Schritten in das ungedämpfte übergeht. Die superlineare Konvergenz der Folge $\{x_k\}$ folgt aus Satz 2.13, dem Satz von Dennis-Moré. Insgesamt ist der Satz bewiesen. □ □

3.2.5 Die Implementation des BFGS-Verfahrens

Sieht man einmal von der Berechnung der Schrittweite ab, so besteht die Hauptarbeit bei der Durchführung des BFGS-Verfahrens darin, die Richtung $p := -B_c^{-1} g_c$ zu berechnen. Hierbei ist natürlich $g_c = \nabla f(x_c)$ der Gradient der Zielfunktion in der aktuellen Näherung x_c und $B_c \in \mathbb{R}^{n\times n}$ die aktuelle Update-Matrix, von der wir voraussetzen, dass sie symmetrisch und positiv definit ist. Alternativ ist die Matrix $H_c = B_c^{-1}$ bekannt. Wir wollen verschiedene Möglichkeiten zur Implementation besprechen[19]. Bekannt sind stets $s := x_+ - x_c$ und $y := g_+ - g_c$, wobei hier möglichst

[19] Siehe z. B. J. E. Dennis, R. B. Schnabel (1983, S. 208 ff.), J. Werner (1992, S. 201 ff.), C. Geiger, C. Kanzow (1999, S. 179 ff), J. Nocedal, S. J. Wright (1999, S. 200).

die Wolfe-Schrittweite benutzt wurde, weil diese im Gegensatz zur Armijo-Schrittweite wenigstens theoretisch die wichtige Beziehung $y^T s > 0$ sichert.

Zuerst geben wir die einfachste Methode an. Hier ist die symmetrische, positiv definite Matrix $H_c \in \mathbb{R}^{n\times n}$ bekannt, es wird die neue Matrix H_+ durch

$$\rho := \frac{1}{y^T s}, \qquad H_+ := (I - \rho s y^T) H_c (I - \rho y s^T) + \rho s s^T$$

berechnet. Anschließend kann dann die neue Richtung einfach durch $p_+ := -H_+ g_+$ erhalten werden. Offenbar ist $O(n^2)$ die Anzahl der benötigten arithmetischen Operationen. In einer kleinen Matlab-Funktion könnte das folgendermaßen aussehen:

```
function H_plus=BFGSa(H_c,y,s);
%Input Parameter:
%           H_c    symmetric, positive definite Matrix
%           y,s    standard
%Output:
%           H_plus =inv(B_plus)
%*******************************************************
rho=1/(y'*s); z=H_c*y;
H_plus=H_c+rho*(1+rho*(y'*z))*(s*s')-rho*(z*s'+s*z');
```

Der Nachteil dieser Methode besteht darin, dass wir keine Kontrolle darüber haben, ob die neue Matrix H_+ (bzw. $B_+ = H_+^{-1}$) wieder positiv definit ist. I, Allg. ist es daher vorzuziehen, eine Cholesky-Faktorisierung von B_c "upzudaten". Genauer sei $B_c = L_c L_c^T$ eine Cholesky-Faktorisierung von B_c, also L_c eine untere Dreiecksmatrix mit positiven Diagonalelementen. Gesucht ist eine Cholesky-Faktorisierung $B_+ = L_+ L_+^T$ von

$$B_+ := B_c - \frac{(B_c s)(B_c s)^T}{s^T B_c s} + \frac{y y^T}{y^T s}.$$

Man speichert also gar nicht B_c und B_+, sondern nur die entsprechenden Cholesky-Faktoren L_c und L_+. Die Vorgehensweise könnte die folgende sein:

- Input: Untere Dreiecksmatrix $L_c \in \mathbb{R}^{n\times n}$ mit positiven Diagonalelementen (Cholesky-Faktor von B_c) sowie $y, s \in \mathbb{R}^n$ mit $y^T s > 0$.
- Berechne
$$u := \sqrt{y^T s}\,\frac{L_c^T s}{\|L_c^T s\|_2}, \qquad J_+^T := L_c^T + \frac{u(y - L_c u)^T}{y^T s}.$$
Dann ist $B_+ = J_+ J_+^T$ (siehe den Beweis zu Satz 2.4).
- Berechne eine QR-Zerlegung $J_+^T = Q_+ R_+$, wobei die obere Dreiecksmatrix R_+ positive Diagonalelemente besitzt. Mit $L_+ := R_+^T$ ist dann
$$B_+ = J_+ J_+^T = R_+^T Q_+^T Q_+ R_+ = L_+ L_+^T$$
die gesuchte Cholesky-Zerlegung von B_+.
- Output: Untere Dreiecksmatrix L_+ mit positiven Diagonalelementen (Cholesky-Faktor von B_+).

Hierbei bleibt noch offen, wie man die QR-Zerlegung einer Matrix $A_+ = R + uv^T$ effizient berechnen kann, wenn R eine obere Dreiecksmatrix mit positiven Diagonalelementen ist. Bei der uns interessierenden Anwendung ist $R = L_c^T$ und z. B. (hier hat man Freiheiten bei der Verteilung der Konstanten)

$$u := \sqrt{y^T s}\frac{L_c^T s}{\|L_c^T s\|_2}, \qquad v := \frac{y - L_c u}{y^T s}.$$

Das Ziel erreicht man, indem man A_+ sukzessive von links mit höchstens $2(n-1)$ Givens-Rotationen multipliziert. Sei $m := \max\{i \in \{1,\dots,n\} : u_i \neq 0\}$. Zunächst führt man den Vektor u durch sukzessive Multiplikation mit $m-1$ geeigneten Givens-Rotationen $G_{m-1,m},\dots,G_{12}$, welche der Reihe nach die Komponenten mit den Indizes $m,\dots,2$ annullieren, in ein Vielfaches $u_1 e_1$ des ersten Einheitsvektors über. Die parallel hierzu durchgeführte Multiplikation der oberen Dreiecksmatrix R mit den Givens-Rotationen $G_{m-1,m},\dots,G_{12}$ transformiert diese in eine obere Hessenberg-Matrix, die wir wieder mit R bezeichnen. Wir veranschaulichen uns diesen ersten Schritt im Falle $n = 4$ und $m = 3$, wobei festbleibende Elemente mit $\bullet$, sich verändernde mit $*$ bezeichnet werden.

$$\left(\begin{array}{cccc|c} \bullet & \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet & \bullet \\ & & \bullet & \bullet & \bullet \\ & & & \bullet & \end{array}\right) \xrightarrow{G_{23}} \left(\begin{array}{cccc|c} \bullet & \bullet & \bullet & \bullet & \bullet \\ & * & * & * & * \\ & * & * & * & \\ & & & \bullet & \end{array}\right) \xrightarrow{G_{12}} \left(\begin{array}{cccc|c} * & * & * & * & * \\ & * & * & * & \\ & \bullet & \bullet & \bullet & \\ & & & \bullet & \end{array}\right)$$

Nach Abschluss dieses ersten Schrittes ist $G_{12}\cdots G_{m-1,m}A_+ = R + u_1 e_1 v^T$ mit einer Hessenberg-Matrix R (deren Subdiagonalelemente in den Spalten $m,\dots,n-1$ verschwinden). In einem Zwischenschritt berechnet man $R := R + u_1 e_1 v^T$, wodurch nur die erste Zeile verändert wird. Durch Multiplikation mit weiteren Givens-Rotationen $G_{12},\dots,G_{m-1,m}$ annulliert man schließlich im letzten Schritt die störenden Subdiagonalelemente in den Spalten $1,\dots,m-1$. Hierbei hat man darauf zu achten, dass die erzeugten Diagonalelemente positiv sind. Auch diesen Schritt wollen wir uns für $n = 4$, $m = 3$ veranschaulichen.

$$\left(\begin{array}{cccc} \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet \\ & & & \bullet \end{array}\right) \xrightarrow{G_{12}} \left(\begin{array}{cccc} * & * & * & * \\ & * & * & * \\ & \bullet & \bullet & \bullet \\ & & & \bullet \end{array}\right) \xrightarrow{G_{23}} \left(\begin{array}{cccc} \bullet & \bullet & \bullet & \bullet \\ & * & * & * \\ & & * & * \\ & & & \bullet \end{array}\right)$$

Damit ist ausführlich beschrieben, wie man aus einer Cholesky-Zerlegung der positiv definiten Matrix B_c eine der BFGS-Update-Matrix B_+ berechnen kann. Die hierzu benötigte Anzahl arithmetischer Operationen ist offenbar im wesentlichen propartional zu n^2. Wir geben nun eine Matlab-Funktion an, in der genau das eben beschriebene Umgesetzt wird:

```
function L_plus=CholBFGS(L_c,y,s);
%Input-Parameter
%       L_c    Cholesky-factor of B_c
%       y,s    standard
```

```
%Output-Parameter
%       L_plus Cholesky-factor of BFGS-update B_plus
%*****************************************************
b=y'*s; v=L_c'*s; u=sqrt(b)*v/norm(v); v=(1/b)*(y-L_c*u);
R=L_c';n=length(y);
m=max(find(u));
for i=m:-1:2
    [G,u(i-1:i)]=planerot(u(i-1:i));
    R(i-1:i,i-1:n)=G*R(i-1:i,i-1:n);
end;
R(1,:)=R(1,:)+u(1)*v';
for i=1:m-1
    [G,R(i:i+1,i)]=planerot(R(i:i+1,i));
    R(i:i+1,i+1:n)=G*R(i:i+1,i+1:n);
end;
L_plus=R';
```

Hierbei haben wir die Matlab-Funktion `planerot` benutzt. Nach `help planerot` erhält man die Information:

```
PLANEROT Givens plane rotation.
    [G,Y] = PLANEROT(X), where X is a 2-component column vector,
    returns a 2-by-2 orthogonal matrix G so that Y = G*X has Y(2) = 0.

    See also QRINSERT, QRDELETE.
```

Dies ist keine sogenannte built-in function, sondern sie ist selbst in Matlab geschrieben, und man kann sie sich mittels `type planerot` oder `edit planerot` ansehen (im Gegensatz zur built-in function `norm`). Sieht man einmal vom Kommentar ab, den wir oben schon angegeben haben, so sieht diese Funktion folgendermaßen aus:

```
function [G,x] = planerot(x)
if x(2) ~= 0
   r = norm(x);
   G = [x'; -x(2) x(1)]/r;
   x = [r; 0];
else
   G = eye(2);
end
```

Wenn die zweite Komponente x_2 des Vektors $x \in \mathbb{R}^2$ also von Null verschieden ist, so wird

$$G := \frac{1}{\|x\|_2} \begin{pmatrix} x_1 & x_2 \\ -x_2 & x_1 \end{pmatrix}$$

gesetzt, so dass in diesem Falle

$$Gx = \frac{1}{\|x\|_2} \begin{pmatrix} x_1 & x_2 \\ -x_2 & x_1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \|x\|_2 \\ 0 \end{pmatrix},$$

ferner wird (bei einem Aufruf `[G,y]=planerot(x)`) $y := (\|x\|_2, 0)^T$ gesetzt. Andernfalls wird G gleich der 2×2-Identität gesetzt (und $y = x$ gesetzt bzw. x nicht verändert).

Man erkennt, dass im obigen Programm die Diagonalelemente des Outputs positiv sind.

Nun wollen wir eine zweite Möglichkeit angeben, die Cholesky-Zerlegung des BFGS-Updates zu berechnen. Diese benutzt die Matlab-Funktion `cholupdate` (eine built-in function, man kann sie sich also nicht ansehen). Die folgende Funktion sollte einfach zu verstehen sein:

```
function L_plus=BFGSb(L_c,y,s);
%Input-Parameter
%       L_c    Cholesky-factor of B_c
%       y,s    standard
%Output-Parameter
%       L_plus Cholesky-factor of BFGS-update B_plus
%*****************************************************
z=y/sqrt(y'*s);
L_plus=cholupdate(L_c',z)';
v=L_c'*s;z=L_c*v/norm(v);
L_plus=cholupdate(L_plus',z,'-')';
```

In zwei Schritten wird also die Cholesky-Zerlegung des BFGS-Updates B_+ aus einer von B_c berechnet.

Einige wenige Worte wollen wir nun noch zur Wahl der Startmatrix B_0 (bzw. $H_0 = B_0^{-1}$) verlieren. Wenn man gar nichts besseres weiß, so setzt man $B_0 = I$ bzw. $L_0 = I$ und $H_0 = I$, so dass am Anfang ein Gradientenschritt durchgeführt wird. Bei J. E. Dennis, R. B. Schnabel (1983, S. 209) wird auch noch die Wahl $B_0 := |f(x_0)|I$ bzw. $L_0 = \sqrt{|f(x_0)|}I$ angegeben. Hiermit ist es nun möglich, ein einfaches function-file `BFGS.m` zu schreiben. Dies könnte z. B. folgendermaßen aussehen:

```
function [x_min,iter]=BFGS(fun,x_0,max_iter,tol);
%Input-Parameter:
%            fun        function to be minimized
%            x_0        starting value
%            max_iter   maximal number of iterations
%            tol        If norm(gradient)<=tol: exit
%Output-Parameter:
%            x_min      (local) solution
%            iter       number of iterations
%*****************************************************
x_c=x_0;
[f_c,g_c]=feval(fun,x_c);iter=0;
L_c=sqrt(abs(f_c))*eye(length(x_0));
while (norm(g_c)>tol)&(iter<max_iter)
     p=-(L_c'\(L_c\g_c)); t=Wolfe(x_c,p,fun); x_plus=x_c+t*p;
     [f_plus,g_plus]=feval(fun,x_plus);
     s=x_plus-x_c; y=g_plus-g_c;
     L_c=CholBFGS(L_c,y,s); g_c=g_plus; x_c=x_plus;
     iter=iter+1;
end;
if (norm(g_c)<=tol)
    x_min=x_c;
end;
```

Z. B. liefert der Aufruf

```
[x,iter]=BFGS('Rosenbrock',[-1.2;1],100,1e-8);
```

$$x = \begin{pmatrix} 1.00000000000062 \\ 1.00000000000132 \end{pmatrix}$$

und die Information, dass hierzu 35 Iterationen benötigt wurden.

3.2.6 Das L-BFGS-Verfahren

Das BFGS-Verfahren (und andere Quasi-Newton-Verfahren) verlangt die Speicherung wenigstens einer $n \times n$-Matrix, nämlich der Approximation an die Hessesche (bzw. ihrer Inversen oder eines Cholesky-Faktors). Das kann für große n problematisch werden. Wir schildern daher in diesem Unterabschnitt das L-BFGS-Verfahren (Limited-memory BFGS-method, BFGS-Verfahren mit beschränktem Gedächtnis). Dieses Verfahren speichert statt einer $n \times n$-Matrix lediglich einige Vektoren der Länge n. Ein Schritt des BFGS-Verfahren für die unrestringierte Optimierungsaufgabe, die glatte Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ auf dem $\mathbb{R}^n$ zu minimieren, hat die Form

$$x_{k+1} := x_k - t_k H_k \nabla f(x_k), \qquad k = 0, 1, \ldots,$$

wobei t_k die Schrittweite ist und H_k in jeder Iteration mit der Formel

$$H_{k+1} = U_k^T H_k U_k + \rho_k s_k s_k^T$$

upgedatet wird, wobei

$$\rho_k := \frac{1}{y_k^T s_k}, \qquad U_k := I - \rho_k y_k s_k^T$$

und (wie immer)

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

Daher ist

$$\begin{aligned}
H_1 &= U_0^T H_0 U_0 + \rho_0 s_0 s_0^T, \\
H_2 &= U_1^T H_1 U_1 + \rho_1 s_1 s_1^T \\
&= U_1^T (U_0^T H_0 U_0 + \rho_0 s_0 s_0^T) U_1 + \rho_1 s_1 s_1^T \\
&= (U_0 U_1)^T H_0 (U_0 U_1) + \rho_0 (U_1^T s_0)(U_1^T s_0)^T + \rho_1 s_1 s_1^T, \\
H_3 &= U_2^T H_2 U_2 + \rho_2 s_2 s_2^T \\
&= U_2^T [(U_0 U_1)^T H_0 (U_0 U_1) + \rho_0 (U_1^T s_0)(U_1^T s_0)^T + \rho_1 s_1 s_1^T] U_2 + \rho_2 s_2 s_2^T \\
&= (U_0 U_1 U_2)^T H_0 (U_0 U_1 U_2) + \rho_0 ((U_1 U_2)^T s_0)((U_1 U_2)^T s_0)^T \\
&\qquad + \rho_1 (U_2^T s_1)(U_2^T s_1)^T + \rho_2 s_2 s_2^T \\
&\vdots \\
H_k &= (U_0 \cdots U_{k-1})^T H_0 (U_0 \cdots U_{k-1})
\end{aligned}$$

$$
\begin{aligned}
&+ \rho_0((U_1 \cdots U_{k-1})^T s_0)((U_1 \cdots U_{k-1})^T s_0)^T \\
&+ \rho_1((U_2 \cdots U_{k-2})^T s_1)((U_2 \cdots U_{k-2})^T s_1)^T \\
&\vdots \\
&+ \rho_{k-2}((U_{k-1})^T s_{k-2})((U_{k-1})^T s_{k-2})^T \\
&+ \rho_{k-1} s_{k-1} s_{k-1}^T.
\end{aligned}
$$

Hierbei ist natürlich

$$\rho_j := \frac{1}{y_j^T s_j}, \qquad U_j := I - \rho_j y_j s_j^T \qquad (j = 0, \ldots, k-1).$$

Also lässt sich H_k aus H_0 und den Vektorpaaren (s_j, y_j), $j = 0, \ldots, k-1$, berechnen. Anschließend kann dann die Richtung

$$p_k := -H_k \nabla f(x_k)$$

berechnet werden. Beim auf J. NOCEDAL (1980)[20] zurückgehenden L-BFGS-Verfahren gibt man sich ein festes $m \in \mathbb{N}$ vor und behält nur die letzten m Vektorpaare (s_j, y_j), $j = k-m, \ldots, k-1$, mit denen die entsprechenden Skalare ρ_j und Matrizen U_j berechnet werden können. Alle vorherigen Informationen, also $\rho_0, \ldots, \rho_{k-m-1}$ und $U_0, \ldots, U_{k-m-1}$, werden vergessen bzw. in der obigen Rekursionsformel durch Nullen (Skalare) bzw. Einheitsmatrizen (Matrizen) ersetzt. Mit $m_k := \min(k, m)$ lautet die Update-Formel des BFGS-Verfahrens mit beschränktem Gedächtnis bzw. der L-BFGS-Update also

$$
\begin{aligned}
H_k &:= (U_{k-m_k} \cdots U_{k-1})^T H_0^{(k)} (U_{k-m_k} \cdots U_{k-1}) \\
&\qquad + \sum_{j=k-m_k}^{k-1} \rho_j ((U_{j+1} \cdots U_{k-1})^T s_j)((U_{j+1} \cdots U_{k-1})^T s_j)^T \\
&= (U_{k-1}^T \cdots U_{k-m_k}^T) H_0^{(k)} (U_{k-m_k} \cdots U_{k-1}) \\
&\qquad + \sum_{j=k-m_k}^{k-1} \rho_j (U_{k-1}^T \cdots U_{j+1}^T) s_j s_j^T (U_{j+1} \cdots U_{k-1}).
\end{aligned}
$$

Hierbei soll die Schreibweise $H_0^{(k)}$ statt H_0 ausdrücken, dass diese Matrix von k abhängig sein kann. Auf die effiziente Berechnung der Suchrichtung $p_k := -H_k \nabla f(x_k)$ wollen wir jetzt eingehen. Diese kann durch den folgenden Algorithmus geschehen. Hierbei bezeichne k die aktuelle Iterationsstufe, ferner sei $m_k := \min(m, k)$.

- Input: Die Paare $(s_j, y_j) \in \mathbb{R}^n \times \mathbb{R}^n$, hiermit die Skalare $\rho_j := 1/y_j^T s_j$, $j = k - m_k, \ldots, k-1$, ferner die (symmetrische und positiv definite) Matrix $H_0^{(k)}$ (gewöhnlich eine Diagonalmatrix, z. B. $H_0^{(k)} = \gamma_k I$ mit $\gamma_k := y_{k-1}^T s_{k-1} / y_{k-1}^T y_{k-1}$) und $q := -\nabla f(x_k)$.
- Für $j = k-1, \ldots, k-m_k$:

[20] J. NOCEDAL (1980) "Updating quasi-Newton matrices with limited storage." Mathematics of Computation 35, 773–782.

  - Berechne $\alpha_j := \rho_j s_j^T q$;
  - Berechne $q := q - \alpha_j y_j$;

- Berechne $p := H_0^{(k)} q$;
- Für $j = k - m_k, \ldots, k-1$:
  - Berechne $\beta := \rho_j y_j^T p$;
  - Berechne $p := p + (\alpha_j - \beta) s_j$;
- Output: Es ist $p = -H_k \nabla f(x_k)$.

Es ist leicht nachzuweisen, dass dieser Algorithmus das Verlangte tut. Durch ihn wird deutlich, dass es nicht nötig ist, $n \times n$-Matrizen zu speichern, sondern dass es genügt, zur Berechnung von p_k die $2m$ Vektoren (y_j, s_j), $j = k-m, \ldots, k-1$, der Länge n im Gedächtnis zu behalten. Wenn also im folgenden etwa von der "Berechnung der Matrix H_k" gesprochen wird, so soll dies nicht suggerieren, das diese Matrix wirklich berechnet wird, sie erzeugt lediglich die Suchrichtung und es ist natürlich wichtig, Eigenschaften dieser Matrizen zu kennen.

Lemma 2.16 *Es gilt die Quasi-Newton-Gleichung $H_k y_{k-1} = s_{k-1}$ und mit $H_0^{(k)}$ ist auch H_k symmetrisch und positiv definit.*

Beweis: Wegen

$$U_{k-1} y_{k-1} = \Big(I - \frac{y_{k-1} s_{k-1}^T}{y_{k-1}^T s_{k-1}} \Big) y_{k-1} = 0$$

ist

$$\begin{aligned} H_k y_{k-1} &= (U_{k-m_k} \cdots U_{k-1})^T H_0^{(k)} U_{k-m_k} \cdots \underbrace{U_{k-1} y_{k-1}}_{=0} \\ &\quad + \sum_{j=k-m_k}^{k-1} \rho_j (U_{k-1}^T \cdots U_{j+1}^T) s_j s_j^T (U_{j+1} \cdots U_{k-1}) y_{k-1} \\ &= \sum_{j=k-m_k}^{k-2} \rho_j (U_{k-1}^T \cdots U_{j+1}^T) s_j s_j^T \underbrace{(U_{j+1} \cdots U_{k-1}) y_{k-1}}_{=0} + \underbrace{\rho_{k-1} s_{k-1}^T y_{k-1}}_{=1} s_{k-1} \\ &= s_{k-1}, \end{aligned}$$

also ist die Quasi-Newton-Gleichung erfüllt. Nun zeigen wir, dass H_k positiv definit ist (die Symmetrie ist offensichtlich). Wegen $\rho_j > 0$, $j = k - m_k, \ldots, k-1$, ist H_k die Summe von positiv semidefiniten Matrizen und daher selbst positiv semidefinit. Wir können annehmen, dass $k > m$ und damit $m_k = m$ (andernfalls stimmt H_k mit einem normalen BFGS-Update überein und die Aussage ist bekanntlich richtig). Ist $x^T H_k x = 0$, so folgt aus der positiven Definitheit von $H_0^{(k)}$, dass

$$(U_{k-m} \cdots U_{k-1}) x = 0, \qquad s_j^T (U_{j+1} \cdots U_{k-1}) x = 0 \qquad (j = k-m, \ldots, k-1).$$

Aus $s_{k-1}^T x = 0$ (setze $j = k-1$) folgt $U_{k-1}x = x$, also ist auch

$$(U_{k-m} \cdots U_{k-2})x = 0, \qquad s_j^T (U_{j+1} \cdots U_{k-2})x = 0 \qquad (j = k-m, \ldots, k-2).$$

In dieser Weise kann man fortfahren und erhält nach endlich vielen Schritten, dass $x = 0$. Damit ist das Lemma bewiesen. □ □

Bemerkung: Die Matrix H_k kann man in $m_k := \min(k, m)$ Schritten aus $H_0^{(k)}$ auf die folgende Weise erhalten:

- Setze $H_k^{(0)} := H_0^{(k)}$.
- Für $i = 0, \ldots, m_k - 1$:
  - Berechne

$$H_k^{(i+1)} := \Big(I - \frac{y_{k-m_k+i} s_{k-m_k+i}^T}{y_{k-m_k+i}^T s_{k-m_k+i}}\Big)^T H_k^{(i)} \Big(I - \frac{y_{k-m_k+i} s_{k-m_k+i}^T}{y_{k-m_k+i}^T s_{k-m_k+i}}\Big) + \frac{s_{k-m_k+i} s_{k-m_k+i}^T}{y_{k-m_k+i}^T s_{k-m_k+i}}.$$

- Setze $H_k := H_k^{(m_k)}$.

Hieraus könnte man erneut einen Beweis dafür erhalten, dass H_k die positive Definitheit von $H_0^{(k)}$ erbt, da man leicht durch Induktion die positive Definitheit von $H_k^{(i)}$, $i = 0, \ldots, m_k$, zeigt. Für uns wichtiger ist, dass wir hiermit auch ein Verfahren zur Berechnung von $B_k := H_k^{-1}$ aus $B_0^{(k)} := (H_0^{(k)})^{-1}$ erhalten. Man beachte nämlich, dass $H_k^{(i+1)}$ die Form eines inversen BFGS-Updates von $H_k^{(i)}$ besitzt. Daher ist das folgende Verfahren naheliegend:

- Mit $B_0^{(k)} := (H_0^{(k)})^{-1}$ setze $B_k^{(0)} := B_0^{(k)}$.
- Für $i = 0, \ldots, m_k - 1$:
  - Berechne

$$B_k^{(i+1)} := B_k^{(i)} - \frac{(B_k^{(i)} s_{k-m_k+i})(B_k^{(i)} s_{k-m_k+i})^T}{s_{k-m_k+i}^T B_k^{(i)} s_{k-m_k+i}} + \frac{y_{k-m_k+i} y_{k-m_k+i}^T}{y_{k-m_k+i}^T s_{k-m_k+i}}.$$

- Setze $B_k := B_k^{(m_k)}$.

Sehr leicht erhält man $B_k^{(i)} = (H_k^{(i)})^{-1}$, $i = 0, \ldots, m_k$, siehe z. B. Satz 2.4. □

Das L-BFGS-Verfahren sieht daher folgendermaßen aus:

- Gegeben $x_0 \in \mathbb{R}^n$ und $m \in \mathbb{N}$. Setze $g_0 := \nabla f(x_0)$.
- Für $k = 0, 1, \ldots$:
  - Falls $g_k = 0$, dann STOP.

- Wähle symmetrische und positiv definite Matrix $H_0^{(k)}$.
- Berechne $p_k := -H_k g_k$ nach obigem Algorithmus.
- Sei $t_k > 0$ die exakte Schrittweite, die Wolfe- oder die Armijo-Schrittweite in x_k in Richtung p_k.
- Setze $x_{k+1} := x_k + t_k p_k$ und berechne $g_{k+1} := \nabla f(x_{k+1})$.
- Falls $k > m$: Vergiss das Paar (y_{k-m}, s_{k-m}).
- Berechne und speichere $s_k := x_{k+1} - x_k$, $y_k := g_{k+1} - g_k$.

Hierbei hat man darauf zu achten, dass $y_k^T s_k > 0$ für alle k gilt, was für eine gleichmäßig konvexe Zielfunktion stets der Fall ist, während es ohne Konvexitätsvoraussetzungen an f jedenfalls noch für die exakte Schrittweite und die Wolfe-Schrittweite gilt.

Nun wollen wir einen globalen Konvergenzsatz für das L-BFGS-Verfahren formulieren und beweisen, siehe D. C. Liu, J. Nocedal (1989, Theorem 7.1)[21].

Satz 2.17 *Gegeben sei die unrestringierte Optimierungsaufgabe (P), die Voraussetzungen (K) aus Satz 2.7 seien erfüllt. Weiter seien die Folgen symmetrischer, positiv definiter Matrizen $\{H_0^{(k)}\}$ und $\{B_0^{(k)}\}$ mit $B_0^{(k)} := (H_0^{(k)})^{-1}$ beschränkt. Dann gilt: Das L-BFGS-Verfahren bricht nach endlich vielen Schritten mit der Lösung x^* von (P) ab oder es liefert eine Folge $\{x_k\}$, die R-linear gegen x^* konvergiert.*

Beweis: Die Durchführbarkeit des Verfahrens ist natürlich gesichert, man vergleiche die entsprechenden Aussagen zu Beginn des Beweises von Satz 2.7. Die Matrizen H_k und daher auch $B_k := H_k^{-1}$ sind symmetrisch und positiv definit. Wir benutzen wieder die auf der Menge der symmetrischen, positiv definiten $n \times n$-Matrizen durch $\psi(A) := \operatorname{tr}(A) - \ln\det(A)$ definierte Funktion ψ. Wir zeigen die Beschränktheit der Matrizenfolgen $\{H_k\}$ und $\{B_k\}$ dadurch, dass wir die Beschränktheit von $\{\psi(B_k)\}$ nachweisen. Hierzu wiederum benutzen wir die obige Bemerkung, dass B_k in m_k Schritten aus $B_0^{(k)}$ berechnet werden kann. Mit den dortigen Bezeichnungen haben wir für $i = 0, \ldots, m_k - 1$ die folgende Ungleichungskette (siehe die entsprechenden Teile im Beweis von Satz 2.7)

$$\begin{aligned}
\psi(B_k^{(i+1)}) &= \psi(B_k^{(i)}) + \underbrace{\ln\Big(\frac{s_{k-m_k+i}^T B_k^{(i)} s_{k-m_k+i}}{\|s_{k-m_k+i}\|\,\|B_k^{(i)} s_{k-m_k+i}\|}\Big)^2}_{\le 0} \\
&\quad + \underbrace{\Big[1 - \frac{\|B_k^{(i)} s_{k-m_k+i}\|^2}{s_{k-m_k+i}^T B_k^{(i)} s_{k-m_k+i}} + \ln\frac{\|B_k^{(i)} s_{k-m_k+i}\|^2}{s_{k-m_k+i}^T B_k^{(i)} s_{k-m_k+i}}\Big]}_{\le 0} \\
&\quad + \Big[\frac{\|y_{k-m_k+i}\|^2}{y_{k-m_k+i}^T s_{k-m_k+i}} - 1 - \ln\frac{y_{k-m_k+i}^T s_{k-m_k+i}}{\|s_{k-m_k+i}\|^2}\Big]
\end{aligned}$$

[21] D. C. Liu, J. Nocedal (1989) "On the limited memory BFGS method for large scale optimization." Mathematical Programming 45, 503–528.

$$\begin{aligned}
&\leq \psi(B_k^{(i)}) + \Big[\frac{\|y_{k-m_k+i}\|^2}{y_{k-m_k+i}^T s_{k-m_k+i}} - 1 - \ln\frac{y_{k-m_k+i}^T s_{k-m_k+i}}{\|s_{k-m_k+i}\|^2}\Big] \\
&\leq \psi(B_k^{(i)}) + \frac{\gamma^2}{c} - 1 - \ln c \\
&\leq \psi(B_k^{(0)}) + (i+1)\Big[\frac{\gamma^2}{c} - 1 - \ln c\Big],
\end{aligned}$$

wobei die Konstanten c, γ in Voraussetzung (K) erklärt sind. Wegen $B_0^{(k)} = B_k^{(0)}$ und $B_k = B_k^{(m_k)}$ sowie $m_k \leq m$ ist

$$\psi(B_k) \leq \psi(B_0^{(k)}) + m\Big[\frac{\gamma^2}{c} - 1 - \ln c\Big].$$

Dies zeigt die Beschränktheit von $\{\psi(B_k)\}$ und damit die von $\{B_k\}$ und $\{H_k\}$. Dies wiederum impliziert, dass

$$\delta_k := \min\Big[-\frac{g_k^T p_k}{\|g_k\|^2}, \Big(\frac{g_k^T p_k}{\|g_k\|\,\|p_k\|}\Big)^2\Big] = \min\Big[\frac{g_k^T H_k g_k}{\|g_k\|^2}, \Big(\frac{g_k^T H_k g_k}{\|g_k\|\,\|H_k g_k\|}\Big)^2\Big]$$

unabhängig von k durch eine positive Konstante nach unten beschränkt ist. Aus dem allgemeinen Konvergenzsatz 1.7 folgt die Behauptung des Satzes. □ □

Bemerkung: Z. B. kann man

$$H_0^{(k+1)} := \frac{y_k^T s_k}{\|y_k\|^2} I$$

setzen. Die an $\{H_0^{(k)}\}$ und $\{B_0^{(k)}\}$ gestellte Beschränktheitsforderung ist bei gleichmäßig konvexer Zielfunktion erfüllt. Ist darüber hinaus $m = 1$ und daher $m_k = 1$ für alle k, so wird

$$\begin{aligned}
H_{k+1} &= \Big(I - \frac{y_k s_k^T}{y_k^T s_k}\Big)^T \Big(\frac{y_k^T s_k}{\|y_k\|^2} I\Big)\Big(I - \frac{y_k s_k^T}{y_k^T s_k}\Big) + \frac{s_k s_k^T}{y_k^T s_k} \\
&= \frac{y_k^T s_k}{\|y_k\|^2} I + \frac{2}{y_k^T s_k}\, s_k s_k^T - \frac{1}{\|y_k\|^2}(s_k y_k^T + y_k s_k^T).
\end{aligned}$$

Daher ist

$$p_{k+1} = -H_{k+1} g_{k+1} = -\frac{y_k^T s_k}{\|y_k\|^2} g_{k+1} - \Big(2\frac{s_k^T g_{k+1}}{y_k^T s_k} - \frac{y_k^T g_{k+1}}{\|y_k\|^2}\Big) s_k + \frac{s_k^T g_{k+1}}{\|y_k\|^2} y_k.$$

Mit $p_0 := -g_0$ ist dies genau die Richtungsstrategie eines Verfahrens von D. F. Shanno (1978)[22]. Daher ist Satz 2.17 Verallgemeinerung eines Konvergenzsatzes von D. F. Shanno (1978, Theorem 6) (siehe auch J. Werner (1992, S. 227)). □

Jetzt wollen wir noch kurz auf eine Matlab-Implementation des L-BFGS-Verfahrens eingehen. Wir geben das Function-File `LBFGS.m` an, dem eine Funktion `LBFGSdir` angehängt ist.

[22] D. F. Shanno (1978) On the convergence of a new conjugate gradient algorithm. SIAM J. Numer. Anal. 15, 1247–1257.

```
function [x_min,iter]=LBFGS(fun,x_0,m,max_iter,tol);
%Input-Parameter:
%            fun        function to be minimized
%            x_0        starting value
%            m          number of vector pairs to be stored
%            max_iter   maximal number of iterations
%            tol        If norm(gradient)<=tol: exit
%Output-Parameter:
%            x_min      (local) solution
%            iter       number of iterations
%************************************************************
x_c=x_0;
[f_c,g_c]=feval(fun,x_c);k=0;
S=[]; Y=[]; rho=[];
while (norm(g_c)>tol)&(k<max_iter)
   if k==0
       p=-g_c;
   else
       p=LBFGSdir(S,Y,rho,g_c);
   end;
   t=Wolfe(x_c,p,fun); x_plus=x_c+t*p;
   [f_plus,g_plus]=feval(fun,x_plus);
   s=x_plus-x_c; y=g_plus-g_c;
   S=[S, s]; Y=[Y, y];rho=[rho, 1/(y'*s)];
   if k>=m
       S=S(:,2:m+1);Y=Y(:,2:m+1);rho=rho(2:m+1);
   end;
   g_c=g_plus; x_c=x_plus;
   k=k+1;
end;
if (norm(g_c)<=tol)
   x_min=x_c;iter=k;
end;
%************************************************************
function p=LBFGSdir(S,Y,rho,g_c);
%Input-Parameter
%          S           Columns are s_{k-m_k},...,s_{k-1}
%          Y           Columns are y_{k-m_k},...,y_{k-1}
%          rho         Components are rho_{k-m_k},...,rho_{k-1}
%          g_c         gradient at the current iterate
%Output-Parameter
%          p           L-BFGS direction
%************************************************************
[n,m_k]=size(S); q=-g_c;
for j=m_k:-1:1
   alpha(j)=rho(j)*(S(:,j)'*q);
   q=q-alpha(j)*Y(:,j);
end;
gamma=S(:,m_k)'*Y(:,m_k)/(Y(:,m_k)'*Y(:,m_k));
p=gamma*q;
for j=1:m_k
   beta=rho(j)*(Y(:,j)'*p);
   p=p+(alpha(j)-beta)*S(:,j);
end;
```

Wir testen die Funktion durch den Aufruf

```
[x,iter]=LBFGS('Rosenbrock',[-1.2;1],1,100,1e-8);
```

und erhalten

$$x = \begin{pmatrix} 0.99999999999710 \\ 0.99999999999091 \end{pmatrix}, \qquad \text{iter} = 44.$$

Mit dem Aufruf

```
[x,iter]=LBFGS('Rosenbrock',[-1.2;1],2,100,1e-8);
```

erhalten wir das Ergebnis

$$x = \begin{pmatrix} 1.00000000024967 \\ 1.00000000050844 \end{pmatrix}, \qquad \text{iter} = 43.$$

Natürlich ist dies eigentlich kein gutes Beispiel für die Anwendung des L-BFGS-Verfahrens.

3.2.7 Aufgaben

1. Mit dem ungedämpften Newton-Verfahren und dem durch die Wolfe- bzw. die Armijo-Schrittweite gedämpften Newton-Verfahren löse man die Aufgabe:

   (P) Minimiere $f(x) := 1.1x_1^2 + 1.2x_2^2 - 2x_1x_2 + \sqrt{1 + x_1^2 + x_2^2} - 7x_1 - 3x_2,$

   wobei man den Startwert $x_0 := (0, 0)^T$ nehme[23].

2. Mit dem durch die Wolfe- bzw. die Armijo-Schrittweite gedämpften Newton-Verfahren löse man die unrestringierte

   (P) Minimiere $f(x), \quad x \in \mathbb{R}^4,$

   wobei

   $$\begin{aligned} f(x) \;:=\; & 100(x_1^2 - x_2)^2 + (1 - x_1)^2 + 90(x_3^2 - x_4)^2 + (1 - x_3)^2 \\ & + 10.1[(1 - x_2)^2 + (1 - x_4)^2] + 19.8(1 - x_2)(1 - x_4) \end{aligned}$$

   (die sogenannte Wood-Funktion). Man nehme $x_0 = (-1.5, -1, -3, -1)^T$ und $x_0 = (-3.1, 8.2, 5.5, -3.5)^T$ als Startwerte.

3. Seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ und die symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n\times n}$ gegeben. Sei $B_\phi \in \mathbb{R}^{n\times n}$ der Broyden-Update zum Parameter $\phi \in \mathbb{R}$. Man berechne die Eigenwerte von $B_c^{-1/2} B_\phi B_c^{-1/2}$. Hierbei benutze man die Abkürzungen

   $$a := y^T B_c^{-1} y, \qquad b := y^T s, \qquad c := s^T B_c s.$$

[23] Siehe P. Spellucci (1993, S. 117).

4. Seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ und eine symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n\times n}$ gegeben. Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n \times n$-Matrizen und $\mathcal{S}^{n\times n}_+ \subset \mathcal{S}^{n\times n}$ die konvexe Teilmenge der positiv definiten Matrizen. Man zeige, dass der BFGS-Update

$$B_+ := B_c - \frac{(B_c s)(B_c s)^T}{s^T B_c s} + \frac{yy^T}{y^T s}$$

Lösung der Aufgabe

$$\text{(P)} \quad \begin{cases} \text{Minimiere} \quad \phi(B) := \operatorname{tr}(B_c^{-1/2} B B_c^{-1/2}) - \ln\det(B_c^{-1/2} B B_c^{-1/2}) \quad \text{auf} \\ M := \{B \in \mathcal{S}^{n\times n} : B \in \mathcal{S}^{n\times n}_+,\ Bs = y\} \end{cases}$$

ist.

Hinweis: Diese Aussage findet man bei R. Fletcher (1991)[24]. Man gebe einen alternativen Beweis an, der die Gateaux-Variation und die Konvexität von ϕ benutzt.

5. Seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ und eine symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n\times n}$ gegeben. Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n \times n$-Matrizen und $\mathcal{S}^{n\times n}_+ \subset \mathcal{S}^{n\times n}$ die konvexe Teilmenge der positiv definiten Matrizen, ferner bezeichne $\|\cdot\|_F$ die Frobenius-Norm. Sei schließlich $V \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit mit $Vs = y$ und $H_c := B_c^{-1}$. Man zeige, dass die Inverse H_+ des BFGS-Updates B_+, also

$$H_+ := H_c + \left(1 + \frac{y^T H_c y}{y^T s}\right)\frac{ss^T}{y^T s} - \frac{s(H_c y)^T + (H_c y)s^T}{y^T s},$$

die Lösung der Aufgabe

$$\text{(P)} \quad \begin{cases} \text{Minimiere} \quad \phi(H) := \dfrac{1}{2}\|V^{1/2}(H - H_c)V^{1/2}\|_F^2 \quad \text{auf} \\ M := \{H \in \mathcal{S}^{n\times n} : H \in \mathcal{S}^{n\times n}_+,\ Hy = s\} \end{cases}$$

ist.

6. Seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ und eine symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n\times n}$ gegeben. Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n \times n$-Matrizen und $\mathcal{S}^{n\times n}_+ \subset \mathcal{S}^{n\times n}$ die konvexe Teilmenge der positiv definiten Matrizen, ferner bezeichne $\|\cdot\|_F$ die Frobenius-Norm. Man zeige, dass

$$B_+ := B_c - \frac{(B_c s - y)^T s}{(s^T B_c s)^2}(B_c s)(B_c s)^T$$

eine Lösung der Aufgabe

$$\text{(P)} \quad \begin{cases} \text{Minimiere} \quad \phi(B) := \dfrac{1}{2}\|B_c^{-1/2}(B - B_c)B_c^{-1/2}\|_F^2 \quad \text{auf} \\ M := \{B \in \mathcal{S}^{n\times n} : B \in \mathcal{S}^{n\times n}_+,\ s^T Bs = y^T s\} \end{cases}$$

ist.

[24] R. Fletcher (1991) "A new variational result for quasi-Newton formulae." SIAM J. Opt. 1, 18–21.

7. Seien $y, s \in \mathbb{R}^n$ und eine symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n\times n}$ gegeben. Es sei $(y - B_c s)^T s \neq 0$. Man bestimme $\gamma \in \mathbb{R}$ und $u \in \mathbb{R}^n$ so, dass die Matrix $B_+ = B_c + \gamma uu^T$ der Quasi-Newton-Gleichung $B_+ s = y$ genügt. Unter welchen Voraussetzungen ist die so bestimmte Matrix positiv definit?

8. Seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ und die symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n\times n}$ gegeben. Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n \times n$-Matrizen und $\mathcal{S}_+^{n\times n} \subset \mathcal{S}^{n\times n}$ die konvexe Teilmenge der positiv definiten Matrizen. Man zeige[25], dass der BFGS-Update der skalierten Matrix

$$\hat{B}_c := \frac{y^T B_c^{-1} y}{y^T s} B_c,$$

also

$$B_+ := \hat{B}_c - \frac{(\hat{B}_c s)(\hat{B}_c s)^T}{s^T \hat{B}_c s} + \frac{yy^T}{y^T s},$$

Lösung der Aufgabe

$$\text{(P)} \qquad \left\{ \begin{array}{l} \text{Minimiere} \quad \phi(B) := \dfrac{\operatorname{tr}(B_c^{-1/2} B B_c^{-1/2})/n}{\det(B_c^{-1/2} B B_c^{-1/2})^{1/n}} \quad \text{auf} \\ \qquad\qquad M := \{B \in \mathcal{S}^{n\times n} : B \in \mathcal{S}_+^{n\times n},\ Bs = y\} \end{array} \right.$$

ist.

Hinweis: Zunächst zeige man, dass $\phi(\cdot)$ in jedem $B \in \mathcal{S}_+^{n\times n}$ in jede Richtung $P \in \mathcal{S}^{n\times n}$ richtungsdifferenzierbar ist und die Gateaux-Variation $\phi'(B;\cdot)\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ gegeben ist durch (siehe auch Aufgabe 7 in Abschnitt 2.2)

$$\phi'(B;P) = \frac{1}{n \det(B_c^{-1} B)^{1/n}} \left(\operatorname{tr}(B_c^{-1} P) - \frac{\operatorname{tr}(B_c^{-1} B)}{n} \operatorname{tr}(B^{-1} P) \right).$$

Anschließend überlege man sich, dass für beliebige $A, B \in \mathcal{S}_+^{n\times n}$ die Implikation

$$\phi'(B; A - B) \geq 0 \Longrightarrow \phi(B) \leq \phi(A)$$

gilt. Im letzten Schritt zeige man, dass $B_+ \in M$ und $\phi'(B_+; B - B_+) \geq 0$ für alle $B \in M$.

9. Man wende das BFGS-Verfahren mit Wolfe-Schrittweite auf die Minimierung der Funktion

$$\begin{aligned} f(x) \quad := \quad & 100(x_1^2 - x_2)^2 + (1 - x_1)^2 + 90(x_3^2 - x_4)^2 + (1 - x_3)^2 \\ & + 10.1[(1 - x_2)^2 + (1 - x_4)^2] + 19.8(1 - x_2)(1 - x_4) \end{aligned}$$

an. Seien $x_0 = (-1.5, -1, -3, -1)^T$ und $x_0 = (-3.1, 8.2, 5.5, -3.5)^T$ die Startwerte.

10. Auf die unrestringierte Optimierungsaufgabe aus Aufgabe 9 wende man das L-BFGS-Verfahren mit $m = 1, 2, 3, 4$ an, wobei man den Startwert $x_0 = (-1.5, -1, -3, -1)^T$ nehme.

[25] Siehe

J. E. Dennis, H. Wolkowicz (1993) "Sizing and least-change secant algorithm." SIAM J. Numer. Anal. 30, 1291–1314.

3.3 Verfahren der konjugierten Gradienten

Die in Abschnitt 3.2 untersuchten Quasi-Newton-Verfahren haben den Nachteil, dass sie Speicherplatz für eine Approximation $B_c \in \mathbb{R}^{n\times n}$ der Hesseschen $\nabla^2 f(x_c)$ der Zielfunktion f in der aktuellen Näherung x_c benötigen. Dies kann für großes n ein Problem werden. Ein Ausweg bietet das in Unterabschnitt 3.2.6 angegebene L-BFGS-Verfahren. Noch weniger Speicherplatz benötigen i. Allg. die jetzt zu besprechenden Verfahren der konjugierten Gradienten. Wie von J. Nocedal, S. J. Wright (1999, S. 101) zu Beginn ihres Kapitels über *Conjugate Gradient Methods* ausgeführt wird, besteht ein zweifaches Interesse an Verfahren der konjugierten Gradienten. Einmal gehören diese Verfahren zu den nützlichsten Techniken zur Lösung großer linearer Gleichungssysteme mit einer symmetrischen, positiv definiten Koeffizientenmatrix bzw. der unrestringierten Minimierung einer quadratischen, gleichmäßig konvexen Zielfunktion. Zum anderen können die Verfahren zur Lösung allgemeiner unrestringierter Optimierungsaufgaben adaptiert werden. Wesentlich ist hierbei, dass keine Matrix gespeichert werden muss und die Verfahren (es gibt etliche Varianten) schneller als das Gradientenverfahren bzw. das Verfahren des steilsten Abstiegs ist. Hingewiesen sei aber darauf, dass die Wahl einer richtigen Schrittweitenstrategie nicht ganz einfach ist.

3.3.1 Quadratische Zielfunktionen

Wir betrachten die Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \frac{1}{2}x^T Ax - b^T x, \quad x \in \mathbb{R}^n,$$

wobei $A \in \mathbb{R}^{n\times n}$ stets als symmetrisch und positiv definit vorausgesetzt wird und $b \in \mathbb{R}^n$. Die Aufgabe (P) besitzt genau eine Lösung x^*, nämlich die Lösung des linearen Gleichungssystems $Ax = b$. Daher können die Verfahren dieses Unterabschnittes auch zur Lösung linearer Gleichungssysteme mit symmetrischer, positiv definiter Koeffizientenmatrix eingesetzt werden.

Wichtig ist die Definition

Definition 3.1 Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Vektoren $p_0, \ldots, p_k \in \mathbb{R}^n$, $k < n$, heißen konjugiert bezüglich A oder auch *A-konjugiert*, wenn sie vom Nullvektor verschieden sind und $p_i^T A p_j = 0$ für $0 \le i < j \le k$ gilt.

Ist ein System $\{p_0, \ldots, p_{n-1}\} \subset \mathbb{R}^n$ von A-konjugierten Richtungen bekannt (diese sind notwendig linear unabhängig), so kann das *Verfahren der konjugierten Richtungen* angewandt werden, siehe Aufgabe 2. In dem folgenden Satz wird das auf H. M. Hestenes, E. Stiefel (1952)[26] zurückgehende *Verfahren der konjugierten Gradienten* angegeben, in dem im Verfahren selber konjugierte Richtungen erzeugt werden.

Satz 3.2 *Zur Lösung von (P) betrachte man das folgende Verfahren:*

[26] H. M. Hestenes, E. Stiefel (1952) "Methods of conjugate gradients for solving linear systems." J. Res. Bur. Standards 48, 409–436.

- *Wähle $x_0 \in \mathbb{R}^n$, berechne $g_0 := Ax_0 - b$ und setze $p_0 := -g_0$.*
- *Für $k = 0, 1, \ldots$:*
  - *Falls $g_k = 0$, dann: $m := k$, STOP. x_m ist die Lösung von (P).*
  - *Andernfalls:*
    - *Berechne*

$$t_k := -\frac{g_k^T p_k}{p_k^T A p_k}, \qquad x_{k+1} := x_k + t_k p_k, \qquad g_{k+1} := g_k + t_k A p_k.$$

    - *Berechne*

$$\beta_k := \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2}, \qquad p_{k+1} := -g_{k+1} + \beta_k p_k.$$

Dann gilt:

1. *Das Verfahren bricht nach $m \leq n$ Schritten ab.*
2. *Es ist $p_i^T g_k = 0$ für $0 \leq i < k \leq m$.*
3. *Es ist $g_k^T p_k = -\|g_k\|_2^2$ für $0 \leq k \leq m$.*
4. *Es ist $g_i^T g_k = 0$ für $0 \leq i < k \leq m$.*
5. *Die Richtungen $p_0, \ldots, p_{m-1}$ sind A-konjugiert.*
6. *Es ist $\operatorname{span}\{p_0, \ldots, p_k\} = \operatorname{span}\{g_0, \ldots, g_k\} = \operatorname{span}\{g_0, Ag_0, \ldots, A^k g_0\}$ für $0 \leq k < m$.*

Beweis: Wir zeigen durch vollständige Induktion nach k: Sind $g_0, \ldots, g_k \neq 0$, wird das Verfahren also bis zum k-ten Schritt nicht abgebrochen, so gilt

(a) $p_i^T g_k = 0$ für $0 \leq i < k$,

(b) $g_k^T p_k = -\|g_k\|_2^2$,

(c) $g_i^T g_k = 0$ für $0 \leq i < k$,

(d) $p_0, \ldots, p_k$ sind A-konjugiert,

(e) $\operatorname{span}\{p_0, \ldots, p_k\} = \operatorname{span}\{g_0, \ldots, g_k\} = \operatorname{span}\{g_0, Ag_0, \ldots, A^k g_0\}$.

Diese fünf Aussagen sind für $k = 0$ (beachte: $p_0 := -g_0$) trivialerweise richtig. Für den Induktionsschluss nehmen wir an, $g_0, \ldots, g_{k+1}$ seien vom Nullvektor verschieden.

Für $0 \leq i < k$ ist

$$p_i^T g_{k+1} = p_i^T (g_k + t_k A p_k) = 0,$$

wobei die Induktionsvoraussetzungen (a) und (d) benutzt wurden. Ferner ist $p_k^T g_{k+1} = 0$ nach Definition der (exakten) Schrittweite t_k. Damit ist der Induktionsschluss für (a) vollzogen.

Es ist

$$g_{k+1}^T p_{k+1} = g_{k+1}^T(-g_{k+1} + \beta_k p_k) = -\|g_{k+1}\|_2^2$$

wegen des gerade eben für $k+1$ bewiesenen Teils (a). Damit ist auch (b) für $k+1$ bewiesen.

Für $1 \le i < k$ ist

$$g_i^T g_{k+1} = (\beta_{i-1} p_{i-1} - p_i)^T g_{k+1} = 0,$$

wobei der für $k+1$ schon bewiesene Teil (a) benutzt wurde, aus welchem auch

$$g_0^T g_{k+1} = -p_0^T g_{k+1} = 0$$

folgt. Berücksichtigt man nun noch, dass

$$g_k^T A p_k = (\beta_{k-1} p_{k-1} - p_k)^T A p_k = -p_k^T A p_k,$$

so erhält man

$$g_k^T g_{k+1} = g_k^T(g_k + t_k A p_k) = \|g_k\|_2^2 - \frac{g_k^T p_k}{p_k^T A p_k} g_k^T A p_k = \|g_k\|_2^2 - \frac{\|g_k\|_2^2}{p_k^T A p_k} p_k^T A p_k = 0,$$

wobei auch noch die Induktionsvoraussetzung (b) benutzt wurde. Insgesamt ist auch (c) für $k+1$ bewiesen.

Wegen der schon für $k+1$ bewiesenen Aussage (b) ist mit g_{k+1} auch p_{k+1} vom Nullvektor verschieden. Für $0 \le i < k$ ist

$$p_i^T A p_{k+1} = p_i^T A(-g_{k+1} + \beta_k p_k) = -g_{k+1}^T A p_i = \frac{1}{t_i} g_{k+1}^T (g_i - g_{i+1}) = 0.$$

Da schließlich

$$p_{k+1}^T A p_k = (-g_{k+1} + \beta_k p_k)^T \frac{1}{t_k}(g_{k+1} - g_k) = \frac{1}{t_k}(-\|g_{k+1}\|_2^2 + \beta_k \|g_k\|_2^2) = 0,$$

ist auch (d) für $k+1$ richtig.

Wegen $p_{k+1} = -g_{k+1} + \beta_k p_k$ und der Induktionsvoraussetzung (e) folgt sofort, dass $\operatorname{span}\{p_0, \ldots, p_{k+1}\} = \operatorname{span}\{g_0, \ldots, g_{k+1}\}$. Weiter ist nach Induktionsannahme

$$g_k \in \operatorname{span}\{g_0, Ag_0, \ldots, A^k g_0\}, \qquad p_k \in \operatorname{span}\{g_0, Ag_0, \ldots, A^k g_0\}.$$

Daher ist

$$\begin{aligned} g_{k+1} &= g_k + t_k A p_k \\ &\in \operatorname{span}\{g_0, Ag_0, \ldots, A^k g_0\} + A(\operatorname{span}\{g_0, Ag_0, \ldots, A^k g_0\}) \\ &\subset \operatorname{span}\{g_0, Ag_0, \ldots, A^{k+1} g_0\} \end{aligned}$$

und folglich

$$\operatorname{span}\{g_0, g_1, \ldots, g_{k+1}\} \subset \operatorname{span}\{g_0, Ag_0, \ldots, A^{k+1} g_0\}.$$

Um die umgekehrte Inklusion zu beweisen, benutzen wir, dass nach Induktionsvoraussetzung

$$A^{k+1}g_0 = A(A^k g_0) \in A(\text{span}\,\{p_0, \ldots, p_k\}) \subset \text{span}\,\{Ap_0, Ap_1, \ldots, Ap_k\}.$$

Wegen $Ap_i = (g_{i+1} - g_i)/t_i$, $i = 0, \ldots, k$, ist

$$A^{k+1}g_0 \in \text{span}\,\{g_0, g_1, \ldots, g_{k+1}\}.$$

Zusammen mit der Induktionsvoraussetzung für (e) erhalten wir

$$\text{span}\,\{g_0, Ag_0, \ldots, A^{k+1}g_0\} \subset \text{span}\,\{g_0, g_1, \ldots, g_{k+1}\}.$$

Damit ist der Induktionsbeweis abgeschlossen.

Insbesondere ist bewiesen worden, dass das Verfahren A-konjugierte Richtungen erzeugt, solange es nicht abbricht. Da A-konjugierte Richtungen linear unabhängig sind, kann es von ihnen nicht mehr als n geben. Der Satz ist daher vollständig bewiesen. □
□

Bemerkungen: Die exakte Schrittweite t_k kann im Verfahren der konjugierten Gradienten auch durch $t_k := \|g_k\|_2^2 / p_k^T A p_k$ (siehe die dritte Behauptung in Satz 3.2) berechnet werden.

Ein Vorteil des Verfahrens der konjugierten Gradienten zur Lösung der unrestringierten Optimierungsaufgabe (P) (mit der quadratischen, gleichmäßig konvexen Zielfunktion $f(x) := \frac{1}{2}x^T Ax - b^T x$) bzw. des äquivalenten linearen Gleichungssystems $Ax = b$ (mit der symmetrischen, positiv definiten Koeffizientenmatrix A) gegenüber Eliminationsverfahren besteht darin, dass die Matrix A in jedem Iterationsschritt nur dadurch eingreift, dass ihre *Wirkung* auf die aktuelle Richtung p_k zu bestimmen ist. Genau das macht das Verfahren attraktiv für hochdimensionale, dünn besetzte Aufgaben vom angegebenen Typ. □

Wendet man das Gradientenverfahren mit exakter Schrittweite auf die Optimierungsaufgabe (P) mit der Zielfunktion $f(x) := \frac{1}{2}x^T Ax - b^T x$ an, wobei A eine symmetrische, positiv definite Matrix mit kleinstem (positiven) Eigenwert $\lambda_{\min}$ und größtem Eigenwert $\lambda_{\max}$ ist, so konnte in Aufgabe 10b in Abschnitt 3.1 mit Hilfe der Ungleichung von Kantorowitsch gezeigt werden, dass

$$f(x_{k+1}) - f(x^*) \le \Big(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\Big)^2 [f(x_k) - f(x^*)], \qquad k = 0, 1, \ldots,$$

wobei natürlich $x^* = A^{-1}b$ die Lösung von (P) bedeutet. Berücksichtigt man, dass die Kondition von A bezüglich der Spektralnorm durch $\kappa_2(A) = \lambda_{\max}/\lambda_{\min}$ gegeben ist, und führt man die (elliptische) Norm $\|\cdot\|_A$ durch $\|x\|_A := \sqrt{x^T Ax}$ auf dem $\mathbb{R}^n$ ein, so erkennt man, dass dieses Ergebnis in

$$\|x_{k+1} - x^*\|_A \le \Big(\frac{\kappa_2(A) - 1}{\kappa_2(A) + 1}\Big) \|x_k - x^*\|_A, \qquad k = 0, 1, \ldots,$$

übergeht. Hieraus wiederum folgt

$$\frac{\|x_k - x^*\|}{\|x_0 - x^*\|} \le \Big(\frac{\kappa_2(A) - 1}{\kappa_2(A) + 1}\Big)^k, \qquad k = 0, 1, \ldots.$$

Hierdurch wird die Aussage "Je kleiner die Kondition von A, desto besser ist die Konvergenz des Gradientenverfahrens" quantifiziert.

Uns kommt es nun darauf an, eine entsprechende Aussage auch für das in Satz 3.2 angegebene Verfahren der konjugierten Gradienten (fast vollständig) zu beweisen.

Satz 3.3 *Seien $x_0, \ldots, x_m$ durch das Verfahren der konjugierten Gradienten gewonnen und x^* die Lösung von $Ax = b$. Dann ist*

$$\frac{\|x_k - x^*\|_A}{\|x_0 - x^*\|_A} \leq 2\Big(\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1}\Big)^k, \qquad k = 0, \ldots, m.$$

Hierbei ist die Norm $\|\cdot\|_A$ durch $\|x\|_A := \sqrt{x^T Ax}$ auf dem $\mathbb{R}^n$ definiert.

Beweis: Wir definieren zunächst für $k = 0, \ldots, m$ den sogenannten *Krylov-Raum*

$$\mathcal{K}_k(A, g_0) := \text{span}\,\{g_0, Ag_0, \ldots, A^{k-1}g_0\},$$

der auch schon in der letzten Ausage von Satz 3.2 vorkam. Nun beweisen wir:

- Für $k = 0, \ldots, m$ gilt:

  1. Die k-te Iterierte x_k ist die Lösung der Aufgabe

     $$(\mathrm{P}_k) \qquad \text{Minimiere} \quad \|x - x^*\|_A, \quad x \in x_0 + \mathcal{K}_k(A, g_0).$$

  2. Es ist

     $$\|x_k - x^*\|_A = \min_{p \in \Pi_k, p(0)=1} \|p(A)(x_0 - x^*)\|_A.$$

  3. Ist $\lambda_{\min}$ der kleinste, $\lambda_{\max}$ der größte Eigenwert der symmetrischen, positiv definiten Matrix A und ist $p \in \Pi_k$ ein beliebiges Polynom mit $p(0) = 1$, so ist

     $$\|x_k - x^*\|_A \leq \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} |p(\lambda)|\, \|x_0 - x^*\|_A.$$

Denn: Zunächst ist $x_k \in x_0 + \mathcal{K}_k(A, g_0)$, wie man durch vollständige Induktion unter Benutzung des letzten Teils von Satz 3.2 nachweist. Bei (P_k) handelt es sich um die Aufgabe, die Lösung x^* bezüglich der durch dass innere Produkt $(x, y)_A := x^T Ay$ induzierten Norm $\|\cdot\|_A$ auf den affinen Teilraum $x_0 + \mathcal{K}_k(A, g_0)$ zu projizieren. Daher ist x_k genau dann diese Projektion bzw. die Lösung von (P_k), wenn $(x_k - x^*, z)_A = 0$ für alle $z \in \mathcal{K}_k(A, g_0)$. Wiederum wegen des letzten Teils von Satz 3.2 ist $\mathcal{K}_k(A, g_0) = \text{span}\,\{g_0, \ldots, g_{k-1}\}$. Folglich ist

$$(x_k - x^*, z)_A = (x_k - x^*)^T Az = (Ax_k - b)^T z = g_k^T z = 0,$$

da g_k wegen des vierten Teiles von Satz 3.2 auf $g_0, \ldots, g_{k-1}$ senkrecht steht. Damit ist der erste Teil der Aussage bewiesen. Die zweite Aussage ist nur eine Umformulierung

der ersten. Ein Element $x \in x_0 + \mathcal{K}_k(A, g_0)$ besitzt eine Darstellung $x = x_0 + q(A)g_0$ mit einem Polynom $q \in \Pi_{k-1}$. Dann ist

$$\begin{aligned} x - x^* &= x_0 + q(A)g_0 - x^* \\ &= x_0 + q(A)(Ax_0 - b) - x^* \\ &= x_0 + q(A)A(x_0 - x^*) - x^* \\ &= (I + q(A)A)(x_0 - x^*) \\ &= p(A)(x_0 - x^*), \end{aligned}$$

wobei $p(t) := 1 + q(t)t$ ein Polynom vom Grad $\leq k$ mit $p(0) = 1$ ist. Aus dem ersten Teil der obigen Aussage folgt dann sofort der zweite. Seien λ_i, $i = 1, \ldots, n$, die Eigenwerte der symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$, mit $\lambda_{\min}$ bezeichnen wir den kleinsten, mit $\lambda_{\max}$ den größten Eigenwert von A. Ferner sei $\{u_1, \ldots, u_n\}$ ein vollständiges Orthonormalsystem von Eigenvektoren und

$$x_0 - x^* = \sum_{i=1}^{n} \alpha_i u_i.$$

Für ein beliebiges Polynom $p \in \Pi_k$ ist dann

$$\begin{aligned} \|p(A)(x_0 - x^*)\|_A^2 &= \sum_{i=1}^{n} \lambda_i p(\lambda_i)^2 \alpha_i^2 \\ &\leq \max_{i=1,\ldots,n} p(\lambda_i)^2 \sum_{i=1,\ldots,n} \lambda_i \alpha_i^2 \\ &= \max_{i=1,\ldots,n} p(\lambda_i)^2 \, \|x_0 - x^*\|_A^2 \\ &\leq \max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} p(\lambda)^2 \, \|x_0 - x^*\|_A^2. \end{aligned}$$

Daher impliziert die zweite Aussage auch die dritte der obigen Aussagen.

Ist $\lambda_{\min} = \lambda_{\max}$, so ist nach dem ersten Schritt die Lösung erreicht und daher die Behauptung trivialerweise richtig. O. B. d. A. können wir daher im weiteren $\lambda_{\min} < \lambda_{\max}$ annehmen. In der dritten der obigen Aussagen setze man als das Polynom p vom Grad $\leq k$ mit $p(0) = 1$ speziell

$$p(\lambda) := T_k\Big(\frac{\lambda_{\min} + \lambda_{\max} - 2\lambda}{\lambda_{\max} - \lambda_{\min}}\Big) \Big/ T_k\Big(\frac{\lambda_{\min} + \lambda_{\max}}{\lambda_{\max} - \lambda_{\min}}\Big).$$

Hierbei bedeutet $T_k(\cdot)$ das k-te Tschebyscheffsche Polynom erster Art, welches für $|t| \leq 1$ die Darstellung

$$T_k(t) = \cos(k \arccos t)$$

und für $|t| \geq 1$ die Darstellung

$$T_k(t) = \frac{1}{2}\Big[\Big(t + \sqrt{t^2 - 1}\Big)^k + \Big(t + \sqrt{t^2 - 1}\Big)^{-k}\Big]$$

besitzt. Für $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ ist

$$\frac{\lambda_{\min} + \lambda_{\max} - 2\lambda}{\lambda_{\max} - \lambda_{\min}} \in [-1, 1], \qquad \frac{\lambda_{\min} + \lambda_{\max}}{\lambda_{\max} - \lambda_{\min}} > 1$$

und daher

$$\begin{aligned} \frac{\|x_k - x^*\|_A}{\|x_0 - x^*\|_A} &\leq \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} |p(\lambda)| \\ &\leq \frac{1}{|T_k((\lambda_{\min} + \lambda_{\max})/(\lambda_{\max} - \lambda_{\min}))|} \\ &= \frac{1}{T_k((\kappa_2(A) + 1)/(\kappa_2(A) - 1))}. \end{aligned}$$

Zur Abkürzung setze man $\eta := 1/(\kappa_2(A) - 1)$. Dann ist

$$\begin{aligned} T_k((\kappa_2(A) + 1)/(\kappa_2(A) - 1)) &= T_k(1 + 2\eta) \\ &\geq \frac{1}{2}[1 + 2\eta + \sqrt{(1 + 2\eta)^2 - 1}]^k \\ &= \frac{1}{2}[1 + 2\eta + 2\sqrt{\eta(\eta + 1)}]^k \\ &= \frac{1}{2}[\sqrt{\eta} + \sqrt{\eta + 1}]^{2k} \\ &= \frac{1}{2}\Big(\frac{\sqrt{\kappa_2(A)} + 1}{\sqrt{\kappa_2(A)} - 1}\Big)^k, \end{aligned}$$

wobei man die letzte Gleichung durch Einsetzen und einfache Manipulationen verifiziert. Daher ist der Satz schließlich bewiesen. □
□

Die Abschätzung in Satz 3.3 für das Verfahren der konjugierten Gradienten ist besser als die entsprechende Abschätzung für das Gradientenverfahren, qualitativ sagt sie wieder aus, dass "gute" Konvergenz bei kleiner Kondition von A zu erwarten ist. Durch eine sogenannte *Präkonditionierung* kann man versuchen, ein äquivalentes Problem zu lösen, bei dem die Hessesche der Zielfunktion (bzw. die Koeffizientenmatrix des linearen Gleichungssystems) immer noch symmetrisch und positiv definit ist, aber eine kleinere Kondition besitzt. Die Idee hierzu besteht darin, mit einer symmetrischen, positiv definiten Matrix $M \in \mathbb{R}^{n \times n}$ zu dem transformierten unrestringierten Optimierungsproblem

$$(\mathrm{P}_M) \qquad \text{Minimiere} \quad \phi(y) := \frac{1}{2} y^T M^{-1/2} A M^{-1/2} y - (M^{-1/2} b)^T y, \quad y \in \mathbb{R}^n,$$

bzw. dem transformierten linearen Gleichungssystem

$$M^{-1/2} A M^{-1/2} y = M^{-1/2} b$$

überzugehen. Ist x^* die Lösung von (P) und y^* die Lösung von (P_M), so ist $x^* = M^{-1/2} y^*$. Nun wende man auf (P_M) das Verfahren der konjugierten Gradienten an, wobei wir bei der Berechnung der Schrittweite t_k eine Bemerkung im Anschluss an den Beweis von Satz 3.2 benutzen. Wir erhalten:

- Wähle $y_0 \in \mathbb{R}^n$, berechne $h_0 := M^{-1/2}(AM^{-1/2}y_0 - b)$ und setze $q_0 := -h_0$.
- Für $k = 0, 1, \ldots$:
  - Falls $h_k = 0$, dann: $m := k$, STOP. y_m ist die Lösung von (P_M) und daher $x_m := M^{-1/2}y_m$ die Lösung von (P).
  - Andernfalls:
    - Berechne

$$t_k := \frac{\|h_k\|_2^2}{q_k^T M^{-1/2} A M^{-1/2} q_k}, \qquad y_{k+1} := y_k + t_k q_k$$

      sowie

$$h_{k+1} := h_k + t_k M^{-1/2} A M^{-1/2} q_k.$$

    - Berechne

$$\beta_k := \frac{\|h_{k+1}\|_2^2}{\|h_k\|_2^2}, \qquad q_{k+1} := -h_{k+1} + \beta_k q_k.$$

In dieser Form ist der Algorithmus natürlich nicht brauchbar, da man nicht bereit ist, $M^{-1/2}$ zu berechnen. Setzt man aber $x_k := M^{-1/2}y_k$, $g_k := M^{1/2}h_k$ und $p_k := M^{-1/2}q_k$, so erhält man das *Verfahren der konjugierten Gradienten mit Präkonditionierung*:

- Wähle $x_0 \in \mathbb{R}^n$, berechne $g_0 := Ax_0 - b$ und $p_0 := -M^{-1}g_0$.
- Für $k = 0, 1, \ldots$:
  - Falls $g_k = 0$, dann: $m = k$, STOP. x_m ist die Lösung von (P).
  - Andernfalls:
    - Berechne Ap_k und anschließend

$$t_k = \frac{g_k^T M^{-1} g_k}{p_k^T A p_k}, \qquad x_{k+1} := x_k + t_k p_k, \qquad g_{k+1} := g_k + t_k A p_k.$$

    - Berechne $M^{-1}g_{k+1}$ und anschließend

$$\beta_k := \frac{g_{k+1}^T M^{-1} g_{k+1}}{g_k^T M^{-1} g_k}, \qquad p_{k+1} := -M^{-1}g_{k+1} + \beta_k p_k.$$

Der Unterschied in der Komplexität zum Verfahren der konjugierten Gradienten ohne Präkonditionierung besteht darin, dass hier in jedem Schritt $M^{-1}g_k$ zu berechnen, also ein lineares Gleichungssystem mit M als Koeffizientenmatrix zu lösen ist. Die Anforderungen an den Präkonditionierer M sind also, dass M symmetrisch und positiv definit ist, ein lineares Gleichungssystem mit M als Koeffizientenmatrix "einfach" zu lösen ist und die Kondition von $M^{-1/2}AM^{-1/2}$ bzw. von $M^{-1}A$ "möglichst klein" sein sollte, also M "möglichst nahe" bei A sein sollte. Auf allgemeine oder spezielle (d. h. vom Problem abhängende) Präkonditionierer wollen wir hier nicht eingehen. Stattdessen geben wir eine Matlab-Funktion zum obigen Verfahren der konjugierten Gradienten mit Präkonditionierung an. Wir hätten diese wesentlich flexibler mit einer variablen Zahl von

In- und Output-Parameterm gestalten können, verzichten aber darauf. Als Parameter geben wir nicht den Präkonditioniere M selber an, sondern den Cholesky-Faktor L, wobei also $M = LL^T$ mit einer unteren Dreiecksmatrix L, die positive Diagonalelemente besitzt.

```
function [x,error,iter] = ConGrad(A,b,x,L,max_iter,tol)
% This function solves the symmetric positive definite linear system Ax=b
% using the Conjugate Gradient method with preconditioning.
% Input   A        symmetric positive definite matrix
%         b        right hand side vector
%         x        initial guess vector
%         L        Cholesky-factor of preconditioner matrix
%         max_iter maximum number of iterations
%         tol      error tolerance
%
% Output  x        solution vector
%         error    error norm
%         iter     number of iterations performed
%**********************************************************************
iter=0; [n,n]=size(A);
bnrm2=norm(b);
if bnrm2==0
    x=zeros(n,1); error=0; return
end;
g=A*x-b; z=L'\(L\g); p=-z; rho_k=g'*z;
error=norm(g)/bnrm2;
while (error>tol)&(iter<max_iter)
    q=A*p; t=rho_k/(p'*q); x=x+t*p; g=g+t*q;
    z=L'\(L\g); rho_kp=g'*z; beta=rho_kp/rho_k;
    p=-z+beta*p; rho_k=rho_kp;
    error=norm(g)/bnrm2; iter=iter+1;
end;
```

Beispiel: Gegeben sei die Poisson-Gleichung $-\Delta u = 1$ im Einheitsquadrat $\Omega := \{(x, y) : 0 < x, y < 1\}$, gesucht ist die (eindeutig existierende) Lösung, die auf dem Rande $\partial\Omega$ von Ω verschwindet. Nach äquidistanter Diskretisierung mit der Maschenweite $h := 1/(N + 1)$ mit einer $N^2 \times N^2$-Koeffizientenmatrix

$$A := \begin{pmatrix} A_N & -I_N & \cdots & 0_N \\ -I_N & A_N & \ddots & \vdots \\ \vdots & \ddots & \ddots & -I_N \\ 0_N & \cdots & -I_N & A_N \end{pmatrix} \in \mathbb{R}^{N^2 \times N^2}$$

mit der $N \times N$-Nullmatrix 0_N, der $N \times N$-Einheitsmatrix I_N und der Tridiagonalmatrix

$$A_N := \begin{pmatrix} 4 & -1 & \cdots & 0 \\ -1 & 4 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ 0 & \cdots & -1 & 4 \end{pmatrix} \in \mathbb{R}^{N \times N}.$$

Ferner ist der Vektor b durch $b := (1/N^2)e_{N^2}$ gegeben, wobei e_{N^2} der Vektor des $\mathbb{R}^{N^2}$ ist, dessen Komponenten alle gleich 1 sind. Die Matrix A wird in einem function-file `Modell.m` berechnet. Es lautet

```
function A=Modell(N);
% A is the N^2-by-N^2 coefficient matrix corresponding to the
% model problem
%*****************************************************************
I=eye(N);
S=4*eye(N)+diag(-ones(N-1,1),1)+diag(-ones(N-1,1),-1);
A=zeros(N^2,N^2);
for i=1:N
   A((i-1)*N+1:i*N,(i-1)*N+1:i*N)=S;
end;
for i=1:N-1
   A((i-1)*N+1:i*N,i*N+1:(i+1)*N)=-I;
   A(i*N+1:(i+1)*N,(i-1)*N+1:i*N)=-I;
end;
```

Als Test benutzen wir das script-file `Test.m` mit dem Inhalt

```
N=100;
A=sparse(Modell(N));b=(1/N)^2*ones(N^2,1);
x0=zeros(N^2,1);
max_iter=1000;tol=0.0000001;
L=speye(N^2);
tic;
[x,error,iter]=ConGrad(A,b,x0,L,max_iter,tol);
toc;
```

Immerhin ist hier ein lineares Gleichungssystem mit 10 000 Gleichungen und Unbekannten zu lösen. Wir sind erfolgreich. Nach 170 Iterationen und `elapsed time=3.8021` ist `error=9.5582e-08`. Wie wir am obigen Programm sehen, haben wir keinen Präkonditionierer benutzt. Ersetzen wir dagegen die Zeile

```
L=speye(N^2);
```

durch

```
L=cholinc(A,1e-3)';
```

(dann wird eine sogenannte unvollständige Cholesky-Faktorisierung mit drop-tolerance 10^{-3} durchgeführt), so erhalten wir nach 15 Iterationen, `elapsed time=0.8985`, dass `error=3.0549e-08`. □

3.3.2 Das Fletcher-Reeves-Verfahren

Von R. Fletcher, C. M. Reeves (1964)[27] stammt eine erste Verallgemeinerung des Verfahrens der konjugierten Gradienten auf unrestringierte Optimierungsaufgaben

$$\text{(P)} \qquad \text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n,$$

[27] R. Fletcher, C. M. Reeves (1964) "Function minimization by conjugate gradients." Computer Journal 7, 149–154.

mit nicht notwendig quadratischer Zielfunktion f. Wir beschränken uns im wesentlichen auf die Beschreibung dieses Verfahrens, Varianten werden in den Aufgaben angesprochen. Wir erinnern an die Voraussetzungen (V) (a)–(c) aus Unterabschnitt 3.1.1 und die (gleichmäßigen) Konvexitätsvoraussetzungen (K) (a)–(c) aus Lemma 1.6. Im folgenden Satz wird das Fletcher-Reeves-Verfahren mit exakter Schrittweitenstrategie angegeben und unter den Voraussetzungen (V) (a)–(c) bzw. (K) (a)–(c) eine globale Konvergenzaussage bewiesen. Auf einen entsprechenden Konvergenzsatz mit einer inexakten Schrittweitenstrategie, nämlich der in Aufgabe 1 in Abschnitt 3.1 angegebenen strengen Wolfe-Schrittweite, werden wir in Aufgabe 6 eingehen, siehe auch M. Al-Baali (1985)[28].

Satz 3.4 *Gegeben sei die Optimierungsaufgabe (P). Die Zielfunktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ genüge den Voraussetzungen (V) (a)–(c). Das Verfahren der konjugierten Gradienten von Fletcher-Reeves ist durch den folgenden Algorithmus gegeben:*

- *Gegeben $x_0 \in \mathbb{R}^n$, berechne $g_0 := \nabla f(x_0)$ und setze $p_0 := -g_0$.*
- *Für $k = 0, 1, \ldots$:*
  - *Falls $g_k = 0$, dann: STOP. x_k ist stationäre Lösung von (P).*
  - *Andernfalls:*
    - *Berechne die exakte Schrittweite $t_k := t^*(x_k, p_k)$.*
    - *Berechne*

$$x_{k+1} := x_k + t_k p_k, \qquad g_{k+1} := \nabla f(x_{k+1})$$

sowie

$$\beta_k := \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2}, \qquad p_{k+1} := -g_{k+1} + \beta_k p_k.$$

Dann gilt: Bricht das Verfahren nicht nach endlich vielen Schritten mit einer stationären Lösung von (P) ab, so liefert es eine Folge $\{x_k\}$ mit $\liminf_{k\to\infty} \|g_k\|_2 = 0$, wenigstens ein Häufungspunkt von $\{x_k\}$ ist also eine stationäre Lösung von (P). Sind sogar die Konvexitätsvoraussetzungen (K) (a)–(c) erfüllt, so konvergiert die gesamte Folge $\{x_k\}$ gegen die dann eindeutige Lösung x^* von (P).

Beweis: Zunächst beachten wir: Da im Verfahren stets die exakte Schrittweite gewählt wird, ist $g_k^T p_k = -\|g_k\|_2^2 < 0$ für $g_k \neq 0$, also p_k eine Abstiegsrichtung in x_k. Das Verfahren breche nicht vorzeitig mit einer stationären Lösung ab. Im Widerspruch zur Behauptung nehmen wir an, es sei $\liminf_{k\to\infty} \|g_k\|_2 > 0$. Dann existiert ein $\epsilon > 0$ mit $\|g_k\|_2 \geq \epsilon$ für alle k. Wegen Satz 1.2 gibt es eine von k unabhängige Konstante $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \geq \theta\Big(\frac{g_k^T p_k}{\|p_k\|_2}\Big)^2 = \theta\,\frac{\|g_k\|_2^4}{\|p_k\|_2^2} = \frac{\theta}{\alpha_k} \qquad \text{mit} \quad \alpha_k := \frac{\|p_k\|_2^2}{\|g_k\|_2^4}.$$

[28] M. Al-Baali (1985) "Descent property and global convergence of the Fletcher-Reeves method with inexact line search." IMA J. Numer. Anal. 5, 121–124.

Für $k \geq 1$ ist

$$\alpha_k = \frac{\|p_k\|_2^2}{\|g_k\|_2^4} = \frac{\|g_k\|_2^2 + \beta_{k-1}^2 \|p_{k-1}\|_2^2}{\|g_k\|_2^4} = \frac{1}{\|g_k\|_2^2} + \alpha_{k-1}.$$

Durch Zurückspulen erhäkt man

$$\alpha_k = \sum_{j=1}^{k} \frac{1}{\|g_j\|_2^2} + \alpha_0 = \sum_{j=0}^{k} \frac{1}{\|g_j\|_2^2} \leq \frac{k+1}{\epsilon^2}, \qquad k = 0, 1, \ldots,$$

und hieraus

$$(*) \qquad f(x_k) - f(x_{k+1}) \geq \frac{\theta \epsilon^2}{k+1}, \qquad k = 0, 1, \ldots.$$

Die harmonische Reihe ist bekanntlich divergent. Daher folgt aus $(*)$, dass $\{f(x_k)\}$ nicht nach unten beschränkt ist, was einen Widerspruch zur vorausgesetzten Kompaktheit der Niveaumenge L_0 darstellt.

Nun seien sogar die Voraussetzungen (K) (a)–(c) erfüllt. Ein $\epsilon > 0$ sei vorgegeben. Wegen der unter den schwächeren Voraussetzungen (V) (a)–(c) bewiesenen Aussage existiert ein $k_0 \in \mathbb{N}$ mit $\|g_{k_0}\|_2 \leq c\epsilon$. Eine Anwendung von Lemma 1.6 liefert für alle $k \geq k_0$ die Ungleichungskette

$$\frac{c}{2}\|x_k - x^*\|_2^2 \leq f(x_k) - f(x^*) \leq f(x_{k_0}) - f(x^*) \leq \frac{1}{2c}\|g_{k_0}\|_2^2 \leq \frac{c\epsilon^2}{2}.$$

Daher ist $\|x_k - x^*\|_2 \leq \epsilon$ für alle $k \geq k_0$, so dass auch der zweite Teil des Satzes bewiesen ist. □ □

Bemerkungen: Spezialisiert man das Fletcher-Reeves-Verfahren auf eine quadratische Zielfunktion, so erhält man genau das in Unterabschnitt 3.3.1 untersuchte Verfahren der konjugierten Gradienten. Insbesondere bricht das Verfahren in diesem Fall nach $m \leq n$ Schritten ab.

Es gibt einige Variantem zum Fletcher-Reeves-Verfahren. Diese unterscheiden sich im wesentlichen in der Definition des Skalars β_k, reduzieren sich für eine quadratische Zielfunktion aber sts auf das Verfahren der konjugierten Gradienten aus 3.3.1. So setzen z. B. E. Polak, G. Ribière (1969)[29]

$$\beta_k := \frac{g_{k+1}^T (g_{k+1} - g_k)}{\|g_k\|_2^2}.$$

Konvergenzaussagen zum Polak-Ribière-Verfahren werden in Aufgabe 5 gemacht.

I. Allg. macht man in einem Verfahren der konjugierten Gradienten alle n Schritte einen sogenannten restart, indem man wieder mit der negativen Gradientenrichtung in der aktuellen Näherung beginnt. □

[29] E. Polak, G. Ribière (1969) "Note sur la convergence de méthodes de directions conjuguées." Rev. Fr. Inf. Rech. Oper. 16, 35–43.

Wir wollen eine einfache Matlab-Funktion angeben, durch die das Fletcher-Reeves-Verfahren implementiert wird. Als Schrittweitenfunktion benutzen wir hierbei die Armijo-Schrittweite, wodurch allerdings nicht garantiert ist, dass man Abstiegsrichtungen erhält. Besser wäre es, die strenge Wolfe-Schrittweite zu implementieren (Hinweise hierzu findet man bei C. Geiger, C. Kanzow (1999, S. 49 ff.)).

```
function [x_min,iter]=FleRee(fun,x_0,max_iter,tol);
%Input-Parameter:
%            fun        function to be minimized
%            x_0        starting value
%            max_iter   maximal number of iterations
%            tol        If norm(gradient)<=tol: exit
%Output-Parameter:
%            x_min      (local) solution
%            iter       number of iterations
%*******************************************************
%The Fletcher-Reeves conjugate gradient method with
%Armijo-line search and restart every n=length(x_0) steps
%is used.
%*******************************************************
x_c=x_0;
[f_c,g_c]=feval(fun,x_c);iter=0;n=length(x_0);
while (norm(g_c)>tol)&(iter<max_iter)
     if (mod(iter,n)==0)
          p=-g_c;
     end;
     t=Armijo(x_c,p,fun); x_plus=x_c+t*p;
     [f_plus,g_plus]=feval(fun,x_plus);
     beta=(norm(g_plus)/norm(g_c))^2;
     p=-g_plus+beta*p; g_c=g_plus; x_c=x_plus;
     iter=iter+1;
end;
if (norm(g_c)<=tol)
    x_min=x_c;
end;
```

Der Aufruf

```
[x,iter]=FleRee('Rosenbrock',[-1.2;1],500,1e-8);
```

liefert z. B.

$$x = \begin{pmatrix} 0.99999999988120 \\ 0.99999999976207 \end{pmatrix}, \qquad \text{iter} = 85.$$

3.3.3 Aufgaben

1. Mit dem Verfahren der konjugierten Gradienten von Hestenes-Stiefel löse man die Aufgabe, die Funktion

   $$f(x) := x_2^2 + 0.3x_1x_2 + 0.975x_2^2 + 0.01x_1x_3 + x_3^2 + 3x_1 - 4x_2 + x_3$$

   auf dem $\mathbb{R}^3$ zu minimieren[30].

[30] Siehe P. Spellucci (1993, S. 164).

2. Gegeben sei die quadratische Zielfunktion $f(x) := \frac{1}{2}x^T Ax - b^T x$ mit einer symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$. Seien $p_0, \ldots, p_{n-1} \in \mathbb{R}^n$ konjugiert bezüglich A. Man betrachte das folgende Verfahren zur Minimierung von $f(x)$ auf dem $\mathbb{R}^n$:

   - Wähle $x_0 \in \mathbb{R}^n$, berechne $g_0 := Ax_0 - b$.
   - Für $k = 0, 1, \ldots$:
     - Falls $g_k = 0$, dann: $m := k$, f nimmt in x_m das Minimum an. STOP.
     - Andernfalls berechne

$$t_k := -\frac{g_k^T p_k}{p_k^T A p_k}, \qquad x_{k+1} := x_k + t_k p_k, \qquad g_{k+1} := g_k + t_k A p_k.$$

   Durch vollständige Induktion nach k zeige man: Sind $g_0, \ldots, g_k \neq 0$, ist das Verfahren im k-ten Schritt also noch nicht abgebrochen, so ist x_{k+1} die Lösung der Aufgabe

$$(\mathrm{P}_k) \qquad \text{Minimiere} \quad f(x), \quad x \in x_0 + \operatorname{span}\{p_0, \ldots, p_k\}.$$

   Wegen $x_0 + \operatorname{span}\{p_0, \ldots, p_{n-1}\} = \mathbb{R}^n$ bricht das Verfahren also nach $m \leq n$ Schritten mit dem Minimum von f ab.

   Hinweis: Nach Konstruktion ist klar, dass $x_{k+1} \in x_0 + \operatorname{span}\{p_0, \ldots, p_k\}$. Man zeige, dass $g_{k+1}^T p_i = 0$, $i = 0, \ldots, k$, und überlege sich, dass dies die Behauptung impliziert.

3. Gegeben sei die quadratische Zielfunktion $f(x) := \frac{1}{2}x^T Ax - b^T x$ mit einer symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$. Zur Lösung der zugehörigen unrestringierten Optimierungsaufgabe bzw. des linearen Gleichungssystems $Ax = b$ betrachte man die folgende Modifikation des Verfahrens der konjugierten Gradienten, welche sich von diesem dadurch unterscheidet, dass nicht notwendig am Anfang $p_0 := -g_0$ gewählt wird.

   - Wähle $x_0 \in \mathbb{R}^n$, berechne $g_0 := Ax_0 - b$. O. B. d. A. sei $g_0 \neq 0$. Wähle $p_0 \in \mathbb{R}^n$ mit $g_0^T p_0 < 0$.
   - Für $k = 0, 1, \ldots$:
     - Berechne

$$t_k := -\frac{g_k^T p_k}{p_k^T A p_k}, \qquad x_{k+1} := x_k + t_k p_k, \qquad q_{k+1} := g_k + t_k A p_k.$$

     - Falls $g_{k+1} = 0$, dann: $m := k + 1$, STOP. x_m ist die Lösung.
     - Andernfalls:
       * Berechne

$$p_{k+1} := \begin{cases} -g_1 - \dfrac{g_1^T(g_1 - g_0)}{g_0^T p_0} p_0, & k = 0, \\ -g_{k+1} + \dfrac{g_{k+1}^T g_0}{g_0^T p_0} p_0 + \dfrac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2} p_k, & k = 1, 2, \ldots. \end{cases}$$

   Durch vollständige Induktion nach k zeige man: Ist $k \geq 1$ und sind $g_1, \ldots, g_k \neq 0$, so gilt:

   (a) Es ist $p_i^T g_k = 0$, $i = 0, \ldots, k-1$,

(b) Es ist $g_k^T p_k = -\|g_k\|_2^2$,

(c) Es ist $g_i^T g_k = 0$, $i = 1, \ldots, k-1$,

(d) Es ist $p_i^T A p_k = 0$, $i = 0, \ldots, k-1$, d. h. die Richtungen $p_0, \ldots, p_k$ sind A-konjugiert.

Insbesondere bricht das Verfahren nach $m \leq n$ Schritten ab.

4. Man zeige: Hat die symmetrische, positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ nur r verschiedene Eigenwerte, so bricht das Verfahren der konjugierte Gradienten, angewandt auf das lineare Gleichungssystem $Ax = b$ bzw. die äquivalente unrestringierte Optimierungsaufgabe, nach spätestens r Iterationsschritten ab.

   Hinweis: Man benutze die im Beweis von Satz 3.3 bewiesene Aussage:

   - Ist x_k die durch das Verfahren der konjugierten Gradienten gewonnene k-te Iterierte, sind $\lambda_1, \ldots, \lambda_n$ die Eigenwerte der symmetrischen, positiv definiten Matrix A und ist $p \in \Pi_k$ ein beliebiges Polynom mit $p(0) = 1$, so ist

   $$\|x_k - x^*\|_A \leq \max_{i=1,\ldots,n} |p(\lambda_i)| \, \|x_0 - x^*\|_A.$$

   Hierbei ist die Norm $\|\cdot\|_A$ durch $\|x\|_A := \sqrt{x^T A x}$ auf dem $\mathbb{R}^n$ definiert.

5. Das Verfahren der konjugierten Gradienten von Polak-Ribière unterscheidet sich von dem Fletcher-Reeves-Verfahren nur darin, dass $\beta_k := g_{k+1}^T (g_{k+1} - g_k)/\|g_k\|_2^2$ (statt $\beta_k := \|g_{k+1}\|_2^2/\|g_k\|_2^2$) gesetzt wird. Man betrachte das dann definierte Polak-Ribière-Verfahren mit exakter Schrittweitenstrategie zur Lösung von

   (P) $$\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n.$$

   Man zeige:

   (a) Sind die Voraussetzungen (K) (a)–(c) erfüllt, so liefert das Verfahren, wenn es nicht vorzeitig mit der Lösung x^* von (P) abbricht, eine Folge $\{x_k\}$, die R-linear gegen x^* konvergiert.

   (a) Die Voraussetzungen (V) (a)–(c) seien erfüllt, das Verfahren breche nicht vorzeitig mit einer stationären Lösung von (P) ab. Für die durch das Verfahren erzeugte Folge $\{x_k\}$ gelte $\lim_{k\to\infty} \|x_{k+1} - x_k\|_2 = 0$. Dann ist $\liminf_{k\to\infty} \|g_k\|_2 = 0$, so dass die Folge $\{x_k\}$ wenigstens eine stationäre Lösung von (P) als Häufungspunkt besitzt.

   Hinweis: Man setze

   $$\delta_k := \left(\frac{g_k^T p_k}{\|g_k\|_2 \, \|p_k\|_2} \right)^2 = \frac{\|g_k\|_2^2}{\|p_k\|_2^2}.$$

   Für den ersten Teil der Aufgabe zeige man die Existenz einer Konstanten $\delta > 0$ mit $\delta_k \geq \delta$, $k = 0, 1, \ldots$, und wende Satz 1.7 an. Für den zweiten Teil der Aufgabe mache man einen Widerspruchsbeweis. Zum einen ist $\sum_{k=0}^{\infty} \delta_k < \infty$ unter Benutzung von Satz 1.2, zum anderen $1/\delta_{k+1} \leq 1 + 1/\delta_k$ für alle hinreichend großen k und daher $\sum_{k=0}^{\infty} \delta_k = \infty$.

6. Unter den Voraussetzungen (V) (a)–(c) aus Unterabschnitt 3.1.1 betrachte man zur Lösung der unrestringierten Optimierungsaufgabe (P) das Fletcher-Reeves-Verfahren mit der strengen Wolfe-Schrittweite (siehe Aufgabe 1 in Abschnitt 3.1):

   - Für die Schrittweitenstrategie seien α und β mit $0 < \alpha < \beta < \frac{1}{2}$ gegeben.
   - Gegeben $x_0 \in \mathbb{R}^n$, berechne $g_0 := \nabla f(x_0)$ und setze $p_0 := -g_0$.
   - Für $k = 0, 1, \ldots$:
     - Falls $g_k = 0$, dann: STOP. x_k ist stationäre Lösung von (P).
     - Andernfalls:
       * Bestimme eine strenge Wolfe-Schrittweite, also ein $t_k > 0$ mit

$$f(x_k + t_k p_k) \le f(x_k) + \alpha t_k g_k^T p_k, \qquad |\nabla f(x_k + t_k p_k)^T p_k| \le -\beta g_k^T p_k.$$

       * Setze bzw. berechne

$$x_{k+1} := x_k + t_k p_k, \qquad g_{k+1} := \nabla f(x_{k+1})$$

         sowie

$$\beta_k := \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2}, \qquad p_{k+1} := -g_{k+1} + \beta_k p_k.$$

   Man zeige:

   (a) Ist im k-ten Schritt noch kein Abbruch erfolgt, ist also $g_0, \ldots, g_k \neq 0$, so ist

$$-\frac{1}{1-\beta} \le -\sum_{j=0}^{k} \beta^j \le \frac{g_k^T p_k}{\|g_k\|_2^2} \le -2 + \sum_{j=0}^{k} \beta^j < -\frac{1-2\beta}{1-\beta}.$$

   Wegen $\beta \in (0, \frac{1}{2})$ ist daher p_k eine Abstiegsrichtung in x_k. Mit Hilfe von Aufgabe 1 in Abschnitt 3.1 folgt die Existenz einer Schrittweite $t_k > 0$ mit den geforderten Eigenschaften.

   (b) Bricht das Verfahren nicht vorzeitig mit einer stationären Lösung ab, so erzeugt es eine Folge $\{x_k\}$ mit $\liminf_{k\to\infty} \|g_k\|_2 = 0$. Sind sogar die Voraussetzungen (K) (a)–(c) erfüllt, so konvergiert die gesamte Folge $\{x_k\}$ gegen die dann eindeutige Lösung x^* von (P).

3.4 Das Gauß-Newton-Verfahren

In diesem Abschnitt untersuchen wir das (durch die Armijo-Schrittweite gedämpfte) Gauß-Newton-Verfahren zur Lösung der diskreten, nichtlinearen Approximationsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n.$$

Hierbei ist $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ und $\|\cdot\|$ eine das Problem bestimmende, fest vorgegebene Norm auf dem $\mathbb{R}^m$ ist. Ist z. B. $\|\cdot\| = \|\cdot\|_2$ die euklidische Norm, so handelt es sich bei (P) um ein nichtlineares Ausgleichsproblem.

3.4.1 Die Konvergenz des Gauß-Newton-Verfahrens

Analog zu den Voraussetzungen (V) (a)–(c) in Unterabschnitt 3.1.1 formulieren wir diesmal als generelle Voraussetzungen:

(V) (a) Mit einem gegebenen $x_0 \in \mathbb{R}^n$ (Startwert des Iterationsverfahrens) ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ kompakt.

(b) Die Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ ist auf einer offenen Obermenge von L_0 stetig differenzierbar.

(c) Die Funktion $F'(\cdot)$ ist auf L_0 lipschitzstetig, d. h. es existiert eine Konstante $\gamma > 0$ mit

$$\|F'(x) - F'(y)\| \leq \gamma \|x - y\| \qquad \text{für alle } x, y \in L_0.$$

(Hierbei ist rechts $\|\cdot\|$ eine feste Norm auf dem $\mathbb{R}^n$, während $\|\cdot\|$ links die der vorgebenen Norm auf dem $\mathbb{R}^m$ und der Norm auf dem $\mathbb{R}^n$ zugeordnete Matrixnorm auf $\mathbb{R}^{m\times n}$ bedeutet[31]).

Die Zielfunktion f von (P) kann auch als

$$f(x) := g \circ F(x) = g(F(x))$$

geschrieben werden, wobei $g\colon \mathbb{R}^m \longrightarrow \mathbb{R}$ durch $g(y) := \|y\|$ gegeben ist. Daher besitzt unter den Voraussetzungen (V) die Zielfunktion f von (P) in jedem $x \in L_0$ eine Gateaux-Variation $f'(x;\cdot)\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, die durch $f'(x;p) = g'(F(x);F'(x)p)$ gegeben ist. Dieses Ergebnis haben wir in Satz 1.8 in Abschnitt 2.1 exemplarisch für $g(y) := \|y\|_\infty$ bewiesen.

Im folgenden Lemma wird gezeigt, wie man in einer aktuellen Näherung $x_c \in L_0$ eine Abstiegsrichtung berechnen oder feststellen kann, dass x_c eine stationäre Lösung von (P) ist.

Lemma 4.1 *Gegeben sei die diskrete, nichtlineare Approximationsaufgabe (P), die Voraussetzungen (V) (a)–(c) seien erfüllt. Bei gegebenem $x_c \in L_0$ betrachte man die in x_c linearisierte Aufgabe*

$$(\mathrm{LP}_c) \qquad \textit{Minimiere} \quad f_c(p) := \|F(x_c) + F'(x_c)p\|, \quad p \in \mathbb{R}^n.$$

Dann existiert eine (nicht notwendig eindeutige) Lösung p^* zu (LP_c). Ferner gilt:

1. *Ist*[32] $f_c(p^*) = f(x)$ *, so ist x_c eine stationäre Lösung von (P), d. h. es ist $f'(x_c;q) \geq 0$ für alle $q \in \mathbb{R}^n$.*

[31] Die Lipschitzstetigkeit ist wegen der Äquivalenz der Normen auf einem endlichdimensionalen linearen Raum eine normunabhängige Eigenschaft. Die eben getroffene Vereinbarung dient der Festlegung der Lipschitzkonstanten γ.

[32] Man beachte: Zwar ist p^* als Lösung von (LP_c) nicht notwendig eindeutig bestimmt, aber $f_c(p^*) = \min_{p\in\mathbb{R}^n} f_c(p)$ ist natürlich eindeutig.

2. *Ist* $f_c(p) < f(x_c)$ *für ein* $p \in \mathbb{R}^n$, *so ist*

$$f'(x_c; p) \leq f_c(p) - f(x_c) < 0,$$

also p *eine Abstiegsrichtung für* f *in* x_c. *Insbesondere ist* p^* *eine Abstiegsrichtung für* f *in* x_c, *wenn* $f_c(p^*) \neq f(x_c)$.

Beweis: Um die Lösbarkeit von (LP_c) nachzuweisen, betrachten wir die Aufgabe

$$\text{Minimiere} \quad g_c(q) := \|F(x) + q\|, \quad q \in \text{Bild}\,(F'(x_c)).$$

Diese Aufgabe besitzt eine Lösung $q^* = F'(x_c)p^*$, da eine zugehörige Niveaumenge offensichtlich kompakt ist. Dann ist p^* eine Lösung von (LP_c).

Ist $f_c(p^*) = f(x_c)$, so ist $p = 0$ eine Lösung und damit auch eine stationäre Lösung von (LP_c). Folglich ist

$$0 \leq f_c'(0; q) = g'(F(x_c); F'(x_c)q) = f'(x_c; q) \qquad \text{für alle } q \in \mathbb{R}^n,$$

und daher x_c eine stationäre Lösung von (P).

Ist $f_c(p) < f(x_c)$, so ist

$$f'(x_c; p) = g'(F(x_c); F'(x_c)p) \leq g(F(x_c) + F'(x_c)p) - g(F(x)) = f_c(p) - f(x_c) < 0,$$

d. h. p ist eine Abstiegsrichtung für die Zielfunktion f der diskreten, nichtlinearen Aooroximationsaufgabe (P) in der aktuellen Näherung x_c. Ist $f_c(p^*) \neq f(x_c) = f_c(0)$, so ist notwendig $f_c(p^*) < f(x_c)$ und damit p^* eine Abstiegsrichtung in x_c. Das Lemma ist damit bewiesen. □

Bemerkung: Für $\|\cdot\| = \|\cdot\|_2$ ist (LP_c) ein lineares Ausgleichsproblem, das unter der Voraussetzung Rang $(F'(x_c)) = n \leq m$ eine eindeutige Lösung besitzt, welche z. B. mit Hilfe einer QR-Zerlegung von $F'(x_c)$ berechnet werden kann. Für Rang $(F'(x_c)) < n \leq m$ ist (LP_c) zwar nicht eindeutig lösbar, es existiert aber genau eine Lösung minimaler euklidischer Norm. Diese kann mit Hilfe einer *Singulärwertzerlegung* von $F'(x_c)$ berechnet werden. Eine Realisierung in Matlab ist denkbar einfach, hierauf gehen wir später ein.

Ist $\|\cdot\| = \|\cdot\|_\infty$, so ist (LP_c) äquivalent der linearen Optimierungsaufgabe

$$\begin{cases} \text{Minimiere} \quad \delta \quad \text{unter den Nebenbedingungen} \\ \qquad -\delta e \leq F(x_c) + F'(x_c)p \leq \delta e, \end{cases}$$

wobei $e \in \mathbb{R}^m$ der Vektor ist, dessen Komponenten alle gleich 1 sind. Mit einem Verfahren für lineare Programme, etwa dem Simplexverfahren, kann also (LP_c) in diesem Falle gelöst werden. Ähnliches gilt auch für $\|\cdot\| = \|\cdot\|_1$. □

Durch Lemma 4.1 ist eine Richtungsstrategie für ein Verfahren zur Lösung von (P) gegeben. Nun kommen wir zur Definition einer Schrittweitenstrategie, wobei wir uns auf die Übertragung der Armijo-Schrittweite vom glatten auf den hier vorliegenden Fall beschränken. Seien eine aktuelle Näherung $x_c \in L_0$ und eine Richtung $p \in \mathbb{R}^n$ mit $f_c(p) := \|F(x_c) + F'(x_c)p\| < f(x_c)$ vorgegeben. Wegen des eben bewiesenen Lemmas 4.1 ist p eine Abstiegsrichtung für die Zielfunktion f von (P) in x_c. Analog zum glatten Fall definieren wir die zugehörige *Armijo-Schrittweite* durch den folgenden Algorithmus:

- Seien $\alpha \in (0, \frac{1}{2})$ und $0 < l \leq u < 1$ gegeben.
- Setze $\rho_0 := 1$.
- Für $j = 0, 1, \ldots$:
  - Falls $f(x_c + \rho_j p) \leq f(x_c) + \alpha\rho_j[f_c(p) - f(x_c)]$, dann: $t := \rho_j$, STOP.
  - Andernfalls: Wähle $\rho_{j+1} \in [l\rho_j, u\rho_j]$.

Es ist klar, dass die Armijo-Schrittweite existiert bzw. der obige Algorithmus nach endlich vielen Schritten abbricht. Denn wäre die zu testende Ungleichung für kein j erfüllt, so wäre $\{\rho_j\} \subset \mathbb{R}$ eine Nullfolge und

$$\alpha[f_c(p) - f(x_c)] \leq \lim_{j\to\infty} \frac{f(x_c + \rho_j p) - f(x_c)}{\rho_j} = f'(x_c; p) \leq f_c(p) - f(x_c)$$

würde den Widerspruch $0 \leq (1 - \alpha)[f_c(p) - f(x_c)]$ ergeben. Wie im glatten Fall kann $l = u =: \rho$ gewählt werden, so dass dann die Armijo-Schrittweite durch $t = \rho^j$ gegeben ist, wobei j die kleinste nichtnegative ganze Zahl mit

$$f(x_c + \rho^j p) \leq f(x_c) + \alpha\rho^j[f_c(p) - f(x_c)]$$

ist. Entsprechend dem glatten Fall kann aber auch $\rho_0 := 1$ und

$$\rho_{j+1} := \max(0.1\rho_j, \rho_j^*) \qquad \text{mit} \quad \rho_j^* := \frac{\rho_j^2[f_c(p) - f(x_c)]}{2[f(x_c + \rho_j p) - (f(x_c) + \rho_j[f_c(p) - f(x_c)])]}$$

gesetzt werden.

Das folgende Lemma wird dazu dienen, ganz entsprechend zu Lemma 1.1, die durch die Armijo-Schrittweite erzielte Verminderung der Zielfunktion nach unten abzuschätzen.

Lemma 4.2 *Gegeben sei die diskrete, nichtlineare Approximationsaufgabe (P), die Zielfunktion f genüge den Voraussetzungen (V) (a)–(c). Sei $x_c \in L_0$ und $p \in \mathbb{R}^n$ eine Richtung mit $f_c(p) := \|F(x_c) + F'(x_c)p\| < f(x_c)$, also p eine Abstiegsrichtung für f in x_c. Mit*

$$t^* := -\frac{2[f(x_c) - f(x_c)]}{\gamma\|p\|^2}$$

ist dann

$$f(x_c + tp) \leq f)x_c) + t[f_c(p) - f(x_c)] + t^2\frac{\gamma}{2}\|p\|^2 \qquad \text{für alle } t \in [0, \min(1, t^*)].$$

Beweis: Wie in Lemma 1.1 sei $\hat{t} = \hat{t}(x_c, p)$ die erste positive Nullstelle der durch $\psi(t) := f(x_c) - f(x_c + tp)$ definierten Abbildung $\psi\colon [0, \infty) \longrightarrow \mathbb{R}$. Dann ist $x_c + sp \in L_0$ für alle $s \in [0, \hat{t}]$. Für $t \in [0, \hat{t}]$ ist

$$\begin{aligned} F(x_c + tp) &= F(x_c) + tF'(x_c)p + \int_0^t [F'(x_c + sp) - F'(x_c)]p\, ds \\ &= (1 - t)F(x_c) + t[F(x_c) + F'(x_c)p] + \int_0^t [F'(x_c + sp) - F'(x_c)]p\, ds. \end{aligned}$$

Nimmt man hier auf beiden Seiten die Norm, wendet die Dreiecksungleichung und die Lipschitzstetigkeit von $F'(\cdot)$ an, so erhält man

$$\begin{aligned} f(x_c + tp) &\leq (1-t)f(x_c) + tf_c(p) + t^2\frac{\gamma}{2}\|p\|^2 \\ &= f(x_c) + t[f_c(p) - f(x_c)] + t^2\frac{\gamma}{2}\|p\|^2 \end{aligned}$$

für alle $t \in [0, \min(1, \hat{t})]$. Ist $\hat{t} \leq 1$, so folgt hieraus (setze $t = \hat{t}$), dass $t^* \leq \hat{t}$. □ □

Entsprechend zu Satz 1.4 kann auch hier die Verminderung der Zielfunktion abgeschätzt werden. Da der Beweis völlig analog verläuft (statt Lemma 1.1 wird Lemma 4.2 benutzt), geben wir ihn nicht an, sondern verweisen auf Aufgabe 1.

Satz 4.3 *Gegeben sei die diskrete, nichtlineare Approximationsaufgabe (P), die Zielfunktion f genüge den Voraussetzungen (V) (a)–(c). Sei $x_c \in L_0$ und $p \in \mathbb{R}^n$ eine Richtung mit $f_c(p) := \|F(x_c) + F'(x_c)p\| < f(x_c)$. Seien $\alpha \in (0, \frac{1}{2})$, $0 < l \leq u < 1$ gegeben und $t := \rho_j$ eine zugehörige Armijo-Schrittweite. Dann existiert eine Konstante $\theta > 0$, die nur von α, γ sowie l und u, nicht aber von x_c oder p abhängt, mit*

$$f(x_c) - f(x_c + tp) \geq \theta\Big[f(x_c) - f_c(p), \Big(\frac{f(x_c) - f_c(p)}{\|p\|}\Big)^2\Big].$$

Im folgenden Satz wird das durch die Armijo-Schrittweite *gedämpfte Gauß-Newton-Verfahren* zur Lösung von (P) angegeben und hierfür eine Konvergenzaussage gemacht.

Satz 4.4 *Gegeben sei die diskrete, nichtlineare Approximationsaufgabe (P), die Zielfunktion f genüge den Voraussetzungen (V) (a)–(c). Zusätzlich sei* Rang $(F'(x)) = n$ *für alle $x \in L_0$. Zur Lösung von (P) betrachte man den folgenden Algorithmus:*

- *Seie $\alpha \in (0, \frac{1}{2})$ und $0 < l \leq u < 1$ (für die Armijo-Schrittweite) vorgegeben.*
- *Sei $x_0 \in \mathbb{R}^n$ ein Startwert (wie in (V)).*
- *Für $k = 0, 1, \ldots$:*
  - *Berechne eine Lösung p_k der in x_k linearisierten Approximationsaufgabe*

    (LP_k) *Minimiere* $f_k(p) := \|F(x_k) + F'(x_k)p\|, \quad p \in \mathbb{R}^n.$

  - *Falls $f_k(p_k) = f(x_k)$, dann: x_k ist stationäre Lösung von (P), STOP.*
  - *Andernfalls:*
    - *Berechne Armijo-Schrittweite t_k.*
    - *Setze $x_{k+1} := x_k + t_k p_k$.*

Dann gilt: Bricht der Algorithmus nicht nach endlich vielen Schritten mit einer stationären Lösung von (P) ab, so liefert er eine Folge $\{x_k\} \subset L_0$ mit $f(x_{k+1}) < f(x_k)$, $k = 0, 1, \ldots$, und der Eigenschaft, dass jeder Häufungspunkt x^* von $\{x_k\}$ eine stationäre Lösung von (P) ist.

Beweis: Die Durchführbarkeit des Algorithmus ist wegen Lemma 4.1 sowie der Existenz der Armijo-Schrittweite klar. O. B. d. A. nehmen wir an, das Verfahren breche nicht vorzeitig mit einer stationären Lösung von (P) ab. Da es sich um ein Abstiegsverfahren handelt, wird eine Folge $\{x_k\} \subset L_0$ mit $f(x_{k+1}) < f(x_k)$, $k = 0, 1, \ldots$, erzeugt. Sei x^* ein wegen der Kompaktheit der Niveaumenge L_0 existierender Häufungspunkt von $\{x_k\}$. Der Nachweis dafür, dass x^* eine stationäre Lösung von (P) ist, erfolgt in drei Schritten.

(a) Die Folge $\{p_k\}$ ist beschränkt.

Denn: Es ist

$$\|F'(x_k)p_k\| \leq \underbrace{\|F(x_k)\|}_{=f(x_k)} + \underbrace{\|F(x_k) + F'(x_k)p_k\|}_{<f(x_k)} < 2f(x_k).$$

Nach Voraussetzung ist Rang $(F'(x)) = n$ für jedes $x \in L_0$, daher ist $F'(x)^T F'(x) \in \mathbb{R}^{n\times n}$ für jedes $x \in L_0$ nichtsingulär. Folglich ist

$$\begin{aligned}
\|p_k\| &= \|[F'(x_k)^T F'(x_k)]^{-1} F'(x_k)^T F'(x_k)p_k\| \\
&\leq \|[F'(x_k)^T F'(x_k)]^{-1} F'(x_k)^T\| \, \|F'(x_k)p_k\| \\
&\leq 2\|[F'(x_k)^T F'(x_k)]^{-1} F'(x_k)^T\| \, f(x_k) \\
&\leq \Big(2 \max_{x\in L_0} \|[F'(x)^T F'(x)]^{-1} F'(x)^T\|\Big) f(x_k) \\
&\leq \Big(2 \max_{x\in L_0} \|[F'(x)^T F'(x)]^{-1} F'(x)^T\|\Big) f(x_0),
\end{aligned}$$

womit die Beschränktheit der Folge $\{p_k\}$ bewiesen ist.

(b) Es ist $\lim_{k\to\infty}[f(x_k) - f_k(p_k)] = 0$.

Denn: Wegen Satz 4.3 existiert eine (von k unabhängige) Konstante $\theta > 0$ mit

$$(*) \quad f(x_k) - f(x_{k+1}) \geq \theta \min\Big[f(x_k) - f_k(p_k), \Big(\frac{f(x_k) - f_k(p_k)}{\|p_k\|}\Big)^2\Big], \qquad k = 0, 1, \ldots.$$

Da $\{f(x_k)\}$ eine monoton fallende, nach unten beschränkte Folge ist, gilt

$$\lim_{k\to\infty} [f(x_k) - f(x_{k+1})] = 0.$$

Hieraus, aus der Beschränktheit von $\{p_k\}$ sowie $(*)$ folgt $\lim_{k\to\infty}[f(x_k) - f_k(p_k)] = 0$.

(c) Da x^* ein Häufungspunkt von $\{x_k\}$ und $\{p_k\}$ nach (a) beschränkt ist, existiert eine unendliche Teilmenge $K \subset \mathbb{N}$ derart, dass $\{x_k\}_{k\in K}$ gegen x^* und $\{p_k\}_{k\in K}$ gegen ein p^* konvergiert. Dann ist p^* eine Lösung von

$$(\mathrm{LP}_*) \qquad \text{Minimiere} \quad f_*(p) := \|F(x^*) + F'(x^*)p\|, \quad p \in \mathbb{R}^n,$$

mit $f_*(p^*) = f(x^*)$, so dass x^* nach Lemma 4.1 eine stationäre Lösung von (P) ist.

Denn: Sei $p \in \mathbb{R}^n$ beliebig. Nach Definition von p_k als Lösung von (LP_k) ist

$$\|F(x_k) + F'(x_k)p_k\| \leq \|F(x_k) + F'(x_k)p\|, \qquad k = 0, 1, \ldots.$$

Lässt man hier $k \in K$ nach ∞ laufen, so erhält man, dass p^* eine Lösung von (LP_*) ist. Mit (b) folgt

$$0 = \lim_{k \in K, k \to \infty} [f(x_k) - f_k(p_k)] = f(x^*) - f_*(p^*).$$

Also ist $f_*(p^*) = f(x^*)$ und der Satz ist bewiesen. □ □

3.4.2 Nichtlineare Ausgleichsprobleme

Ziel dieses Unterabschnittes ist es, eine einfache Matlab-Funktion zur Lösung des nichtlinearen Ausgleichsproblems

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \|F(x)\|_2, \quad x \in \mathbb{R}^n,$$

anzugeben. Die obigen Voraussetzungen (V) (a)–(c) seien erfüllt. Einfacher wird die Umsetzung des Gauss-Newton-Verfahrens vor allem dadurch, dass die Lösung des in der aktuellen Näherung x_c linearisierten Problems

$$(\mathrm{LP}_c) \qquad \text{Minimiere} \quad f_c(p) := \|F(x_c) + F'(x_c)p\|_2, \quad p \in \mathbb{R}^n,$$

in Matlab außerordentlich einfach erhalten werden kann. Ein File `GauNew.m` hat bei uns den folgenden Inhalt:

```
function [x,iter]=GauNew(Fun,x_0,max_iter,tol);
%****************************************************
%The Gauss-Newton method with Armijo line search is used
%to solve the nonlinear least squares problem
%       minimize   f(x):=||F(x)||_2
%****************************************************
%Input parameter:
%             Fun         function, [F(x),F'(x)]=Fun(x)
%             x_0         initial iterate
%             max_iter    maximal number of iterations
%             tol         tolerance
%Output parameter:
%             x           approximate solution
%             iter        number of iterations performed
%****************************************************
x_c=x_0; iter=0; [F_c,J_c]=feval(Fun,x_c); p=-J_c\F_c;
f=norm(F_c); f_c=norm(F_c+J_c*p);
error=f-f_c;
while (error>tol)&(iter<max_iter)
    iter=iter+1;
    t=ArmGauNew(x_c,p,Fun,f,f_c);
    x_c=x_c+t*p;[F_c,J_c]=feval(Fun,x_c); p=-J_c\F_c;
    f=norm(F_c); f_c=norm(F_c+J_c*p);
    error=f-f_c;
```

```
end;
if (error<=tol)
    x=x_c;
end;
%*******************************************************
function t=ArmGauNew(x_c,p,Fun,f,f_c);
%*******************************************************
%The Armijo steplength is computed at an current iterate
%x_c in the direction p.
%*******************************************************
%Input parameter:
%               x_c         current iterate
%               p           direction
%               Fun         F(x)=Fun(x)
%               f           ||F(x_c)||_2
%               f_c         ||F(x_c)+F'(x_c)p||_2
%Output parameter:
%               t           Armijo steplength
%*******************************************************
alpha=0.0001;rho=1.0;
x_plus=x_c+rho*p;F_plus=feval(Fun,x_plus);f_plus=norm(F_plus);
while (f_plus>f+alpha*rho*(f_c-f))
    rho_star=0.5*rho^2*(f_c-f)/(f_plus-f-rho*(f_c-f));
    rho=max(0.1*rho,rho_star);
    x_plus=x_c+rho*p;F_plus=feval(Fun,x_plus);f_plus=norm(F_plus);
end;
t=rho;
```

Als Test wollen wir die Rosenbrock-Funktion minimieren. Hier ist

$$F(x) := \begin{pmatrix} 10(x_2 - x_1^2) \\ 1 - x_1 \end{pmatrix}.$$

Wir schreiben ein fuction file `Rose.m` mit dem Inhalt

```
function [F,J]=Rose(x);
F=[10*(x(2)-x(1)^2);1-x(1)]
if nargout>1
    J=[-20*x(1),10;-1,0];
end;
```

Ein Aufruf

```
>> [x,iter]=GauNew('Rose',[-1.2;1],100,1e-8)
```

liefert die exakte Lösung $(1,1)^T$, die in 18 Iterationsschritten erreicht wurde.

Beispiel: Wir kehren zu einem Beispiel aus dem ersten Kapitel zurück. Hierbei handelt es sich um die Aufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(a_1,a_2,a_3,\alpha_1,\alpha_2) := \displaystyle\sum_{i=1}^{9} [a_1 + a_2 e^{\alpha_1 t_i} + a_3 e^{\alpha_2 t_i} - z_i]^2, \\ (a_1,a_2,a_3,\alpha_1,\alpha_2) \in \mathbb{R}^5, \end{cases}$$

wobei die Daten durch

t_i	0.0	0.5	1.0	1.5	2.0	3.0	5.0	8.0	10.0
z_i	3.85	2.95	2.63	2.33	2.24	2.05	1.82	1.80	1.75

gegeben. Da wir das Beispiel aus dem Lehrbuch H. R. SCHWARZ (1988, S. 318) genommen haben, schreiben wir ein function file `Schwarz.m` mit dem Inhalt:

```
function [F,J]=Schwarz(x);
t=[0.;0.5;1.0;1.5;2.0;3.0;5.0;8.0;10.0];
z=[3.85;2.95;2.63;2.33;2.24;2.05;1.82;1.80;1.75];
F=x(1)+x(2)*exp(x(4)*t)+x(3)*exp(x(5)*t)-z;
if nargout>1
   J=[ones(9,1),exp(x(4)*t),exp(x(5)*t),x(2)*t.*(exp(x(4)*t)),x(3)*t.*exp(x(5)*t)];
end;
```

Im folgenden Aufruf werden die von Schwarz vorgeschlagenen (ziemlich guten) Näherungswerte genommen:

```
[x,iter]=GauNew('Schwarz',[1.75;1.20;0.8;-0.5;-2],100,1e-8);
```

Wir erhalten

$$x = \begin{pmatrix} 1.7577 \\ 1.4208 \\ 0.6709 \\ -0.5552 \\ -3.3816 \end{pmatrix}, \qquad \text{iter} = 4.$$

□

3.4.3 Nichtlineare Tschebyscheff-Approximation

Nur kurz wollen wir auf die diskrete, nichtlineare Tschebyscheffsche Approximationsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \|F(x)\|_\infty, \quad x \in \mathbb{R}^n,$$

eingehen, wobei wir wieder annehmen, dass (V) (a)–(c) gilt. Der einzige Unterschied besteht in der Methode zur Lösung des Hilfsproblems, also von

$$(\text{LP}_c) \qquad \text{Minimiere} \quad f_c(p) := \|F(x_c) + F'(x_c)p\|_\infty, \quad p \in \mathbb{R}^n.$$

Wir hatten uns früher schon überlegt, dass diese Aufgabe äquivalent der linearen Optimierungsaufgabe

$$\begin{cases} \text{Minimiere} \quad \delta \quad \text{unter den Nebenbedingungen} \\ \qquad -\delta e \le F(x_c) + F'(x_c)p \le \delta e, \end{cases}$$

ist, wobei $e \in \mathbb{R}^m$ der Vektor ist, dessen Komponenten alle gleich 1 sind. Diese lineare Optimierungsaufgabe hat die Form

$$\begin{cases} \text{Minimiere} \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} p \\ \delta \end{pmatrix} \quad \text{unter der Nebenbedingung} \\ \qquad\qquad \begin{pmatrix} J_c & -e \\ -J_c & -e \end{pmatrix} \begin{pmatrix} p \\ \delta \end{pmatrix} \leq \begin{pmatrix} -F_c \\ F_c \end{pmatrix}. \end{cases}$$

Eine Umsetzung in Matlab ist am einfachsten, wenn die Optimization Toolbox zur Verfügung steht. In der folgenden Funktion `DisTsch` gehen wir davon aus. Wieder haben wir uns nicht sonderlich um Effizienz bemüht. Vielmehr kam es uns darauf an, obige Funktion `GauNew` zur Lösung nichtlinearer Ausgleichsprobleme möglichst wenig zu ändern. Die Änderungen betreffen eigentlich nur die Lösung des Hilfsproblems, ferner wird die Funktion `norm` mit dem Zusatz `inf` benutzt, um ihr zu sagen, dass wir diesmal die Maximum-Norm benutzen.

```
function [x,iter]=DisTsch(Fun,x_0,max_iter,tol);
%*******************************************************
%The Gauss-Newton method with Armijo line search is used
%to solve the nonlinear Tschebyscheff problem
%       minimize   f(x):=||F(x)||_inf
%*******************************************************
%Input parameter:
%             Fun         function, [F(x),F'(x)]=Fun(x)
%             x_0         initial iterate
%             max_iter    maximal number of iterations
%             tol         tolerance
%Output parameter:
%             x           approximate solution
%             iter        number of iterations performed
%*******************************************************
options=optimset('LargeScale','off','Display','off');
x_c=x_0; iter=0; [F_c,J_c]=feval(Fun,x_c);f=norm(F_c,inf);
[m,n]=size(J_c);e=ones(m,1);c=[zeros(n,1);1];
A=[J_c,-e;-J_c,-e];b=[-F_c;F_c];
[z,f_c]=linprog(c,A,b,[],[],[],[],[],options);
p=z(1:n);
error=f-f_c;
while (error>tol)&(iter<max_iter)
    iter=iter+1;
    t=ArmGauNew(x_c,p,Fun,f,f_c);
    x_c=x_c+t*p;
    [F_c,J_c]=feval(Fun,x_c);f=norm(F_c,inf);
    A=[J_c,-e;-J_c,-e];b=[-F_c;F_c];
    [z,f_c]=linprog(c,A,b,[],[],[],[],[],options);
    p=z(1:n);
    error=f-f_c;
end;
if (error<=tol)
    x=x_c;
end;
%*******************************************************
```

```
function t=ArmGauNew(x_c,p,Fun,f,f_c);
%*****************************************************
%The Armijo steplength is computed at an current iterate
%x_c in the direction p.
%*****************************************************
%Input parameter:
%               x_c         current iterate
%               p           direction
%               Fun         F(x)=Fun(x)
%               f           ||F(x_c)||_inf
%               f_c         ||F(x_c)+F'(x_c)p||_inf
%Output parameter:
%               t           Armijo steplength
%*****************************************************
alpha=0.0001;rho=1.0;
x_plus=x_c+rho*p;F_plus=feval(Fun,x_plus);f_plus=norm(F_plus,inf);
while (f_plus>f+alpha*rho*(f_c-f))
    rho_star=0.5*rho^2*(f_c-f)/(f_plus-f-rho*(f_c-f));
    rho=max(0.1*rho,rho_star);
    x_plus=x_c+rho*p;F_plus=feval(Fun,x_plus);f_plus=norm(F_plus,inf);
end;
t=rho;
```

Nach dem Aufruf

```
[x,iter]=DisTsch('Schwarz',[1.75;1.20;0.8;-0.5;-2],100,1e-8);
```

erhalten wir

$$x = \begin{pmatrix} 1.7812 \\ 1.4025 \\ 0.6316 \\ -0.5971 \\ -2.7079 \end{pmatrix}, \qquad \text{iter} = 4.$$

In Abbildung 3.3 haben wir die (t_i, z_i), $i = 1, \ldots, 9$, durch ein Kreuz markiert und die Funktion $y(s) = x_1 + x_2 e^{x_4 s} + x_3 e^{x_5 s}$ geplottet. Diesen Plot (man könnte die Kreuze vergrößern, einen Text einfügen, die Achsen beschriften usw.) haben wir gewonnen durch

```
>> t=[0.;0.5;1.0;1.5;2.0;3.0;5.0;8.0;10.0];
>> z=[3.85;2.95;2.63;2.33;2.24;2.05;1.82;1.80;1.75];
>> plot(t,z,'x');s=linspace(0,10);
>> hold on
>> y=x(1)+x(2)*exp(x(4)*s)+x(3)*exp(x(5)*s);
>> plot(s,y);
```

wobei natürlich x durch die Lösung besetzt ist.

Beispiel: Wir wollen auf ein Beispiel aus Unterabschnitt 2.1.1 zurückkommen. Und zwar betrachten wir das diskrete Tschebyscheffsche Approximationsproblem

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \max_{i=1,2} |F_i(x)|, \quad x \in \mathbb{R}^2,$$

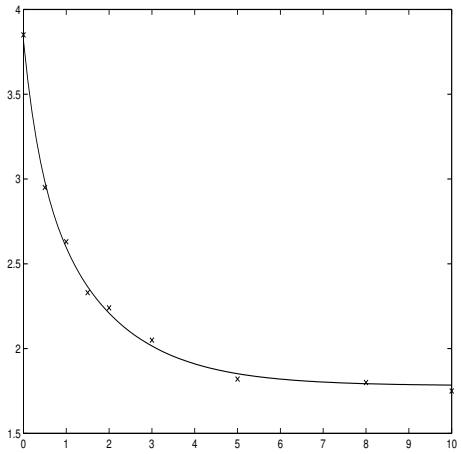

Abbildung 3.3: Lösung eines diskreten Tschebyscheff-Problems

wobei

$$\begin{aligned} F_1(x) &:= x_1 - x_2^3 + 5x_2^2 - 2x_2 - 13, \\ F_2(x) &:= x_1 + x_2^3 + x_2^2 - 14x_2 - 29. \end{aligned}$$

Da dies Beispiel von Fletcher-Watson stammt, schreiben wir ein file `FleWat.m` mit naheliegendem Inhalt. Der Aufruf

```
[x,iter]=DisTsch('FleWat',[2;3],100,1e-8);
```

liefert (nach `format long`) das Resultat

$$x = \begin{pmatrix} 4.99999999999654 \\ 4.00000000000042 \end{pmatrix}, \qquad \text{iter} = 6.$$

Bei der Berechnung des lokalen Minimums bei $(11.4, -0.9)$ waren wir allerdings nicht erfolgreich. □

3.4.4 Starke Eindeutigkeit, Superlineare Konvergenz

In diesem Unterabschnitt wollen wir annehmen, das gedämpfte Gauß-Newton-Verfahren aus Satz 4.4 liefere eine gegen ein x^* konvergente Folge $\{x_k\}$. Uns interessiert die Frage, ob unter geeigneten Voraussetzungen $t_k = 1$ für alle hinreichend großen k, ob also das gedämpfte Gauß-Newton-Verfahren nach endlich vielen Schritten in das ungedämpfte übergeht. Eine weiter Frage besteht darin, ob etwas über die Konvergenzgeschwindigkeit ausgesagt werden kann. Von entscheidender Bedeutung bei der Beantwortung dieser Fragen ist die folgende Definition.

Definition 4.5 Gegeben sei die diskrete, nichtlineare Approximationsaufgabe

(P) $$\text{Minimiere} \quad f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n.$$

Ein $x^* \in \mathbb{R}^n$ heißt *lokal stark eindeutige Lösung* von (P), wenn es positive Konstanten σ und δ mit

$$f(x) \geq f(x^*) + \sigma\|x - x^*\| \qquad \text{für alle } x \in \mathbb{R}^n \text{ mit } \|x - x^*\| \leq \delta$$

gibt.

Ohne Beweis (siehe J. WERNER (1992, S. 179 ff.)) wollen wir den folgenden Satz zitieren.

Satz 4.6 *Gegeben sei die diskrete, nichtlineare Approximationsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n.$$

Hierbei sei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$, $\|\cdot\|$ eine Norm auf dem $\mathbb{R}^m$ und $m \geq n$. Das Gauß-Newton-Verfahren aus Satz 4.4 liefere eine gegen ein $x^* \in \mathbb{R}^n$ konvergente Folge $\{x_k\}$. Sei x^* eine lokal stark eindeutige Lösung von (P), F stetig differenzierbar und $F'(\cdot)$ lipschitzstetig auf einer Umgebung von x^*. Dann gilt:

1. *Für alle hinreichend großen k ist $t_k = 1$.*
2. *Die Folge $\{x_k\}$ konvergiert von mindestens zweiter Ordnung gegen x^*, d. h. es existiert eine Konstante $C > 0$ derart, dass $\|x_{k+1} - x^*\| \leq C\|x_k - x^*\|^2$ für alle hinreichend großen k.*

Die entscheidende Voraussetzung in Satz 4.6 ist die der lokal starken Eindeutigkeit von x^*, des Grenzwertes einer durch das gedämpfte Gauß-Newton-Verfahren aus Satz 4.4 erzeugten Folge $\{x_k\}$. In dem folgenden Satz werden einige Aussagen im Zusammenhang mit dieser Voraussetzung gemacht.

Satz 4.7 *Gegeben sei die diskrete, nichtlineare Approximationsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n.$$

Hierbei sei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ in einem $x^* \in \mathbb{R}^n$ stetig differenzierbar, $\|\cdot\|$ eine Norm auf dem $\mathbb{R}^m$ und $m \geq n$. Dann gilt:

1. *Ist $F(x^*) = 0$ und $\operatorname{Rang}(F'(x^*)) = n$, so ist x^* eine lokal stark eindeutige Lösung von (P).*
2. *x^* ist genau dann eine lokal stark eindeutige Lösung von (P), wenn $p^* := 0$ eine (global) stark eindeutige Lösung der linearisierten Aufgabe*

   $$(\text{LP}_*) \qquad \textit{Minimiere} \quad f_*(p) := \|F(x^*) + F'(x^*)p\|, \quad p \in \mathbb{R}^n,$$

   ist.
3. *Sei $x^* \in \mathbb{R}^n$ eine lokal stark eindeutige Lösung des linearen Ausgleichsproblems*

   $$\text{(P)} \qquad \textit{Minimiere} \quad f(x) := \|F(x)\|_2, \quad x \in \mathbb{R}^n.$$

   Dann ist $F(x^*) = 0$.

Beweis: Sei $F(x^*) = 0$ und Rang $(F'(x^*)) = n$. Man definiere $\tau > 0$ durch

$$\tau := \min_{p \in \mathbb{R}^n : \|p\| = 1} \|F'(x^*)p\|$$

und wähle anschließend $\delta > 0$ so klein, dass

$$\|F(x) - \underbrace{F(x^*)}_{=0} - F'(x^*)(x - x^*)\| \leq \frac{\tau}{2}\|x - x^*\|$$

für alle $x \in \mathbb{R}^n$ mit $\|x - x^*\| \leq \delta$. Für diese x ist dann aber

$$\begin{aligned} \tau\|x - x^*\| &\leq \|F'(x^*)(x - x^*)\| \\ &\qquad \text{(Definition von } \tau) \\ &\leq f(x) + \|F(x) - F'(x^*)(x - x^*)\| \\ &\leq f(x) + \frac{\tau}{2}\|x - x^*\| \end{aligned}$$

und folglich

$$\frac{\tau}{2}\|x - x^*\| + \underbrace{f(x^*)}_{=0} \leq f(x),$$

also x^* eine lokal stark eindeutige Lösung von (P).

Sei x^* eine lokal stark eindeutige Lösung von (P), es existieren also positive Konstanten σ und δ mit $f(x) \geq f(x^*) + \sigma\|x - x^*\|$ für alle x mit $\|x - x^*\| \leq \delta$. Die Abbildung $f = g \circ F$, wobei $g(y) := \|y\|$, besitzt eine Gateaux-Variation in x^* (siehe Satz 1.8 in Abschnitt 2.1), welche durch $f'(x^*; p) = g'(F(x^*); F'(x^*)p)$ gegeben ist. Für ein beliebiges $p \in \mathbb{R}^n$ ist für alle hinreichend kleinen $t > 0$ (genauer: für alle $t > 0$ mit $t\|p\| \leq \delta$) einerseits

$$\frac{f(x^* + tp) - f(x^*)}{t} \geq \sigma\|p\|$$

und (mit $t \to 0+$) daher

$$f'(x^*; p) \geq \sigma\|p\|,$$

andererseits

$$\begin{aligned} f'(x^*; p) &= g'(F(x^*); F'(x^*)p) \\ &\leq g(F(x^*) + F'(x^*)p) - g(F(x^*)) \\ &= \|F(x^*) + F'(x^*)p\| - \|F(x^*)\|, \end{aligned}$$

insgesamt also

$$(*) \qquad \|F(x^*) + F'(x^*)p\| \geq \|F(x^*)\| + \sigma\|p\| \qquad \text{für alle } p \in \mathbb{R}^n.$$

Also ist $p^* := 0$ eine (global) stark eindeutige Lösung der linearisierten Aufgabe (LP_*). Nun zur Umkehrung dieser Aussage. Sei also $p^* := 0$ (global) stark eindeutige Lösung von (LP_*), mit einem $\sigma > 0$ gelte also $(*)$. Man bestimme ein so kleines $\delta > 0$, dass

$$\|F(x) - F(x^*) - F'(x^*)(x - x^*)\| \leq \frac{\sigma}{2}\|x - x^*\|$$

für alle x mit $\|x - x^*\| \leq \delta$. Für alle solche x ist dann wegen $(*)$ (angewandt mit $p := x - x^*$)

$$\sigma\|x - x^*\| + \|F(x^*)\| \leq \|F(x^*) + F'(x^*)(x - x^*)\| \leq \|F(x)\| + \frac{\sigma}{2}\|x - x^*\|,$$

also x^* eine lokal stark eindeutige Lösung von (P).

Nun betrachten wir ein nichtlineares Ausgleichsproblem, in (P) sei also $\|\cdot\| = \|\cdot\|_2$ die euklidische Norm, und nehmen an, x^* sei eine lokal stark eindeutige Lösung von (P). Wie wir uns im zweiten Teil dieses Satzes überlegt haben, ist dann $p^* := 0$ eine global stark eindeutige Lösung der linearisierten Aufgabe (LP_*), es existiert also ein $\sigma > 0$ mit

$$\|F(x^*) + F'(x^*)p\|_2 \geq \|F(x^*)\| + \sigma\|p\|_2 \qquad \text{für alle } p \in \mathbb{R}^n.$$

Hieraus wollen wir schließen, dass $F(x^*) = 0$. Diesen Schluss stellen wir als Aufgabe, siehe Aufgabe 3. □ □

Bemerkung: Ist $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ mit $m \geq n$ in x^* stetig differenzierbar und

$$\text{Rang } (F'(x^*)) = n,$$

so ist x^*, wie wir gerade eben gesehen haben, genau dann eine lokal stark eindeutige Lösung des nichtlinearen Ausgleichsproblems

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \|F(x)\|_2, \quad x \in \mathbb{R}^n,$$

wenn $F(x^*) = 0$. Nur in diesem Falle ist also durch Satz 4.6 gesichert, dass eine durch das gedämpfte Gauß-Newton-Verfahren erzeugte, gegen x^* konvergente Folge $\{x_k\}$ sogar quadratisch gegen x^* konvergiert. Das ist nicht verwunderlich, wenn man sich den Unterschied zwischen dem Gauß-Newton-Verfahren und dem auf

$$\text{Minimiere} \quad h(x) := \frac{1}{2}\|F(x)\|_2^2, \quad x \in \mathbb{R}^n,$$

angewandten Newton-Verfahren genauer ansieht. Unter geeigneten Rangvoraussetzungen lautet ersteres

$$x_{k+1} := x_k - t_k[F'(x_k)^T F'(x_k)]^{-1} F'(x_k)^T F(x_k),$$

während letzteres (in der ungedämpften Form) durch

$$x_{k+1} := x_k - [F'(x_k)^T F'(x_k) + S(x_k)]^{-1} F'(x_k)^T F(x_k)$$

gegeben ist, wobei

$$S(x) := \sum_{i=1}^{m} F_i(x)\nabla^2 F_i(x)$$

gesetzt ist. Nur für $S(x^*) = 0$ wird man daher hoffen können, dass sich die lokale quadratische Konvergenz des Newton-Verfahrens auf das Gauß-Newton-Verfahren überträgt. Ferner wird man erwarten, dass für "kleines" $S(x^*)$ (wenn also F nur "schwach

nichtlinear" oder $F(x^*)$ "klein" ist) die lokale Konvergenz des Gauß-Newton-Verfahrens befriedigend sein wird, weil es dann sozusagen nicht weit entfernt vom Newton-Verfahren ist. □

Anders als bei nichtlinearen Ausgleichsproblemen sind die Verhältnisse beim diskreten, nichtlinearen Tschebyscheffschen Approximationsproblem. Hier kann auch für $F(x^*) \neq 0$ in x^* eine lokal stark eindeutige Lösung vorliegen. Dies wird im folgenden Satz präzisiert.

Satz 4.8 *Gegeben sei die diskrete, nichtlineare Tschebyscheffsche Approximationsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) := \|F(x)\|_\infty, \quad x \in \mathbb{R}^n.$$

Sei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ mit $m \geq n$ in $x^* \in \mathbb{R}^n$ stetig differenzierbar, x^* eine stationäre Lösung von (P) und die sogenannte Haarsche Bedingung in x^* erfüllt, d. h. jede $n \times n$-Untermatrix von $F'(x^*)$ sei nichtsingulär. Dann ist x^* eine lokal stark eindeutige Lösung von (P).

Beweis: O. B. d. A. kann $F(x^*) \neq 0$ angenommen werden (andernfalls wende man den ersten Teil von Satz 4.7 an). Definiert man zu der stationären Lösung x^* von (P) die Indexmenge

$$I(x^*) := \{i \in \{1, \ldots, m\} : |F_i(x^*)| = \|F(x^*)\|_\infty\},$$

so sagt Satz 1.9, dass reelle Zahlen λ_i^*, $i \in I(x^*)$, existieren mit

$$\lambda_i^* \geq 0 \quad (i \in I(x^*)), \qquad \sum_{i \in I(x^*)} \lambda_i^* = 1, \qquad \sum_{i \in I(x^*)} \lambda_i^* \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*) = 0.$$

Mit anderen Worten liegt der Nullvektor 0 des $\mathbb{R}^n$ in der *konvexen Hülle* der Vektoren $\{\operatorname{sign}(F_i(x^*))\nabla F_i(x^*) : i \in I(x^*)\}$, er lässt sich also als eine *Konvexkombination* dieser Vektoren darstellen. Nun wenden wir einen bekannten Satz von Carathéodory an (mit den Hinweisen in Aufgabe 5 wird man diesen selbst beweisen können):

- *Sei $S \subset \mathbb{R}^n$ und*

$$\operatorname{co}(S) := \Big\{ \sum_{i=1}^m \lambda_i x_i : x_i \in S,\ \lambda_i \geq 0\ (i = 1, \ldots, m),\ \sum_{i=1}^m \lambda_i = 1,\ m \in \mathbb{N} \Big\}$$

  die Menge aller Konvexkombinationen von Punkten aus S. Dann lässt sich jedes $x \in \operatorname{co}(S)$ als Konvexkombination von höchstens $n+1$ Punkten aus S darstellen. Zu jedem $x \in \operatorname{co}(S)$ existiert also ein $m \in \mathbb{N}$ mit $m \leq n+1$ sowie $\mu_i \geq 0$, $x_i \in S$, $i = 1, \ldots, m$, mit $\sum_{i=1}^m \mu_i = 1$ und $x = \sum_{i=1}^m \mu_i x_i$.

Daher existieren eine Indexmenge $I^* \subset I(x^*)$ mit höchstens $n+1$ Elementen und reelle Zahlen μ_i^*, $i \in I^*$, mit

$$\mu_i^* > 0 \quad (i \in I^*), \qquad \sum_{i \in I^*} \mu_i^* = 1, \qquad \sum_{i \in I^*} \mu_i^* \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*) = 0.$$

Hätte I^* *weniger* als $n+1$ Elemente, so ließe sich der Nullvektor des $\mathbb{R}^n$ als nichttriviale Linearkombination von n Zeilen von $F'(x^*)$ darstellen, was einen Widerspruch dazu bedeutet, dass jede $n \times n$-Untermatrix von $F'(x^*)$ nichtsingulär ist. Daher enthält I^* *genau* $n+1$ Elemente. Das Ziel besteht darin, die Existenz einer Konstanten $\sigma > 0$ mit

$$(*) \qquad \sigma\|p\|_\infty + \|F(x^*)\|_\infty \le \|F(x^*) + F'(x^*)p\|_\infty \qquad \text{für alle } p \in \mathbb{R}^n$$

zu zeigen, woraus wegen des zweiten Teils von Satz 4.7 die Behauptung folgt.

Sei $q \in \mathbb{R}^n \setminus \{0\}$. Dann ist

$$\max_{i \in I^*} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)^T q > 0,$$

denn andernfalls wäre

$$0 \ge \sum_{i \in I^*} \underbrace{\mu_i^*}_{>0} \underbrace{\operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)^T q}_{\le 0} = \underbrace{\Big(\sum_{i \in I^*} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*) \Big)^T}_{=0} q = 0,$$

so dass $q \neq 0$ auf den $n+1$ Vektoren $\nabla F_i(x^*)$, $i \in I^*$, von denen je n linear unabhängig sind, senkrecht stehen würde, was einen Widerspruch bedeutet. Daher ist

$$0 < \sigma := \min_{\|q\|_\infty = 1} \max_{i \in I^*} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)^T q.$$

Wir wollen zeigen, dass $(*)$ gilt. Für $p = 0$ ist das trivialerweise der Fall, so dass wir $p \neq 0$ annehmen können. Zu $q := p/\|p\|_\infty$ existiert nach Definition von σ ein $k \in I^*$ mit $\sigma \le \operatorname{sign}(F_k(x^*)) \nabla F_k(x^*)^T q$. Dann ist

$$\begin{aligned}
\|F(x^*) + F'(x^*)p\|_\infty &\ge |F_k(x^*) + \nabla F_k(x^*)^T p| \\
&= |\operatorname{sign}(F_k(x^*))[F_k(x^*) + \nabla F_k(x^*)^T p]| \\
&= |\,|F_k(x^*)| + \operatorname{sign}(F_k(x^*)) \nabla F_k(x^*)^T p\,| \\
&= \|F(x^*)\|_\infty + \operatorname{sign}(F_k(x^*)) \nabla F_k(x^*)^T p \\
&\ge \|F(x^*)\|_\infty + \sigma\|p\|_\infty,
\end{aligned}$$

womit die Gültigkeit von $(*)$ und schließlich die lokal starke Eindeutigkeit von x^* bewiesen ist. □ □

Beispiel: Wir kommen auf ein Beispiel aus Kapitel 2 zurück. Dort hatten wir das diskrete Tschebyscheffsche Approximationsproblem

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \max_{i=1,2} |F_i(x)|, \quad x \in \mathbb{R}^2$$

betrachtet, wobei

$$\begin{aligned}
F_1(x) &:= x_1 - x_2^3 + 5x_2^2 - 2x_2 - 13, \\
F_2(x) &:= x_1 + x_2^3 + x_2^2 - 14x_2 - 29.
\end{aligned}$$

$x^* = (5, 4)$ ist eine globale Lösung mit $F(x^*) = 0$. Da

$$F'(x^*) = \begin{pmatrix} 1 & -10 \\ 1 & 42 \end{pmatrix}$$

nichtsingulär ist, ist x^* lokal stark eindeutige Lösung. Eine isolierte lokale Lösung ist aber auch $x^* = (x_1^*, x_2^*)$ mit

$$x_1^* := \frac{1}{3}(53 - 4\sqrt{22}), \qquad x_2^* := \frac{1}{3}(2 - \sqrt{22}),$$

wie wir durch Nachprüfen der hinreichenden Bedingungen zweiter Ordnung gesehen haben. Da aber $\nabla F_1(x^*) = \nabla F_2(x^*)$, ist $F'(x^*)$ singulär. Nun ist Rang $(F'(x^*)) = n$ eine *notwendige* Bedingung für lokale starke Eindeutigkeit. Daher ist obige lokale Lösung nicht lokal stark eindeutig und es ist kein Wunder, dass das Gauß-Newton-Verfahren bei diesem Beispiel Schwierigkeiten hat. □

3.4.5 Aufgaben

1. Man beweise Satz 4.3, also die folgende Aussage: Gegeben sei die diskrete, nichtlineare Approximationsaufgabe

   (P) $$\text{Minimiere} \quad f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n,$$

   die Zielfunktion f genüge den Voraussetzungen (V) (a)–(c). Sei $x_c \in L_0$ und $p \in \mathbb{R}^n$ eine Richtung mit $f_c(p) := \|F(x_c) + F'(x_c)p\| < f(x_c)$. Seien $\alpha \in (0, \frac{1}{2})$, $0 < l \le u < 1$ gegeben und $t := \rho_j$ eine zugehörige Armijo-Schrittweite. Dann existiert eine Konstante $\theta > 0$, die nur von α, γ sowie l und u, nicht aber von x_c oder p abhängt, mit

   $$f(x_c) - f(x_c + tp) \ge \theta \left[f(x_c) - f_c(p), \left(\frac{f(x_c) - f_c(p)}{\|p\|} \right)^2 \right].$$

2. Gegeben sei das nichtlineare Ausgleichsproblem (siehe P. Spellucci (1993, S. 199))

   $$\text{Minimiere} \quad f(x) := \|F(x)\|_2, \quad x \in \mathbb{R}^4,$$

   wobei $F\colon \mathbb{R}^4 \longrightarrow \mathbb{R}^{11}$ definiert ist durch

   $$F_i(x) := \frac{x_1 + x_2 t_i}{1 + x_3 t_i + x_4 t_i^2} - y_i, \qquad i = 1, \ldots, 11,$$

   mit den Daten

i	t_i	y_i
1	4.0000	0.0489250
2	2.0000	0.0973500
3	1.0000	0.1735000
4	0.5000	0.3200000
5	0.2500	0.3376000
6	0.1670	0.3754491
7	0.1250	0.3648000
8	0.1000	0.3420000
9	0.0823	0.3924666
10	0.0714	0.3291317
11	0.0625	0.3936000

   Man berechne eine Lösung mit Hilfe des gedämpften Gauß-Newton-Verfahrens.

3. Sei $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$. Es existiere ein $\sigma > 0$ mit

$$\sigma\|p\|_2 + \|b\|_2 \le \|b + Ap\|_2 \qquad \text{für alle } p \in \mathbb{R}^n.$$

Man zeige, dass dann Rang $(A) = n$ und $b = 0$.

4. Sei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ mit $m \ge n$ eine stetig differenzierbare Abbildung. Die Abbildungen $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ seien definiert durch

$$f(x) := \|F(x)\|_2, \qquad h(x) := \frac{1}{2}\|F(x)\|_2^2.$$

Bei gegebenem $x_c \in \mathbb{R}^n$ sei $p^* \in \mathbb{R}^n$ eine Lösung des in x_c linearisierten Ausgleichsproblems

$$(\mathrm{LP}_c) \qquad \text{Minimiere} \quad f_c(p) := \|F(x_c) + F'(x_c)p\|_2, \quad p \in \mathbb{R}^n.$$

Man zeige:

(a) Es ist $\nabla h(x_c)^T p^* = f_c(p^*)^2 - f(x_c)^2$.
Hinweis: Es gilt die Normalgleichung $F'(x_c)^T[F(x_c) + F'(x_c)p^*] = 0$.

(b) Ist $f_c(p^*) < f(x_c)$, ist also p^* eine Abstiegsrichtung (für f und h) in x_c, und sind $\alpha, \rho \in (0,1)$, so impliziert

$$h(x_c + \rho p^*) \le h(x_c) + \alpha\rho\nabla h(x_c)^T p^*,$$

dass

$$f(x_c + \rho p^*) \le f(x_c) + \alpha\rho[f_c(p^*) - f(x_c)].$$

5. Man beweise den Satz von Carathéodory: Sei $S \subset \mathbb{R}^n$ und

$$\mathrm{co}(S) := \left\{ \sum_{i=1}^m \lambda_i x_i : x_i \in S,\ \lambda_i \ge 0\ (i = 1,\dots,m),\ \sum_{i=1}^m \lambda_i = 1,\ m \in \mathbb{N} \right\}$$

die Menge aller Konvexkombinationen von Punkten aus S. Dann lässt sich jedes $x \in \mathrm{co}(S)$ als Konvexkombination von höchstens $n+1$ Punkten aus S darstellen. Zu jedem $x \in \mathrm{co}(S)$ existiert also ein $m \in \mathbb{N}$ mit $m \le n+1$ sowie $\mu_i \ge 0$, $x_i \in S$, $i = 1,\dots,m$, mit $\sum_{i=1}^m \mu_i = 1$ und $x = \sum_{i=1}^m \mu_i x_i$.

Hinweis: Sei $x \in \mathrm{co}(S)$ Konvexkombination von m Punkten $x_1,\dots,x_m$ aus S. Man zeige: Ist $m > n+1$, so ist x auch Konvexkombination von $m-1$ Punkten aus S (nach endlich vielen Schritten hat man dann die Behauptung bewiesen). Hierzu benutze man, dass $n+1$ (und mehr) Vektoren des $\mathbb{R}^n$, etwa $\{x_1 - x_m, \dots, x_{m-1} - x_m\}$, linear abhängig sind und folglich der Nullvektor des $\mathbb{R}^n$ sich als nichttriviale Linearkombination von $x_1,\dots,x_m$ darstellen lässt.

Kapitel 4

Trust-Region-Verfahren

In Kapitel 3 haben wir Schrittweitenverfahren zur Lösung unrestringierter Optimierungsaufgaben analysiert. Diese bestanden aus einer Schrittweiten- und einer Richtungsstrategie. Insbesondere eine vernünftige Implementation der Schrittweitenstrategie ist nicht ganz einfach. In diesem Kapitel untersuchen wir Trust-Region-Verfahren[1], die insbesondere bei hochdimensionalen Problemen den Schrittweitenverfahren häufig überlegen sind.

4.1 Ein Modellalgorithmus

Gegeben sei wieder die unrestringierte Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n.$$

Die Idee bei den Trust-Region besteht darin, die Zielfunktion f lokal auf einer Kugel (bezüglich einer geeigneten Norm) um eine aktuelle Näherung durch ein einfacheres "Modell" zu ersetzen, etwa einer linearen oder quadratischen Approximation der Zielfunktion. Dann bestimmt man ein Minimum (oder auch nur eine Approximation an das Minimum) des Modells bzw. der vereinfachten Zielfunktion auf der Kugel. Wird eine Verminderung des Zielfunktionswertes entweder nicht erreicht, oder ist diese eher enttäuschend gering, so hat man dem Modell auf einer zu großen Kugel um die aktuelle Näherung "vertraut", diese wird daher verkleinert und auf dieser kleineren Kugel erneut ein Minimum der Modellfunktion bestimmt. Andernfalls wird dieses Minimum als neue aktuelle Näherung akzeptiert und der Radis der Kugel wird vergrößert, wenn ein verschärfter Test auf hinreichende Verminderung erfolgreich bestanden wird. Das ist, sehr lax ausgedrückt, die Idee der Trust-Region-Verfahren.

Nun soll diese Idee etwas genauer gefasst werden. In den folgenden beiden Abschnitten werden wir dann auf glatte (unrestringierte) Optimierungsaufgaben sowie auf diskrete Approximationsaufgaben eingehen.

[1] Als Lehrbuch, welches bisher noch nicht genannt wurde, sei hier vor allem auf

A. R. Conn, N. I. M. Gould, Ph. L. Toint (2000) *Trust-Region Methods*. SIAM-MPS, Philadelphia

hingewiesen.

Bei gegebener aktueller Näherung $x_c \in \mathbb{R}^n$ sei ein "einfaches Modell" $f_c\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ für die i. Allg. komplizierte Funktion $p \mapsto f(x_c + p)$ gegeben. Eine Minimalforderung an die Modellfunktion f_c wird $f_c(0) = f(x_c)$ sein. Ist z. B. f in x_c zweimal stetig differenzierbar, so liegt es nahe,

$$f_c(p) := f(x_c) + \nabla f(x_c)^T p + \frac{1}{2} p^T \nabla^2 f(x_c) p$$

zu setzen, wonei $\nabla^2 f(x_c)$ durch eine symmetrische (nicht notwendig positiv definite) Matrix $B_c \in \mathbb{R}^{n\times n}$ ersetzt sein kann. Ist dagegen $f(x) := \|F(x)\|$ mit einer stetig differenzierbaren Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ und einer Norm $\|\cdot\|$ auf dem $\mathbb{R}^m$, so könnte die Modellfunktion f_c durch

$$f_c(p) := \|F(x_c) + F'(x_c)p\|$$

gegeben sein. Entsprechendes gilt für andere "halbglatte" Zielfunktionen $f = g \circ F$ mit einer stetig differenzierbaren Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ und einer konvexen Funktion $g \circ \mathbb{R}^m \longrightarrow \mathbb{R}$.

Ein Schritt eines Modellalgorithmus für Trust-Region-Verfahren könnte dann folgendermaßen aussehen:

- Unabhängig vom aktuellen Iterationsschritt seien Konstanten $0 < \rho_1 < \rho_2 < 1$, $\sigma_1 \in (0, 1)$ und $\sigma_2 > 1$ gegeben. Z. B. sei $\rho_1 := 0.01$, $\rho_2 := 0.9$, $\sigma_1 := 0.5$ und $\sigma_2 := 2$.

- Gegeben sei ein aktuelles Paar (x_c, Δ_c), wobei x_c eine Näherung für eine (stationäre, lokale, globale) Lösung von (P) ist und $\Delta_c > 0$ der Radius einer Kugel um 0 bzw. x_c ist, die bezüglich einer Norm $\|\cdot\|_c$ zu verstehen ist. Diese Norm ist gewöhnlich die euklidische Norm oder die Maximumnorm, welche noch abhängig von dem aktuellen Iterationsschritt mit Gewichten versehen sein kann. Mit einer positiv definiten Diagonalmatrix $D_c \in \mathbb{R}^{n\times n}$ könnte daher $\|\cdot\|_c$ durch $\|p\|_c := \|D_c p\|_2$ bzw. $\|p\|_c := \|D_c p\|_\infty$ gegeben sein.

- Bestimme eine globale Lösung $p^* \in \mathbb{R}^n$ der Aufgabe

  $$(\mathrm{P}_c) \qquad \text{Minimiere} \quad f_c(p), \quad \|p\|_c \le \Delta_c.$$

  Natürlich sollte dieses Trust-Region-Hilfsproblem "einfach" lösbar sein.

- Falls $f(x_c) = f_c(p^*)$ (dies ist genau dann der Fall, wenn 0 eine Lösung von (P_c) ist), dann: STOP.

  Ist die Modellfunktion richtig gewählt, so wird x_c in diesem Falle wenigstens eine stationäre Lösung von (P) sein.

- Andernfalls berechne

  $$r_c := \frac{f(x_c) - f(x_c + p^*)}{f(x_c) - f_c(p^*)}.$$

  Falls $r_c \ge \rho_1$, dann setze $x_+ := x_c + p^*$ als neue Näherung und bezeichne den Iterationsschritt als erfolgreich. Andernfalls setze $x_+ := x_c$.

- Falls $r_c < \rho_1$, dann wähle $\Delta_+ \in (0, \sigma_1 \Delta_c]$.
- Falls $r_c \in [\rho_1, \rho_2)$, dann wähle $\Delta_+ \in [\sigma_1 \Delta_c, \Delta_c]$.
- Falls $r_c \geq \rho_2$, dann wähle $\Delta_+ \in [\Delta_c, \sigma_2 \Delta_c]$.

Einige Bemerkungen zu den Tests im letzten Schritt sind angebracht. In ihm wird angenommen, dass $f(x_c) \neq f_c(p^*)$ (andernfalls wäre ein Ausstieg im vorherigen Schritt erfolgt). Dann ist aber $f(x_c) > f_c(p^*)$, da von der Modellfunktion $f_c(0) = f(x_c)$ angenommen wurde. Entscheidend für den Test ist die Größe r_c, der Quotient aus der tatächlichen und der durch das Modell vorhergesagten Verminderung des Zielfunktionswertes. Je näher r_c bei 1 liegt, desto besser "stimmt" das Modell.

Ist $r_c < \rho_1$, so hat sich keine Verminderung eingestellt oder diese ist, verglichen mit der vorhergesagten, zu gering. Das wird darauf zurückgeführt, dass dem Modell auf einer zu großen Kugel vertraut wurde. Diese wird daher entsprechend verkleinert und mit derselben aktuellen Näherung ein erneuter Versuch unternommen.

Ist sogar der verschärfte Test $r_c \geq \rho_2$ erfolgreich, so stimmen die tatsächliche und die vorhergesagte Verminderung hinreichend gut überein, so dass im nächsten Schritt dem Modell auf einer i. Allg. größeren Kugel vertraut wird. Für $r_c \in [\rho_1, \rho_2)$ ist man mit der neuen Näherung zufrieden, vergrößert den Bereich aber nicht. In beiden Fällen ist man aber erfolgreich und berechnet eine neue aktuelle Näherung.

Beispiel: Ist f in x_c stetig differenzierbar, so ist $f_c(p) := f(x_c) + \nabla f(x_c)^T p$ sicher das einfachste Modell für die Abbildung $p \mapsto f(x+p)$. Nimmt man im Hilfsproblem (P_c) als Norm $\|\cdot\|_c$ unabhängig vom Iterationsschritt die euklidische Norm, so lautet das entsprechende Trust-Region-Hilfsproblem

$$(\mathrm{P}_c) \qquad \text{Minimiere} \quad f_c(p) := f(x_c) + \nabla f(x_c)^T p, \quad \|p\|_2 \leq \Delta_c.$$

Ist $\nabla f(x_c) = 0$ (dies ist äquivalent dazu, dass $p^* := 0$ eine Lösung von (P_c) ist bzw. $\min(\mathrm{P}_c) = f(x_c)$ gilt), so ist x_c eine stationäre Lösung von (P). Andernfalls ist

$$p^* := -\Delta_c \frac{\nabla f(x_c)}{\|\nabla f(x_c)\|_2}$$

die Lösung von (P_c). Also ist p^* bis auf den positiven Faktor Δ_c die negative, normierte Gradientenrichtung. □

Bemerkung: Bei unrestringierten Optimierungsaufgaben, und mit diesen beschäftigen wir uns ja, wird man als Norm im Hilfsproblem eigentlich immer die (gewichtete) euklidische Norm nehmen. Wenn lineare Nebenbedingungen auftreten würden, könnte es geschickter sein, etwa die Maximumnorm zu wählen, weil man dann insgesamt lineare Nebenbedingungen hat. □

4.1.1 Aufgaben

1. Sei $f \in \mathbb{R}$, $g \in \mathbb{R}^n \setminus \{0\}$ und $\Delta > 0$. Man gebe eine Lösung von

   $$(\mathrm{P}) \qquad \text{Minimiere} \quad \phi(p) := f + g^T p, \quad \|p\|_\infty \leq \Delta$$

   an und begründe dies.

2. Man betrachte die unrestringierte Optimierungsaufgabe

   (P) $$\text{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T B p, \quad p \in \mathbb{R}^n,$$

   wobei $f \in \mathbb{R}$, $g \in \mathbb{R}^n$ und $B \in \mathbb{R}^{n \times n}$ eine symmetrische Matrix ist. Man zeige:

   (a) (P) besitzt genau dann eine Lösung, wenn B positiv semidefinit und $g \in \text{Bild}\,(B)$ ist.

   (b) (P) besitzt genau dann eine eindeutige Lösung, wenn B positiv definit ist.

3. Sei[2] $f(x) := 10(x_2 - x_1^2)^2 + (1 - x_1)^2$. Für $x_c := (0, -1)^T$ mache man einen Contour-Plot des quadratischen Modells

   $$f_c(p) := f(x_c) + \nabla f(x_c)^T p + \frac{1}{2} p^T \nabla^2 f(x_c) p.$$

   Man zeichne in den Contour-Plot ferner noch Kreise um $(0, 0)$ mit dem Radius 0.5, 1 und 2. Man wiederhole das ganze mit $x_c = (0, 0.5)^T$.

4. Sei p^* eine Lösung von

   (P) $$\text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2} p^T B p, \quad \|p\|_\infty \leq \Delta.$$

   Hierbei sind $g \in \mathbb{R}^n$, die symmetrische Matrix $B \in \mathbb{R}^{n \times n}$ und $\Delta > 0$ gegeben. Man zeige:

   (a) Ist $-\Delta < p_j^* < \Delta$, so ist $(g + Bp^*)_j = 0$.

   (b) Ist $p_j^* = \Delta$, so ist $(g + Bp^*)_j \leq 0$.

   (c) Ist $p_j^* = -\Delta$, so ist $(g + Bp^*)_j \geq 0$.

5. Sei

   $$g := \begin{pmatrix} -2 \\ -20 \end{pmatrix}, \qquad B := \begin{pmatrix} 42 & 0 \\ 0 & 20 \end{pmatrix}.$$

   Für $\Delta := \frac{1}{2}$, 1, 2 berechne man eine Lösung von

   (P) $$\text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2} p^T B p, \quad \|p\|_\infty \leq \Delta.$$

4.2 Trust-Region-Verfahren bei glatter Zielfunktion

In diesem Abschnitt betrachten wir nach wie vor die unrestringierte Optimierungsaufgabe

(P) $$\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n,$$

wobei wir voraussetzen werden, dass die Zielfunktion f zweimal stetig differenzierbar ist. Als Modellfunktion in einer aktuellen Näherung x_c werden wir stets eine quadratische Approximation verwenden, also

$$f_c(p) := f(x_c) + g_c^T p + \frac{1}{2} p^T B_c p.$$

[2] Diese Aufgabe findet man bei J. Nocedal, S. J. Wright (1999, S. 97).

Hierbei ist stets $g_c := \nabla f(x_c)$, während die symmetrische Matrix B_c eine Approximation an die Hessesche $\nabla^2 f(x_c)$ (oder mit ihr übereinstimmt). Wichtig ist, dass B_c nicht als positiv semidefinit vorausgesetzt wird, so dass f_c nicht notwendig konvex ist.

4.2.1 Das Trust-Region-Hilfsproblem

Die wesentliche Arbeit im Modellalgorithmus für Trust-Region-Verfahren erfolgt bei der Lösung des Hilfsproblems, das (in unserem Falle: quadratische) Modell auf einer Kugel zu minimieren. Daher beschäftigen wir uns in diesem Unterabschnitt mit der numerischen Lösung des Problems

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T B p, \quad \|p\|_2 \leq \Delta,$$

wobei $f \in \mathbb{R}$, $g \in \mathbb{R}^n$, die symmetrische Matrix $B \in \mathbb{R}^{n \times n}$ und $\Delta > 0$ gegeben sind. Wir verzichten in diesem Unterabschnitt auf den Index $_c$, da wir sozusagen ein festes Problem lösen. Sind die Variablen durch eine nichtsinguläre Diagonalmatrix $D \in \mathbb{R}^{n \times n}$ skaliert, ist also $\phi(\cdot)$ unter der Restriktion $\|Dp\|_2 \leq \Delta$ zu lösen, so mache man die Variablentransformation $q = Dp$, gewinnt q^* als Lösung von

$$\text{Minimiere} \quad \psi(q) := f + (D^{-1}g)^T q + \frac{1}{2} q^T D^{-1} B D^{-1} q, \quad \|q\|_2 \leq \Delta$$

und berechnet anschließend die Lösung $p^* = D^{-1}q^*$ von (P). Das ist der Grund, weshalb wir in diesem Unterabschnitt keine Skalierung der Variablen betrachten.

Ganz entscheidend für die numerische Lösung von (P) ist, dass man eine globale Lösung von (P) charakterisieren kann, also notwendige und hinreichende Bedingungen dafür angeben kann, dass ein $p^* \in \mathbb{R}^n$ mit $\|p^*\|_2 \leq \Delta$ eine globale Lösung von (P) ist. Da wir *nicht* voraussetzen, dass B positiv semidefinit ist, es sich bei (P) also um eine konvexe Optimierungsaufgabe (konvexe Zielfunktion ist auf konvexer Menge zu minimieren) handelt, ist dies ein durchaus bemerkenswertes Ergebnis.

Satz 2.1 *Genau dann ist ein $p^* \in \mathbb{R}^n$ mit $\|p^*\|_2 \leq \Delta$ eine globale Lösung von*

$$\text{(P)} \qquad \textit{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T B p, \quad \|p\|_2 \leq \Delta,$$

wenn ein $\lambda^* \geq 0$ mit

(a) $(B + \lambda^* I)p^* = -g$,

(b) $\lambda^*(\Delta - \|p^*\|_2) = 0$,

(c) $B + \lambda^* I$ *ist positiv semidefinit*

existiert. Darüberhinaus ist p^* eindeutige globale Lösung von (P), wenn $B + \lambda^* I$ sogar positiv definit ist.

Beweis: Wir zeigen zunächst die einfache Richtung, dass nämlich die Existenz eines $\lambda^* \geq 0$ mit (a)–(c) eine hinreichende Optimalitätsbedingung ist. Seien also $p^* \in \mathbb{R}^n$ mit $\|p^*\|_2 \leq \Delta$ und ein $\lambda^* \geq 0$ mit (a)–(c) gegeben. Für ein beliebiges $p \in \mathbb{R}^n$ mit $\|p\|_2 \leq \Delta$ ist

$$\begin{aligned}
\phi(p) - \phi(p^*) &= (\underbrace{g + Bp^*}_{-\lambda^* p^*})^T (p - p^*) + \frac{1}{2}(p - p^*)^T B (p - p^*) \\
&= -\lambda^* (p^*)^T (p - p^*) + \frac{1}{2} \underbrace{(p - p^*)^T (B + \lambda^* I)(p - p^*)}_{\geq 0} - \frac{\lambda^*}{2} \|p - p^*\|_2^2 \\
&\geq -\lambda^* (p^*)^T (p - p^*) - \frac{\lambda^*}{2} \|p - p^*\|_2^2 \\
&= \frac{\lambda^*}{2} (\|p^*\|_2^2 - \|p\|_2^2) \\
&= \frac{\lambda^*}{2} (\Delta^2 - \|p\|_2^2) \\
&\geq 0,
\end{aligned}$$

und daher p^* eine globale Lösung von (P). Ist $B + \lambda^* I$ sogar positiv definit, so entnimmt man der obigen Gleichungs-Ungleichungskette, dass $p = p^*$ aus $\phi(p) = \phi(p^*)$ folgt, und das bedeutet die Eindeutigkeit der globalen Lösung.

Nun kommen wir zu der tiefer liegenden Richtung, dass nämlich die Existenz eines $\lambda^* \geq 0$ mit (a)–(c) eine notwendige Optimalitätsbedingung ist. Wir nehmen also an, p^* sei eine globale Lösung von (P) (und damit p^* natürlich zulässig, also $\|p^*\|_2 \leq \Delta$). Ist $\|p^*\|_2 < \Delta$, so ist bei p^* ein unrestringiertes Minimum von ϕ und daher notwendigerweise $\nabla \phi(p^*) = g + Bp^* = 0$ und $\nabla^2 \phi(p^*) = B$ positiv semidefinit. Mit $\lambda^* := 0$ sind also (a)–(c) erfüllt. Wir können daher annehmen, dass p^* eine globale Lösung von (P) mit $\|p^*\|_2 = \Delta$ ist. Zunächst nutzen wir nur aus, dass p^* als globale Lösung auch eine lokale Lösung von (P) ist und beweisen:

- Sei p^* eine lokale Lösung von (P) mit $\|p^*\|_2 = \Delta$. Dann existiert ein $\lambda^* \geq 0$ mit $(B + \lambda^* I)p^* = -g$ und der Eigenschaft, dass $B + \lambda^* I$ auf dem orthogonalen Komplement von span $\{p^*\}$ positiv semidefinit ist, also $h^T (B + \lambda^* I) h \geq 0$ für alle $h \in \mathbb{R}^n$ mit $(p^*)^T h = 0$ gilt.

Denn: Zunächst zeigen wir, dass es ein $\lambda^* \geq 0$ mit $(B + \lambda^* I)p^* = -g$ gibt. Angenommen, dies wäre nicht der Fall, es würde also kein $\lambda^* \geq 0$ mit $\lambda^* p^* = -(g + Bp^*)$ geben. Dann ist

$$-(p^*)^T (g + Bp^*) < \underbrace{\|p^*\|_2}_{=\Delta} \|g + Bp^*\|_2.$$

Man setze

$$q := -\|p^*\|_2 (g + Bp^*) - \|g + Bp^*\|_2 p^*.$$

Dann ist

$$(p^*)^T q = -\|p^*\|_2 (p^*)^T (g + Bp^*) - \|g + Bp^*\|_2 \|p^*\|_2^2$$

$$\begin{aligned} &= \underbrace{\|p^*\|_2}_{>0}\underbrace{[-(p^*)^T(g+Bp^*) - \|g+Bp^*\|_2\,\|p^*\|_2]}_{<0} \\ &< 0, \end{aligned}$$

und

$$\begin{aligned} (g+Bp^*)^T q &= -\|p^*\|_2\,\|g+Bp^*\|_2^2 - \|g+Bp^*\|_2 (p^*)^T(g+Bp^*) \\ &= \underbrace{\|g+Bp^*\|_2}_{>0}\underbrace{[-\|p^*\|_2\,\|g+Bp^*\|_2 - (p^*)^T(g+Bp^*)]}_{<0} \\ &< 0. \end{aligned}$$

Daher ist q wegen $\nabla\phi(p^*)^T q < 0$ eine Abstiegsrichtung für ϕ in p^*, anderseits ist $\|p^* + tq\|_2 < \|p^*\|_2 = \Delta$ für alle hinreichend kleinen $|t|$. Beides zusammen ergibt einen Widerspruch zur Optimalität von p^*. Damit ist die Existenz eines $\lambda^* \geq 0$ mit $(B + \lambda^* I)p^* = -g$ nachgewiesen. Nun sei $h \in \mathbb{R}^n$ ein Vektor mit $(p^*)^T h = 0$. Dann ist $p(t) \in \mathbb{R}^n$ durch

$$p(t) := \Delta \frac{p^* + th}{\|p^* + th\|_2} = \Delta \frac{p^* + th}{\sqrt{\Delta^2 + t^2\|h\|_2^2}}$$

für alle hinreichend kleinen $|t|$ definiert. Die Funktion $\psi(t) := \phi(p(t))$ nimmt bei $t = 0$ ein lokales Minimum an. Daher ist $\psi'(0) = 0$ und $\psi''(0) \geq 0$. Nun ist

$$\psi'(t) = \nabla\phi(p(t))^T p'(t), \qquad \psi''(t) = \nabla\phi(p(t))^T p''(t) + p'(t)^T \nabla^2\phi(p(t))p'(t).$$

Weiter ist

$$p'(0) = h, \qquad p''(0) = -\frac{\|h\|^2}{\Delta^2} p^*.$$

Für diese etwas mühsame Rechnung haben wir Maple benutzt. Daher ist

$$\begin{aligned} 0 &\leq \psi''(0) \\ &= \nabla\phi(p(0))^T p''(0) + p'(0)^T \nabla^2\phi(p(0)p'(0) \\ &= -\frac{\|h\|_2^2}{\Delta^2}(g+Bp^*)^T p^* + h^T B h \\ &= h^T(B + \lambda^* I)h. \end{aligned}$$

Damit ist die obige Zwischenbehauptung bewiesen. Man beachte, dass bisher lediglich benutzt wurde, dass p^* eine *lokale* Lösung von (P) ist.

Nun zeigen wir, dass $B + \lambda^* I$ (auf dem ganzen $\mathbb{R}^n$) positiv semidefinit ist. Hierzu genügt es, den Fall $\|p^*\|_2 = \Delta$ zu betrachten und nachzuweisen, dass $h^T(B + \lambda^* I)h \geq 0$ für alle $h \in \mathbb{R}^n$ mit $(p^*)^T h \neq 0$. Ein solches h sei gegeben. Mit

$$t := -\frac{2(p^*)^T h}{\|h\|_2^2}$$

ist dann $t \neq 0$ und nach Konstruktion $\|p^* + th\|_2 = \|p^*\|_2$. Zur Abkürzung setze man $p := p^* + th$. Da p^* eine *globale* Lösung von (P) ist, erhalten wir

$$0 \leq \phi(p) - \phi(p^*)$$

$$
\begin{aligned}
&= (\underbrace{g + Bp^*}_{=-\lambda^* p^*})^T (\underbrace{p - p^*}_{=th}) + \frac{1}{2} t^2 h^T B h \\
&= -t\lambda^* (p^*)^T h + \frac{1}{2} t^2 h^T B h \\
&= \frac{1}{2} \lambda^* t^2 \|h\|_2^2 + \frac{1}{2} t^2 h^T B h \\
&= \frac{t^2}{2} h^T (B + \lambda^* I) h.
\end{aligned}
$$

Damit ist der Satz schließlich bewiesen. □ □

Nun kommen wir zur numerischen Berechnung einer oder der Lösung p^* von (P). Hierzu benutzen wir die notwendige und hinreichende Optimalitätsbedingung in Satz 2.1 und führen diese Berechnung auf die Lösung einer nichtlinearen Gleichung in einer Unbekannten zurück.

Seien

$$\lambda_1 = \cdots = \lambda_r < \lambda_{r+1} \leq \cdots \leq \lambda_n$$

die Eigenwerte der symmetrischen Matrix $B \in \mathbb{R}^{n \times n}$. Wir suchen nach einem $\lambda^* \geq \max(0, -\lambda_1)$ (da nur dann $\lambda^* \geq 0$ und $B + \lambda^* I$ positiv semidefinit ist) und einem $p^* \in \mathbb{R}^n$ mit $(B + \lambda^* I)p^* = -g$ und $\lambda^*(\|p^*\|_2 - \Delta) = 0$. Wir definieren die Funktion $p\colon (-\lambda_1, \infty) \longrightarrow \mathbb{R}$ durch

$$p(\lambda) := -(B + \lambda I)^{-1} g.$$

Sei $\Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ und $U = (\ u_1 \ \cdots \ u_n\ )$ eine orthogonale Matrix mit einem zugehörigen Orthonormalsystem von Eigenvektoren als Spalten. Dann ist $U^T B U = \Lambda$ und folglich

$$\|p(\lambda)\|_2^2 = \|(\Lambda + \lambda I)^{-1} U^T g\|_2^2 = \sum_{i=1}^{n} \frac{(u_i^T g)^2}{(\lambda_i + \lambda)^2} = \frac{1}{(\lambda_1 + \lambda)^2} \sum_{i=1}^{r} (u_i^T g)^2 + \sum_{i=r+1}^{n} \frac{(u_i^T g)^2}{(\lambda_i + \lambda)^2}.$$

Hieraus liest man ab, dass $\|p(\cdot)\|_2$ auf $(-\lambda_1, \infty)$ beliebig oft differenzierbar sowie (für $g \neq 0$) monoton fallend ist und $\lim_{\lambda \to \infty} \|p(\lambda)\|_2 = 0$ gilt. Ist $g \notin \operatorname{span}\{u_1, \ldots, u_r\}^\perp$, steht g also nicht senkrecht auf dem Eigenraum zum kleinsten Eigenwert λ_1, so ist $\lim_{\lambda \to -\lambda_1 +} \|p(\lambda)\| = \infty$. Wir können nun die folgenden Fälle unterscheiden:

- Es ist $\lambda_1 > 0$, also B positiv definit.

Ist $\|B^{-1} g\|_2 \leq \Delta$, so ist $p^* := -B^{-1} g$ die Lösung von (P), der zugehörige Multiplikator ist $\lambda^* := 0$. Andernfalls ist $\|p(0)\|_2 > \Delta$, es gibt genau eine Lösung $\lambda^* > 0$ von $\|p(\lambda)\|_2 = \Delta$ und $p^* := p(\lambda^*)$ ist die Lösung von (P).

- Es ist $\lambda_1 \leq 0$ und $g \notin \operatorname{span}\{u_1, \ldots, u_r\}^\perp$.

Dann ist $\lim_{\lambda \to -\lambda_1 +} \|p(\lambda)\| = \infty$, $\lim_{\lambda \to \infty} \|p(\lambda)\|_2 = 0$. Da $\|p(\cdot)\|_2$ auf $(-\lambda_1, \infty)$ monoton fallend ist, gibt es genau eine Lösung $\lambda^* > -\lambda_1$ von $\|p(\lambda)\|_2 = \Delta$, damit ist $p^* := p(\lambda^*)$ die Lösung von (P).

- Es ist $\lambda_1 \leq 0$ und $g \in \operatorname{span}\{u_1, \ldots, u_r\}^\perp$. Dies ist der sogenannte schwere Fall, während die beiden ersten Fälle einfach genannt werden.

Dann existiert

$$\Delta_{\text{cri}} := \lim_{\lambda \to -\lambda_1 +} \|p(\lambda)\|_2 = \Big(\sum_{i=r+1}^{n} \frac{(u_i^T g)^2}{(\lambda_i - \lambda_1)^2} \Big)^{1/2},$$

ferner existiert auch

$$p_{\text{cri}} := \lim_{\lambda \to -\lambda_1 +} p(\lambda) = -U(\Lambda - \lambda_1 I)^+ U^T g,$$

wobei

$$(\Lambda - \lambda_1 I)^+ := \text{diag}\,(0, \ldots, 0, 1/(\lambda_{r+1} - \lambda_1), \ldots, 1/(\lambda_n - \lambda_1))$$

die Moore-Penrose verallgemeinerte Inverse von $\Lambda - \lambda_1 I$ ist. Ist $\Delta < \Delta_{\text{cri}}$, so hat $\|p(\lambda)\|_2 = \Delta$ eine Nullstelle $\lambda^* \in (-\lambda_1, \infty)$ und $p^* := p(\lambda^*)$ löst (P). Ist dagegen $\Delta \geq \Delta_{\text{cri}}$, so hat man $\lambda^* := -\lambda_1$ zu setzen. Für ein beliebiges $\alpha \in \mathbb{R}$ ist $(B+\lambda^* I)(p_{\text{cri}}+\alpha u_1) = -g$. Bestimmt man daher α^* als Lösung von $\|p_{\text{cri}} + \alpha u_1\|_2 = \Delta$, so ist $p^* := p_{\text{cri}} + \alpha^* u_1$ eine Lösung von (P).

Beispiel: Durch einfache Beispiele wollen wir die obigen Fälle illustrieren. Sei

$$g := \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad B := \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}.$$

Das Problem (P) ist also ein konvexes Problem. In Abbildung 4.1 links geben wir einen Plot von $\psi(\lambda) := \|p(\lambda)\|_2$ auf $[0, 4]$ an. Wenn Δ groß ist (genauer: wenn $\Delta > \|p(0)\|_2 =$

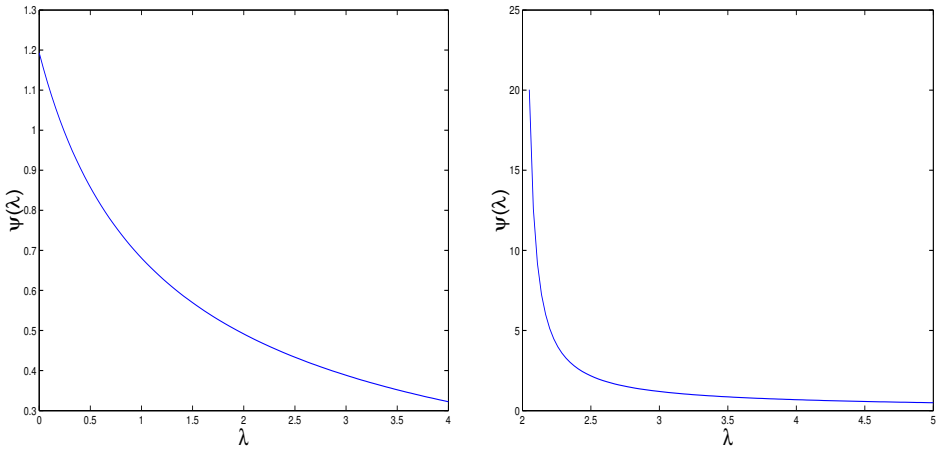

Abbildung 4.1: Die Funktion $\psi(\lambda) := \|p(\lambda)\|_2$: Einfacher Fall

$\sqrt{\frac{205}{144}} \approx 1.19$), ist $\psi(\lambda) < \Delta$ für alle $\lambda \geq 0$ und folglich $\lambda^* = 0$. Ist $\Delta < \psi(0)$, so hat die Gleichung $\psi(\lambda) = \Delta$ genau eine Lösung $\lambda^* > 0$.

Jetzt betrachten wir ein nichtkonvexes Beispiel. Hier sei

$$g := \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad B := \begin{pmatrix} -2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Der kleinste Eigenwert ist jetzt $\lambda_1 = -2$, also $p(\cdot)$ definiert auf $(2, \infty)$. In Abbildung 4.1 rechts geben wir einen Plot von $\psi(\lambda) := \|p(\lambda)\|$ auf $(2, 5]$ an. Man erkennt, dass es hier zu jedem $\Delta > 0$ genau eine Lösung $\lambda^* > 2$ von $\psi(\lambda) = \Delta$ gibt. Nun verändern wir das Beispiel nur geringfügig, indem wir die erste Komponente von g auf 0 setzen. Es sei also

$$g := \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad B := \begin{pmatrix} -2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Dann kann $\psi(\cdot)$ durch

$$\psi(\lambda) = \Big(\frac{1}{(-1+\lambda)^2} + \frac{1}{\lambda^2} + \frac{1}{(1+\lambda)^2}\Big)^{1/2}$$

auf $(1, \infty)$ fortgesetzt werden. In Abbildung 4.2 links geben wir $\psi(\cdot)$ auf $(1, 5)$ an (auch wenn man nicht sehr viel erkennt), in derselben Abbildung rechts auf dem Intervall $[2, 5]$. Hier gibt es den kritischen Wert $\psi(2) = \frac{7}{6} \approx 1.17$: Ist $\Delta < \psi(2)$, so besitzt die

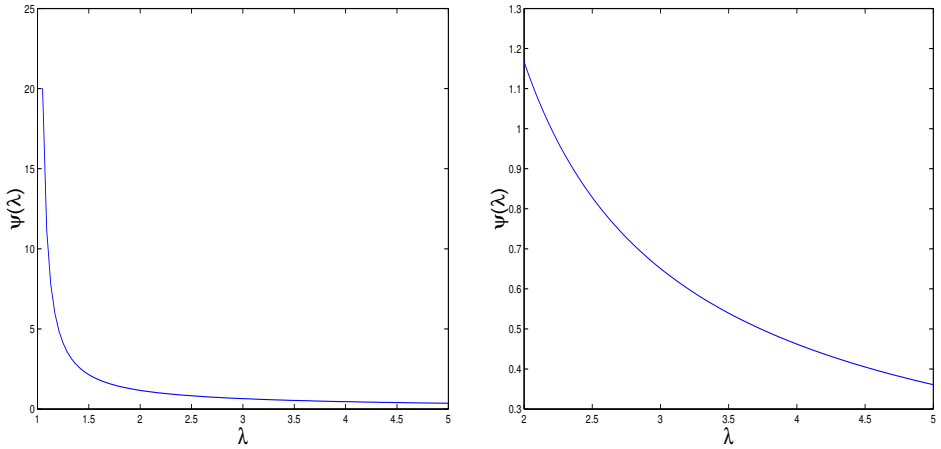

Abbildung 4.2: Die Funktion $\psi(\lambda) := \|p(\lambda)\|_2$: Schwerer Fall

Gleichung $\psi(\lambda) = \Delta$ genau eine Lösung $\lambda^* \in (2, \infty)$. Für $\Delta \geq \psi(2)$ ist notwendigerweise $\lambda^* = -\lambda_1 = 2$. Wir müssen eine Lösung p^* von $(B + \lambda^* I)p^* = -g$ mit $\|p^*\|_2 = \Delta$ bestimmen. In unserem Fall genügt $p^* = (p_1^*, 1, \frac{1}{2}, \frac{1}{3})^T$ mit beliebigem p_1^* dem Gleichungssystem $(B + \lambda^* I)p^* = -g$. Die Gleichung $\|p^*\|_2 = \Delta$ führt auf $\sqrt{(p_1^*)^2 + \frac{49}{36}} = \Delta$ bzw. $p_1^* = \pm\sqrt{\Delta^2 - \frac{49}{36}}$. □

Nun geht es um die numerische Lösung der Gleichung $\|p(\lambda)\|_2 - \Delta = 0$. Es ist keine gute Idee, auf diese Gleichung direkt das Newton-Verfahren anzuwenden. Denn $\psi(\cdot)$ fällt in der Nähe von $-\lambda_1$ sehr rasch ab bzw. die Ableitung $\psi'(\cdot)$ ist dort sehr klein, die Funktion selber also stark nichtlinear. Besser ist es, auf die Gleichung

$$\chi(\lambda) := \frac{1}{\|p(\lambda)\|_2} - \frac{1}{\Delta} = 0$$

das Newton-Verfahren anzuwenden.

Beispiel: Sei

$$g := \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad B := \begin{pmatrix} -2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

(dies ist ein nichtkonvexer einfacher Fall, siehe im obigen Beispiel der zweite Fall). In der Abbildung 4.3 geben wir $1/\|p(\lambda)\|_2$ auf $[2, 5]$ an. Man erkennt, dass zumindestens

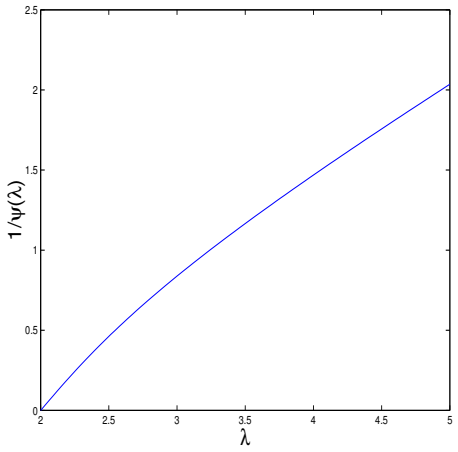

Abbildung 4.3: $1/\|p(\lambda)\|_2$ ist fast linear

in diesem Fall $1/\|p(\cdot)\|_2$ und damit auch $\chi(\cdot)$ in großer Näherung linear ist, so dass man vom auf $\chi(\lambda) = 0$ angewandten Newton-Verfahren sehr gute Ergebnisse erwarten darf. □

Im folgenden Lemma formulieren wir einige Eigenschaften von $\chi(\cdot)$.

Lemma 2.2 *Sei $g \in \mathbb{R}^n \setminus \{0\}$ und $B \in \mathbb{R}^{n\times n}$ symmetrisch mit kleinstem Eigenwert λ_1, ferner sei $\Delta > 0$. Auf $(-\lambda_1, \infty)$ definiere man*

$$p(\lambda) := -(B + \lambda I)^{-1} g, \qquad \chi(\lambda) := \frac{1}{\|p(\lambda)\|_2} - \frac{1}{\Delta}.$$

Dann ist $\chi(\cdot)$ auf $(-\lambda_1, \infty)$ beliebig oft differenzierbar. Die ersten beiden Ableitungen sind gegeben durch

$$\begin{aligned} \chi'(\lambda) &= -\frac{p(\lambda)^T p'(\lambda)}{\|p(\lambda)\|_2^3}, \\ \chi''(\lambda) &= \frac{3[(p(\lambda)^T p'(\lambda))^2 - \|p(\lambda)\|_2^2 \, \|p'(\lambda)\|_2^2]}{\|p(\lambda)\|_2^5}, \end{aligned}$$

wobei

$$p'(\lambda) = -(B + \lambda I)^{-1} p(\lambda).$$

Weiter ist $\chi(\cdot)$ streng monoton wachsend und konkav (d. h. $-\chi$ ist konvex) auf dem Intervall $(-\lambda_1, \infty)$.

Beweis: Sei wieder $U^T BU = \text{diag}\,(\lambda_1, \ldots, \lambda_n)$ eine orthoginale Ähnlichkeitstransformation der symmetrischen Matrix B auf Diagonalgestalt.Dann ist

$$\frac{1}{\|p(\lambda)\|_2} = \Big(\sum_{i=1}^{n} \frac{(u_i^T g)^2}{(\lambda_i + \lambda)^2}\Big)^{-1/2},$$

woraus man die Aussage über die Differenzierbarkeit von $\chi(\cdot)$ auf $(-\lambda_1, \infty)$ abliest. Differentiation von

$$\chi(\lambda) = [p(\lambda)^T p(\lambda)]^{-1/2} - \frac{1}{\Delta}$$

liefert

$$\chi'(\lambda) = -[p(\lambda)^T p(\lambda)]^{-3/2} p(\lambda)^T p'(\lambda) = -\frac{p(\lambda)^T p'(\lambda)}{\|p(\lambda)\|_2^3}.$$

Eine erneute Differentiation führt auf

$$\chi''(\lambda) = \frac{3(p(\lambda)^T p'(\lambda))^2}{\|p(\lambda)\|_2^5} - \frac{p(\lambda)p''(\lambda) + \|p'(\lambda)\|_2^2}{\|p(\lambda)\|_2^3}.$$

Durch zweimaliges Differenzieren der Gleichung $(B + \lambda I)p(\lambda) = -g$ erhält man

$$p(\lambda) + (B + \lambda I)p'(\lambda) = 0, \qquad 2p'(\lambda) + (B + \lambda I)p''(\lambda) = 0.$$

Daher ist

$$p'(\lambda) = -(B + \lambda I)^{-1} p(\lambda), \qquad p''(\lambda) = -2(B + \lambda I)^{-1} p'(\lambda).$$

Einsetzen ergibt

$$\chi'(\lambda) = \frac{p(\lambda)^T (B + \lambda I)^{-1} p(\lambda)}{\|p(\lambda)\|_2^3} > 0,$$

also ist $\chi(\cdot)$ monoton wachsend, und

$$p(\lambda)^T p''(\lambda) = [-(B + \lambda I)p'(\lambda)]^T [-2(B + \lambda I)p'(\lambda)] = 2\|p'(\lambda)\|_2^2,$$

folglich wegen der Cauchy-Schwarzschen Ungleichung

$$\chi''(\lambda) = \frac{3[(p(\lambda)^T p'(\lambda))^2 - \|p(\lambda)\|_2^2 \, \|p'(\lambda)\|_2^2]}{\|p(\lambda)\|_2^5} \leq 0.$$

Daher ist $\chi(\cdot)$ auf $(-\lambda_1, \infty)$ konkav, das Lemma ist bewiesen. □ □

Wir nehmen nun an, es liege der einfache Fall vor und es sei eine Lösung $\lambda^* \in (-\lambda_1, \infty)$ von $\chi(\lambda) = 0$ zu bestimmen. Ist $\lambda_c > -\lambda_1$ eine aktuelle Näherung, so ist

$$\lambda_+ := \lambda_c - \frac{\chi(\lambda_c)}{\chi'(\lambda_c)}$$

die durch das Newton-Verfahren gewonnene, (hoffentlich verbesserte) neue Näherung. Hierzu haben wir $\chi(\lambda_c)$ und $\chi'(\lambda_c)$ bzw. $p(\lambda_c)$ und $p'(\lambda_c)$ auszurechnen. Wegen

$$(B + \lambda_c I)p(\lambda_c) = -g, \qquad (B + \lambda_c I)p'(\lambda_c) = -p(\lambda_c)$$

hat man hierzu zwei lineare Gleichungssysteme mit ein und derselben symmetrischen, positiv definiten Koeffizientenmatrix $B + \lambda_c I$. Es liegt nahe, eine Cholesky-Zerlegung $B + \lambda_c I = L_c L_c^T$ zu berechnen[3] (L_c ist also eine untere Dreiecksmatrix mit positiven Einträgen in der Diagonalen). Dann gewinnt man $p_c = p(\lambda_c)$ durch Rückwärts- und Vorwärtseinsetzen aus $L_c L_c^T p_c = -g$. Wegen

$$\chi'(\lambda_c) = \frac{p(\lambda_c)^T (B + \lambda_c I)^{-1} p(\lambda_c)}{\|p(\lambda_c)\|_2^3} = \frac{p_c^T L_c^{-T} L_c^{-1} p_c}{\|p_c\|_2^3} = \frac{\|L_c^{-1} p_c\|_2^2}{\|p_c\|_2^3}$$

ist die neue Newton-Iterierte gegeben durch

$$\begin{aligned} \lambda_+ &:= \lambda_c - \frac{\chi(\lambda_c)}{\chi'(\lambda_c)} \\ &= \lambda_c - \Big(\frac{1}{\|p_c\|_2} - \frac{1}{\Delta}\Big) \frac{\|p_c\|_2^3}{\|L_c^{-1} p_c\|_2^2} \\ &= \lambda_c + \Big(\frac{\|p_c\|_2 - \Delta}{\Delta}\Big) \Big(\frac{\|p_c\|_2}{\|L_c^{-1} p_c\|}\Big)^2. \end{aligned}$$

Zusammenfassend wird also ein Newton-Schritt folgendermaßen realisiert:

- Input: $\Delta > 0$ und aktuelle Näherung $\lambda_c > -\lambda_1$.
- Berechne Cholesky-Faktorisierung $B + \lambda_c I = L_c L_c^T$.
- Berechne p_c durch Vorwärts- und Rückwärtseinsetzen aus $L_c L_c^T p_c = -g$.
- Berechne w_c durch Vorwärtseinsetzen aus $L_c w_c = p_c$.
- Output: Die neue Newton-Iterierte

$$\lambda_+ := \lambda_c + \Big(\frac{\|p_c\|_2 - \Delta}{\Delta}\Big) \Big(\frac{\|p_c\|_2}{\|w_c\|_2}\Big)^2.$$

Das folgende Lemma zeigt, dass die Konkavität von $\chi(\cdot)$ erfreuliche Konsequenzen hat.

Lemma 2.3 *Für den Startwert $\lambda^{(0)}$ des Newton-Verfahrens gelte $\lambda^{(0)} > -\lambda_1$ und $\chi(\lambda^{(0)}) < 0$. Sei $\{\lambda^{(k)}\}$ die durch das Newton-Verfahren*

$$\lambda^{(k+1)} := \lambda^{(k)} - \frac{\chi(\lambda^{(k)})}{\chi'(\lambda^{(k)})}$$

gewonnene Folge. Es sei $\chi(\lambda^{(k)}) \neq 0$, $k = 0, 1, \ldots$, es erfolge also kein vorzeitiger Abbruch. Dann gilt:

1. *Es ist $\lambda^{(k)} > -\lambda_1$ und $\chi(\lambda^{(k)}) < 0$, $k = 0, 1, \ldots$.*

[3] Hierdurch kann gleichzeitig überprüft werden, ob $\lambda_c > -\lambda_1$.

2. *Die Folge* $\{\lambda^{(k)}\}$ *ist monoton wachsend und konvergiert quadratisch gegen die eindeutige Lösung* $\lambda^* \in (-\lambda_1, \infty)$ *von* $\chi(\lambda) = 0$.

Beweis: Sei $\lambda_c > -\lambda_1$ und $\chi(\lambda_c) < 0$. Dann ist $\lambda_+ := \lambda_c - \chi(\lambda_c)/\chi'(\lambda_c) > \lambda_c$. Die Folge $\{\lambda^{(k)}\}$ ist also monoton wachsend und insbesondere $\lambda^{(k)} > -\lambda_1$, $k = 0, 1, \ldots$. Wegen der Konkavität von $\chi(\cdot)$ ist weiter

$$\chi(\lambda_+) \leq \chi(\lambda_c) + (\lambda_+ - \lambda_c)\chi'(\lambda_c) = 0,$$

womit auch $\chi(\lambda^{(k)}) < 0$, $k = 0, 1, \ldots$, bewiesen ist. Da $\chi(\lambda^{(0)}) < 0$, $\lim_{\lambda \to \infty)} \chi(\lambda) = +\infty$ und $\chi(\cdot)$ monoton wachsend ist, gibt es genau ein $\lambda^* \in (\lambda^{(0)}, \infty)$ mit $\chi(\lambda^*) = 0$. Es ist $\lambda^{(k)} < \lambda^*$, also $\{\lambda^{(k)}\}$ konvergent. Wegen $\chi'(\lambda^*) \neq 0$ ist die Konvergenz quadratisch. Damit ist das einfache Lemma schon bewiesen. □ □

Bemerkung: Startet man das Newton-Verfahren also mit einem Startwert $\lambda^{(0)}$ zwischen $-\lambda_1$ und λ^*, so ist Konvergenz gesichert. Es ist dann

$$\begin{aligned} 0 &< \lambda^* - \lambda^{(k+1)} \\ &= \lambda^* - \lambda^{(k)} + \frac{\chi(\lambda^{(k)})}{\chi'(\lambda^{(k)})} \\ &= (\lambda^* - \lambda^{(k)})\Big(1 - \frac{\chi'(\tilde{\lambda}^{(k)})}{\chi'(\lambda^{(k)})}\Big) \\ &\qquad (\text{mit } \tilde{\lambda}^{(k)} \in (\lambda^{(k)}, \lambda^*)) \\ &\leq (\lambda^* - \lambda^{(k)})\Big(1 - \frac{\chi'(\lambda^*)}{\chi'(\lambda^{(0)})}\Big). \end{aligned}$$

Daher ist das Newton-Verfahren global Q-linear konvergent mit dem Konvergenzfaktor $\gamma := 1 - \chi'(\lambda^*)/\chi'(\lambda^{(0)}) \in [0, 1)$. □

Im folgenden Lemma wird der Fall betrachtet, dass man sich mit einer aktuellen Näherung rechts von dem gesuchten λ^* befindet.

Lemma 2.4 *Sei* $\lambda_c > -\lambda_1$ *und* $\chi(\lambda_c) > 0$. *Ist dann* $\lambda_+ := \lambda_c - \chi(\lambda_c)/\chi'(\lambda_c)$ *die nächste Newton-Iterierte, so ist* $\lambda_+ \leq \lambda^*$. *Ferner ist* $\lambda_+ > -\lambda_1$ *und* $\chi(\lambda_+) \leq 0$ *oder* $\lambda_+ \leq -\lambda_1$.

Beweis: Wegen $\chi(\lambda_c) > 0$ und $\chi'(\lambda_c) > 0$ ist natürlich $\lambda_+ < \lambda_c$. Ist $\lambda_+ > -\lambda_1$, so folgt wieder aus der Konkavität von $\chi(\cdot)$ auf $(-\lambda_1, \infty)$, dass

$$\chi(\lambda_+) \leq \chi(\lambda_c) + (\lambda_+ - \lambda_c)\chi'(\lambda_c) = 0.$$

Damit ist das Lemma bewiesen. □ □

Bemerkung: Für eine effiziente, sichere Implementation wären noch einige Fragen zu klären. Gute Implementationen benutzen z. B. ein sogenanntes "Safeguarding" (Sicherheitsmaßnahmen), d. h. es wird ein "interval of uncertainty" $[\lambda^L, \lambda^U]$ konstruiert, von dem man weiß, dass die gesuchte Lösung λ^* in ihm liegt. Die aktuelle Näherung muss in diesem "Unsicherheitsintervall" liegen, welches natürlich möglichst von Schritt zu

Schritt verkleinert werden sollte. Außerdem müsste geklärt werden, wie das Newton-Verfahren gestartet werden sollte und wann es abgebrochen wird. Weiter ist die Berechnung von Cholesky-Faktorisierungen bei hochdimensionalen Problemen kaum möglich. Auf alle diese Fragen wird ausführlich bei A. R. Conn, N. I. M. Gould, Ph. L. Toint (2000, S. 176 ff.) eingegangen, ferner findet man dort Hinweise auf die relevante Originalliteratur. □

4.2.2 Globale Konvergenz

Ziel in diesem Unterabschnitt ist es, einen globalen Konvergenzsatz für ein Trust-Region-Verfahren zu formulieren und zu beweisen. Im quadratischen Modell wird die auftretende symmetrische Matrix B bzw. B_k noch nicht spezifiziert. Als Hilfssatz benötigen wir das folgende Lemma.

Lemma 2.5 *Man betrachte das Trust-Region-Hilfsproblem*

$$\textit{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T B p, \quad \|p\|_2 \le \Delta,$$

wobei $f \in \mathbb{R}$, $g \in \mathbb{R}^n$, die symmetrische Matrix $B \in \mathbb{R}^{n\times n}$ und $\Delta > 0$ gegeben sind. Sei p^* eine globale Lösung dieser Aufgabe. Dann gilt:

1. *Es ist $\phi(p^*) = f$ genau dann, wenn $g = 0$ und B positiv semidefinit ist.*
2. *Es ist*
$$f - \phi(p^*) \ge \frac{1}{2}\|g\|_2 \min\Big(\Delta, \frac{\|g\|_2}{\|B\|_2}\Big).$$

Beweis: Ist $\phi(p^*) = f$, so ist auch $p^{**} := 0$ eine Lösung. Aus Satz 2.1 folgt sofort, dass $g = 0$ und B positiv semidefinit ist. Die Umkehrung ist trivial.

Für den Beweis des zweiten Teils des Lemmas können wir o. B. d. A. $g \neq 0$ annehmen. Für ein beliebiges p mit $\|p\|_2 \le \Delta$ ist wegen der Optimalität von p^* offenbar

$$(*) \qquad f - \phi(p^*) \ge f - \phi(p) = -g^T p - \frac{1}{2} p^T B p \ge -g^T p - \frac{1}{2}\|p\|_2^2 \, \|B\|_2.$$

Ist $\Delta\|B\|_2 \le \|g\|_2$, so ist wegen $(*)$ (setze $p := -(\Delta/\|g\|_2)g$)

$$f - \phi(p^*) \ge \Delta\|g\|_2 - \frac{1}{2}\Delta^2\|B\|_2 \ge \frac{1}{2}\Delta\|g\|_2.$$

Ist dagegen $\Delta\|B\|_2 > \|g\|_2$, so setze man $p := -(1/\|B\|_2)g$ und erhalte aus $(*)$

$$f - \phi(p^*) \ge \frac{\|g\|_2^2}{\|B\|_2} - \frac{\|g\|_2^2}{2\|B\|_2} = \frac{\|g\|_2^2}{2\|B\|_2},$$

insgesamt ist die behauptete Abschätzung bewiesen. □ □

Nun folgt der angekündigte globale Konvergenzsatz.

Satz 2.6 *Gegeben sei die unrestringierte Optimierungsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n.$$

Die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ sei kompakt, wobei x_0 der Startvektor für das gleich anzugebende Verfahren ist. Die Zielfunktion f sei auf einer offenen Obermenge von L_0 stetig differenzierbar und der Gradient $\nabla f(\cdot)$ auf L_0 lipschitzstetig. Ferner sei $\{B_k\} \subset \mathbb{R}^{n\times n}$ eine beschränkte Folge symmetrischer Matrizen. Man betrachte das folgende Verfahren.

- *Gegeben seien Konstanten $0 < \rho_1 < \rho_2 < 1$, $\sigma_1 \in (0,1)$ und $\sigma_2 > 1$.*
- *Seien $x_0 \in \mathbb{R}^n$ und $\Delta_0 > 0$ gegeben. Berechne $g_0 := \nabla f(x_0)$.*
- *Für $k = 0, 1, \ldots$:*
  - *Berechne eine globale Lösung p_k der Aufgabe*

    $$(\mathrm{P}_k) \qquad \textit{Minimiere} \quad f_k(p) := f(x_k) + g_k^T p + \frac{1}{2} p^T B_k p, \quad \|p\|_2 \leq \Delta_k.$$

  - *Falls $f(x_k) = f_k(p_k)$, dann: STOP. Es ist $g_k = 0$ und daher x_k eine stationäre Lösung von (P).*
  - *Berechne*

    $$r_k := \frac{f(x_k) - f(x_k + p_k)}{f(x_k) - f_k(p_k)}.$$

  - *Falls $r_k \geq \rho_1$, dann setze $x_{k+1} := x_k + p_k$ und berechne $g_{k+1} := \nabla f(x_{k+1})$. In diesem Falle nennen wir den Iterationsschritt k erfolgreich.*
  - *Andernfalls setze $x_{k+1} := x_k$ sowie $g_{k+1} := g_k$.*
  - *Update des Trust-Region-Radius:*
    - *Falls $r_k < \rho_1$, dann wähle $\Delta_{k+1} \in (0, \sigma_1 \Delta_k]$.*
    - *Falls $r_k \in [\rho_1, \rho_2)$, dann wähle $\Delta_{k+1} \in [\sigma_1 \Delta_k, \Delta_k]$.*
    - *Falls $r_k \geq \rho_2$, dann wähle $\Delta_{k+1} \in [\Delta_k, \sigma_2 \Delta_k]$.*

Das Verfahren breche nicht vorzeitig ab. Dann liefert es eine Folge $\{x_k\}$ mit $\lim_{k\to\infty} g_k = 0$, insbesondere ist jeder Häufungspunkt der Folge $\{x_k\}$ eine stationäre Lösung von (P)[4].

Beweis: Als Vorbereitung zeigen wir zunächst, dass $\liminf_{k\to\infty} \|g_k\|_2 = 0$. Angenommen, das sei nicht der Fall. Dann existiert ein $\epsilon > 0$ mit $\|g_k\|_2 \geq \epsilon$ für alle k. Wir zeigen, dass einerseits $\sum_{k=0}^{\infty} \Delta_k < \infty$, insbesondere die Folge $\{\Delta_k\}$ gegen Null konvergiert, und andererseits $\lim_{k\to\infty} r_k = 1$ ist. Da letzteres zeigt, dass $r_k \geq \rho_2$ und damit $\Delta_{k+1} \geq \Delta_k$ für alle hinreichend großen k gilt, wird man einen Widerspruch erreicht haben.

[4]Die folgende Aussage ist leicht zu beweisen: Sei $\{x_k\} \subset L_0$ und L_0 kompakt. Dann ist $\lim_{k\to\infty} \nabla f(x_k) = 0$ genau dann, wenn jeder Häufungspunkt von $\{x_k\}$ ein stationärer Punkt von f ist.

Gibt es nur endlich viele erfolgreiche Iterationsschritte, so ist $\Delta_{k+1} \le \sigma_1 \Delta_k$ für alle hinreichend großen k, so dass $\sum_{k=0}^{\infty} \Delta_k < \infty$ wegen $\sigma_1 \in (0,1)$ trivialerweise richtig ist. Ist der Iterationsschritt k erfolgreich, so ist wegen der Abschätzung in Lemma 2.5

$$f(x_k) - f(x_{k+1}) \ge \rho_1 [f(x_k) - f_k(p_k)] \ge \frac{\rho_1}{2} \|g_k\|_2 \min\Big(\Delta_k, \frac{\|g_k\|_2}{\|B_k\|_2}\Big).$$

Wegen $\|g_k\|_2 \ge \epsilon$ und mit $\beta := \sup_{k=0,1,\ldots} \|B_k\|_2$ (hier geht die Beschränktheit der Folge $\{B_k\}$ ein) ist daher für jeden erfolgreichen Iterationsschritt k:

$$(*) \qquad f(x_k) - f(x_{k+1}) \ge \frac{\rho_1 \epsilon}{2} \min\Big(\Delta_k, \frac{\epsilon}{\beta}\Big).$$

Nun nehmen wir an, dass es unendlich viele erfolgreiche Iterationsschritte gibt. Diese seien in der Indexmenge E zusammengefasst. Da $\{f(x_k)\}$ eine monoton nicht wachsende, nach unten beschränkte Folge ist, erhält man aus $(*)$

$$\frac{\rho_1 \epsilon}{2} \sum_{k \in E} \min\Big(\Delta_k, \frac{\epsilon}{\beta}\Big) \le \sum_{k \in E} [f(x_k) - f(x_{k+1})] \le \sum_{k=0}^{\infty} [f(x_k) - f(x_{k+1})] < \infty$$

und hieraus $\sum_{k \in E} \Delta_k < \infty$. Nun betrachten wir die k, die zwischen zwei aufeinanderfolgenden erfolgreichen Iterationsschritten $i < j$ liegen. Dann ist $\Delta_{i+1} \le \sigma_2 \Delta_i$ (da i erfolgreich) und $\Delta_{k+1} \le \sigma_1 \Delta_k$ für $k = i+1, \ldots, j-1$. Hieraus folgt durch Zurückspulen und Summieren

$$\sum_{k=i+1}^{j-1} \Delta_k \le \frac{\sigma_2}{1-\sigma_1} \Delta_i.$$

Also ist auch $\sum_{k \notin E} \Delta_k < \infty$ und insgesamt $\sum_{k=0}^{\infty} \Delta_k < \infty$. Aus

$$\sum_{k=0}^{\infty} \|x_{k+1} - x_k\|_2 \le \sum_{k=0}^{\infty} \|p_k\|_2 \le \sum_{k=0}^{\infty} \Delta_k < \infty$$

folgt, dass $\{x_k\}$ eine Cauchy-Folge ist[5] und damit gegen ein x^* konvergiert. Nun wird $\lim_{k \to \infty} r_k = 1$ nachgewiesen. Für alle hinreichend großen k ist

$$\begin{aligned}
|r_k - 1| &= \frac{|f(x_k) + g_k^T p_k + \frac{1}{2} p_k^T B_k p_k - f(x_k + p_k)|}{f(x_k) - f_k(p_k)} \\
&\le \frac{2}{\epsilon \Delta_k} \Big| f(x_k) + g_k^T p_k + \frac{1}{2} p_k^T B_k p_k - f(x_k + p_k) \Big| \\
&\qquad \text{(erneute Anwendung der Abschätzung in Lemma 2.5)} \\
&\le \frac{2}{\epsilon \|p_k\|_2} \Big\{ |[\nabla f(x_k) - \nabla f(x_k + \theta_k p_k)]^T p_k| + \frac{1}{2} \|B_k\|_2 \, \|p_k\|_2^2 \Big\} \\
&\le \frac{2}{\epsilon} \Big\{ \|\nabla f(x_k) - \nabla f(x_k + \theta_k p_k)\|_2 + \frac{1}{2} \beta \|p_k\|_2 \Big\}
\end{aligned}$$

[5] Denn: Sei $\epsilon > 0$ vorgegeben. Wegen $\sum_{k=0}^{\infty} \|x_{k+1} - x_k\|_2 < \infty$ existiert ein $k_0 = k_0(\epsilon)$ mit $\sum_{j=k_0}^{\infty} \|x_{j+1} - x_j\|_2 \le \epsilon$. Für $k_0 \le k < l$ ist dann

$$\|x_l - x_k\|_2 \le \sum_{j=k}^{l-1} \|x_{j+1} - x_j\|_2 \le \sum_{j=k_0}^{\infty} \|x_{j+1} - x_j\|_2 \le \epsilon.$$

Dies zeigt, dass $\{x_k\}$ eine Cauchy-Folge ist.

mit $\theta_k \in (0,1)$. Wegen $p_k \to 0$ folgt $r_k \to 1$. Das aber ist, wie wir am Anfang schon festgestellt haben, ein Widerspruch zu $\Delta_k \to 1$. Damit ist $\liminf_{k\to\infty} \|g_k\|_2 = 0$ bewiesen.

Auch die Behauptung des Satzes, dass nämlich $\lim_{k\to\infty} g_k = 0$ gilt, zeigen wir durch Widerspruch. Ist das nicht der Fall, so gibt es ein $\epsilon > 0$ und eine Teilfolge $\{x_{k(i)}\} \subset \{x_k\}$ mit $\|g_{k(i)}\|_2 \geq 2\epsilon$, $i = 1, 2, \ldots$. Wegen $\liminf_{k\to\infty} \|g_k\|_2 = 0$ gibt es unendlich viele k mit $\|g_k\|_2 < \epsilon$. Für $i = 1, 2, \ldots$ existiert daher ein $l(i) > k(i)$ mit

$$\|g_k\|_2 \geq \epsilon \quad (k(i) \leq k < l(i)), \qquad \|g_{l(i)}\|_2 < \epsilon \qquad (i = 1, 2, \ldots).$$

Hieraus erkennen wir: Ist $k(i) \leq k < l(i)$ und der Iterationsschritt k erfolgreich, so ist wegen $\|x_{k+1} - x_k\|_2 \leq \Delta_k$ und der Abschätzung in Lemma 2.5

$$f(x_k) - f(x_{k+1}) \geq \rho_1[f(x_k) - f_k(p_k)] \geq \frac{\rho_1\epsilon}{2} \min\Big(\|x_{k+1} - x_k\|_2, \frac{\epsilon}{\beta}\Big),$$

wobei wie oben wieder $\beta := \sup_{k=0,1,\ldots} \|B_k\|_2$ gesetzt wurde. Da die Folge $\{f(x_k)\}$ (als monoton nicht wachsende, nach unten beschränkte Folge) konvergiert, schließen wir hieraus, dass

$$f(x_k) - f(x_{k+1}) \geq \frac{\rho_1\epsilon}{2}\|x_{k+1} - x_k\|_2, \qquad k(i) \leq k < l(i),$$

für alle hinreichend großen i (für nicht erfolgreiche k ist diese Ungleichung trivial). Daher ist

$$\begin{aligned} \frac{\rho_1\epsilon}{2}\|x_{k(i)} - x_{l(i)}\|_2 &\leq \frac{\rho_1\epsilon}{2} \sum_{k=k(i)}^{l(i)-1} \|x_{k+1} - x_k\|_2 \\ &\leq \sum_{k=k(i)}^{l(i)-1} [f(x_k) - f(x_{k+1})] \\ &= f(x_{k(i)}) - f(x_{l(i)}) \end{aligned}$$

für alle hinreichend großen i. Da die Folge $\{f(x_k)\}$ konvergiert, insbesondere also eine Cauchy-Folge ist, konvergiert hier die rechte Seite gegen Null. Daher konvergiert $\{\|x_{k(i)} - x_{l(i)}\|_2\}$ und wegen der vorausgesetzten Lipschitzstetigkeit von $\nabla f(\cdot)$ auf L_0 auch $\{\|g_{k(i)} - g_{l(i)}\|_2\}$ gegen Null. Andererseits ist

$$\|g_{k(i)} - g_{l(i)}\|_2 \geq \|g_{k(i)}\|_2 - \|g_{l(i)}\|_2 \geq 2\epsilon - \epsilon = \epsilon,$$

womit wir den gewünschten Widerspruch erhalten haben. Der Satz ist bewiesen. □ □

Bemerkung: Die Voraussetzung im vorigen Satz, dass p_k eine (exakte) globale Lösung des Trust-Region-Hilfsproblems ist, kann wesentlich abgeschwächt werden. So würde es genügen, dass mit Konstanten $c_1 \in (0,1)$ und $\gamma \geq 1$ die Bedingungen

$$f(x_k) - f_k(p_k) \geq c_1\|g_k\|_2 \min\Big(\Delta_k, \frac{\|g_k\|_2}{\|B_k\|_2}\Big)$$

und

$$\|p_k\|_2 \leq \gamma \|\Delta_k\|$$

gelten. Beide Bedingungen gelten natürlich für die exakte Lösung von (P_k). In Aufgabe 6 kann gezeigt werden, dass dies auch für den wesentlich leichter zu berechnenden *Cauchy-Punkt* gilt. In Aufgabe 7 findet man ein entsprechendes Resultat für die "Dog Leg Strategie". □

4.2.3 Das Trust-Region-Newton-Verfahren

Im letzten Unterabschnitt wurde über die Matrixfolge $\{B_k\} \subset \mathbb{R}^{n\times n}$, die im quadratischen Modell auftreten, lediglich vorausgesetzt, dass sie beschränkt ist. In diesem Unterabschnitt setzen wir $B_k := \nabla^2 f(x_k)$ (natürlich müssen wir jetzt entsprechend stärkere Glattheitsvoraussetzungen an die Zielfunktion f stellen) und sprechen dann vom *Trust-Region-Newton-Verfahren*.

Um ein Blättern zu vermeiden, formulieren wir den Algorithmus aus Satz 2.6 im folgenden Satz noch einmal neu, auch wenn der Unterschied nur darin besteht, dass diesmal $B_k = \nabla^2 f(x_k)$.

Satz 2.7 *Gegeben sei die unrestringierte Optimierungsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n.$$

Die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ sei kompakt, wobei x_0 der Startvektor für das gleich anzugebende Verfahren ist. Die Zielfunktion f sei auf einer offenen Obermenge von L_0 zweimal stetig differenzierbar und der Gradient $\nabla f(\cdot)$ auf L_0 lipschitzstetig. Man betrachte das folgende Verfahren.

- *Gegeben seien Konstanten $0 < \rho_1 < \rho_2 < 1$, $\sigma_1 \in (0,1)$ und $\sigma_2 > 1$.*
- *Seien $x_0 \in \mathbb{R}^n$ und $\Delta_0 > 0$ gegeben. Berechne $g_0 := \nabla f(x_0)$, $B_0 := \nabla^2 f(x_0)$.*
- *Für $k = 0, 1, \ldots$:*
  - *Berechne eine globale Lösung p_k der Aufgabe*

$$(P_k) \qquad \textit{Minimiere} \quad f_k(p) := f(x_k) + g_k^T p + \frac{1}{2} p^T B_k p, \quad \|p\|_2 \leq \Delta_k.$$

  - *Falls $f(x_k) = f_k(p_k)$, dann: STOP. In x_k sind die notwendigen Bedingungen zweiter Ordnung erfüllt, siehe erster Teil von Lemma 2.5..*
  - *Berechne*

$$r_k := \frac{f(x_k) - f(x_k + p_k)}{f(x_k) - f_k(p_k)}.$$

  - *Falls $r_k \geq \rho_1$, dann setze $x_{k+1} := x_k + p_k$ und berechne $g_{k+1} := \nabla f(x_{k+1})$, $B_{k+1} := \nabla^2 f(x_{k+1})$. In diesem Falle nennen wir den Iterationsschritt k erfolgreich.*
  - *Andernfalls setze $x_{k+1} := x_k$ sowie $g_{k+1} := g_k$, $B_{k+1} := B_k$.*

- *Update des Trust-Region-Radius:*
  - *Falls $r_k < \rho_1$, dann wähle $\Delta_{k+1} \in (0, \sigma_1 \Delta_k]$.*
  - *Falls $r_k \in [\rho_1, \rho_2)$, dann wähle $\Delta_{k+1} \in [\sigma_1 \Delta_k, \Delta_k]$.*
  - *Falls $r_k \geq \rho_2$, dann wähle $\Delta_{k+1} \in [\Delta_k, \sigma_2 \Delta_k]$.*

Das Verfahren breche nicht vorzeitig ab. Dann liefert es eine Folge $\{x_k\}$ mit:

1. *Die Folge $\{x_k\}$ besitzt mindestens einen Häufungspunkt x^*, in dem die notwendigen Optimalitätsbedingungen zweiter Ordnung erfüllt sind, für den also $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ positiv semidefinit ist.*

2. *Ist x^* ein Häufungspunkt der Folge $\{x_k\}$ und ist $\nabla^2 f(x^*)$ positiv definit, so konvergiert die Folge $\{x_k\}$ gegen x^*. Ferner sind nach endlich vielen Schritten alle Iterationsschritte erfolgreich, es ist $\|p_k\|_2 < \Delta_k$ für alle hinreichend großen k und $\inf_{k=0,1,\ldots} \Delta_k > 0$. Ist darüberhinaus $\nabla^2 f(\cdot)$ auf einer Kugel um x^* in x^* lipschitzstetig, so konvergiert die Folge $\{x_k\}$ von mindestens zweiter Ordnung gegen x^*.*

Beweis: Mit $\lambda_{\min}(A)$ bezeichnen wir den kleinsten Eigenwert der symmetrischen Matrix $A \in \mathbb{R}^{n \times n}$. Wir zeigen im Anschluss durch Widerspruch, dass

$$\lambda^* := \limsup_{k \to \infty} \lambda_{\min}(B_k) \geq 0.$$

Ist dies gelungen, so existiert eine Teilfolge $\{x_{k(i)}\} \subset \{x_k\}$ mit

$$\lambda^* = \lim_{i \to \infty} \lambda_{\min}(\nabla^2 f(x_{k(i)})) \geq 0.$$

Ein Häufungspunkt x^* der Folge $\{x_{k(i)}\}$ (ein solcher existiert, da $\{x_{k(i)}\} \subset L_0$ und L_0 kompakt ist) ist wegen Satz 2.6 notwendig stationär (d. h. $\nabla f(x^*) = 0$), wegen $\lambda_{\min}(\nabla^2 f(x^*)) = \lambda^* \geq 0$ ist $\nabla^2 f(x^*)$ positiv semidefinit. Damit wird der erste Teil des Satzes bewiesen sein.

Angenommen, es gibt ein $\epsilon > 0$ mit $\lambda_{\min}(B_k) \leq -\epsilon$ für alle hinreichend großen k. Sei q_k ein durch $\|q_k\|_2 = \Delta_k$ und $g_k^T q_k \leq 0$ normierter Eigenvektor zum Eigenwert $\lambda_{\min}(B_k)$. Da p_k eine globale Lösung von (P_k) ist, erhalten wir

$$f_k(p_k) \leq f_k(q_k) = f(x_k) + \underbrace{g_k^T q_k}_{\leq 0} + \frac{1}{2} q_k^T B_k q_k \leq f(x_k) - \frac{\epsilon}{2} \Delta_k^2$$

und damit

$$(*) \qquad f(x_k) - f_k(p_k) \geq \frac{\epsilon}{2} \Delta_k^2$$

für alle hinreichend großen k. Hieraus folgt $\sum_{k=0}^{\infty} \Delta_k^2 < \infty$ mit derselben Argumentation wie zu Beginn des Beweises von Satz 2.6, als $\sum_{k=0}^{\infty} \Delta_k < \infty$ nachgewiesen wurde. Mit $\{\Delta_k\}$ ist wegen $\|p_k\|_2 \leq \Delta_k$ auch $\{\|p_k\|_2\}$ eine Nullfolge. Hieraus folgern wir, wieder

ähnlich einem entsprechenden Teil des Beweises von Satz 2.6, dass $\lim_{k\to\infty} r_k = 1$. Denn wegen $(*)$ ist für alle hinreichend großen k

$$\begin{aligned} |r_k - 1| &= \frac{|f(x_k) + g_k^T p_k + \frac{1}{2} p_k^T B_k p_k - f(x_k + p_k)|}{f(x_k) - f_k(p_k)} \\ &\leq \frac{2}{\epsilon \|p_k\|_2^2} |f(x_k) + g_k^T p_k + \frac{1}{2} p_k^T B_k p_k - f(x_k + p_k)| \\ &= \frac{1}{\epsilon \|p_k\|_2^2} |p_k^T [\nabla^2 f(x_k) - \nabla^2 f(x_k + \theta_k p_k)] p_k| \quad \text{mit } \theta_k \in (0,1) \\ &\leq \frac{1}{\epsilon} \|\nabla^2 f(x_k) - \nabla^2 f(x_k + \theta_k p_k)\|_2. \end{aligned}$$

Hieraus folgt $\lim_{k\to\infty} r_k = 1$, damit $r_k \geq \rho_2$ und $\Delta_{k+1} \geq \Delta_k$ für alle hinreichend großen k. Das ist ein Widerspruch zu $\lim_{k\to\infty} \Delta_k = 0$. Der erste Teil des Satzes ist damit bewiesen.

Für den Rest des Beweises sei x^* ein Häufungspunkt der Folge $\{x_k\}$ mit der Eigenschaft, dass $\nabla^2 f(x^*)$ positiv definit ist bzw. die hinreichende Optimalitätsbedingung zweiter Ordnung erfüllt ist.

Zunächst wird die Konvergenz der Folge $\{x_k\}$ gegen x^* nachgewiesen. Da $\nabla^2 f(x^*)$ positiv definit ist, gibt es positive Konstanten c und δ mit $c \leq \lambda_{\min}(\nabla^2 f(x))$ für alle x aus der (euklidischen) Kugel um x^* mit dem Radius δ. Ein $\epsilon \in (0, \delta]$ sei beliebig vorgegeben. Da x^* ein Häufungspunkt von $\{x_k\}$ ist und $\lim_{k\to\infty} \|g_k\|_2 = 0$ wegen Satz 2.6, gibt es ein $l \in \mathbb{N}$ mit

$$\|x_l - x^*\|_2 \leq \frac{\epsilon}{2}, \qquad \|g_k\|_2 \leq \frac{c\epsilon}{4} \qquad \text{für alle } k \geq l.$$

Wir wollen $\|x_k - x^*\|_2 \leq \frac{1}{2}\epsilon$ für alle $k \geq l$ zeigen, womit die Konvergenz der Folge $\{x_k\}$ gegen x^* bewiesen sein wird. Dies geschieht durch vollständige Induktion nach k, wobei der Induktionsanfang bei $k = l$ gesichert ist. Für den Induktionsschluss sei also $k \geq l$ und $\|x_k - x^*\|_2 \leq \frac{1}{2}\epsilon$. Da p_k Lösung des Hilfsproblems (P_k) und $p = 0$ hierfür zulässig ist, erhalten wir

$$g_k^T p_k + \frac{c}{2} \|p_k\|_2^2 \leq g_k^T p_k + \frac{1}{2} p_k^T B_k p_k \leq 0$$

und hieraus mit Hilfe der Cauchy-Schwarzschen Ungleichung

$$\frac{c}{2} \|p_k\|_2 \leq \|g_k\|_2 \leq \frac{c\epsilon}{4},$$

also $\|p_k\|_2 \leq \frac{1}{2}\epsilon$. Daher ist

$$\|x_{k+1} - x^*\|_2 \leq \|x_{k+1} - x_k\|_2 + \|x_k - x^*\|_2 \leq \|p_k\|_2 + \frac{\epsilon}{2} \leq \epsilon.$$

Aus (gleichmäßige Konvexität von f auf der δ-Kugel um x^*)

$$\frac{c}{2} \|x^* - x_{k+1}\|_2^2 + g_{k+1}^T (x^* - x_{k+1}) \leq f(x^*) - f(x_{k+1}) \leq 0$$

erhält man ebenso $\|x_{k+1} - x^*\|_2 \leq \frac{1}{2}\epsilon$. Damit ist die Konvergenz der Folge $\{x_k\}$ gegen x^* bewiesen.

Nun zeigen wir $\lim_{k\to\infty} r_k = 1$, womit bewiesen sein wird, dass sogar der verschärfte Test $r_k \geq \rho_2$ für alle hinreichend großen k erfüllt ist. Eben haben wir u. a. erhalten, dass eine Konstante $c > 0$ mit $\frac{1}{2}c\|p_k\|_2 \leq \|g_k\|_2$ für alle hinreichend großen k existiert. Berücksichtigt man noch $\|p_k\|_2 \leq \Delta_k$, die Abschätzung in Lemma 2.5 und die Beschränktheit von $\{\|B_k\|_2\}$, so erhält man die Existenz einer Konstanten $c_0 > 0$ mit

$$f(x_k) - f_k(p_k) \geq \frac{1}{2}\|g_k\|_2 \min\Big(\Delta_k, \frac{\|g_k\|_2}{\|B_k\|_2}\Big) \geq c_0\|p_k\|_2^2$$

für alle hinreichend großen k. Wie oben folgt hieraus $\lim_{k\to\infty} r_k = 1$.

Bis auf endlich viele vergebliche Versuche sind alle Iterationsschritte erfolgreich. Für alle hinreichend großen k ist sogar $r_k \geq \rho_2$. Daher wird der Trust-Region-Radius nach endlich vielen Schritten nicht mehr verkleinert, insbesondere ist $\inf_{k=0,1,\ldots} \Delta_k > 0$. Andererseits ist $x_{k+1} = x_k + p_k$ für alle hinreichend großen k. Die Konvergenz der Folge $\{x_k\}$ liefert $\lim_{k\to\infty} \|p_k\|_2 = 0$, also ist $\|p_k\| < \Delta_k$ für alle hinreichend großen k. Daher (siehe Satz 2.1) ist $B_k p_k = -g_k$ für alle hinreichend großen k. Nach endlich vielen Schritten geht die Trust-Region-Modifikation des Newton-Verfahrens in das ungedämpfte Newton-Verfahren über und erbt deswegen dessen Konvergenzeigenschaften.

Damit ist der Satz bewiesen. □ □

Bemerkung: Einige Modifikationen (Skalierung, andere Update-Strategie für den Trust-Region-Radius u. a.) sind denkbar, wir wollen hierauf nicht mehr eingehen. □

4.2.4 Aufgaben

1. Man berechne die Lösung der Aufgabe

   (P) $$\text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2}p^T B p, \quad \|p\|_2 \leq \Delta,$$

   wobei
   $$g := \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad B := \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \Delta := \frac{1}{2}.$$

2. Sei $B \in \mathbb{R}^{n\times n}$ symmetrisch mit kleinstem Eigenwert λ_1 und $g \in \mathbb{R}^n \setminus \{0\}$. Man zeige, dass die durch
   $$\psi(\lambda) := \|(B + \lambda I)^{-1} g\|_2$$
   definierte Funktion $\psi\colon (-\lambda_1, \infty) \longrightarrow \mathbb{R}$ auf $(-\lambda_1, \infty)$ monoton fallend und konvex ist.

3. Sei $B \in \mathbb{R}^{n\times n}$ symmetrisch mit kleinstem Eigenwert λ_1, sei weiter $\lambda \in \mathbb{R}$.

   (a) Für jedes $u \in \mathbb{R}^n$ mit $\|u\|_2 = 1$ ist
   $$\lambda_+ := \lambda - u^T(B + \lambda I)u \leq -\lambda_1.$$

   (b) Sei $\lambda > -\lambda_1$ und $B + \lambda I = LL^T$ eine Cholesky-Zerlegung, also $L \in \mathbb{R}^{n\times n}$ eine untere Dreiecksmatrix mit positiven Diagonalelementen. Um eine möglichst gute untere Schranke für $-\lambda_1$ zu erhalten, bestimmt man einen Vektor $u \in \mathbb{R}^n$ mit

$\|u\|_2 = 1$, für den $u^T(B+\lambda I)u = \|L^T u\|_2^2$ "klein" ist. Eine mögliche Heuristik zur Bestimmung von u ist die folgende (siehe A. R. Conn, N. I. M. Gould, Ph. L. Toint (2000, S. 191)): Mit einem Vektor $v \in \mathbb{R}^n$ mit $v_i \in \{-1, 1\}$, $i = 1, \ldots, n$, setze

$$u := \frac{(B+\lambda I)^{-1} v}{\|(B+\lambda I)^{-1} v\|_2}.$$

Dann ist

$$u^T(B+\lambda I)u = \frac{v^T(B+\lambda I)^{-1} v}{\|(B+\lambda I)^{-1} v\|_2^2}.$$

Daher sollte man v so wählen, dass

$$\|(B+\lambda I)^{-1} v\|_2 = \|L^{-T} L^{-1} v\|_2$$

bzw. $\|L^{-1}v\|$ möglichst groß ist. Hierzu bestimme man $w = Lv$ folgendermaßen. Angenommen, $w_1, \ldots, w_{k-1}$ sind schon bestimmt. Man setze

$$v_k := \begin{cases} -1, & \text{falls} \quad \sum_{i=1}^{k-1} l_{ki} w_i > 0, \\ +1, & \text{falls} \quad \sum_{i=1}^{k-1} l_{ki} w_i \leq 0 \end{cases}$$

und anschließend

$$w_k := \frac{1}{l_{kk}} \left(v_k - \sum_{i=1}^{k-1} l_{ki} w_i \right).$$

Weiter ist

$$u = \frac{(B+\lambda I)^{-1} v}{\|(B+\lambda I)^{-1} v\|_2} = \frac{L^{-T} w}{\|L^{-T} w\|_2}.$$

Man schreibe eine Matlab-function `Estimate.m`, welche den Overhead hat:

```
function [lambda_plus,info]=Estimate(B,lambda);
%Input-parameter:
%    B           symmetric matrix
%    lambda      real number, usually lambda>-lambda_1
%Output-parameter:
%    lambda_plus lower bound of -lambda_1, if successfull
%    info        info=0, if successfull
%                info=1, if the Cholesky decomposition of
%                B+lambda*I does not exist (because lambda is
%                not an upper bound of -lambda_1)
%************************************************************
```

Anschließend erprobe man die Funktion an den Daten

$$B := \begin{pmatrix} 4 & -3 & 5 & -1 \\ -3 & 2 & -6 & 4 \\ 5 & -6 & 1 & 5 \\ -1 & 4 & 5 & -8 \end{pmatrix}, \qquad \lambda = 9,\ 8,\ 7.7.$$

4. Sei $B \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $g \in \mathbb{R}^n$ und $\Delta > 0$. Man schreibe eine Matlab-Funktion `TrustStep` zur Berechnung der Lösung des Trust-Region-Hilfsproblems

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2} p^T B p, \quad \|p\|_2 \le \Delta.$$

Anschließend teste man die Funktion für den Spezialfall, dass

$$B := \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{n\times n}, \qquad g := \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} \in \mathbb{R}^n$$

mit $n := 10$ und $\Delta := 1$.

Hinweis: Wegen der Voraussetzung, dass B positiv definit ist, kann der schwere Fall nicht eintreten, so dass die Aufgabenstellung einfach ist.

5. Sei $f \in \mathbb{R}$, $g \in \mathbb{R}^n$, $B \in \mathbb{R}^{n\times n}$ symmetrisch, $D \in \mathbb{R}^n$ nichtsingulär und $\Delta > 0$. Man gebe notwendige und hinreichende Bedingungen dafür an, dass ein $p^* \in \mathbb{R}^n$ mit $\|Dp^*\|_2 \le \Delta$ eine globale Lösung von

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T B p, \quad \|Dp\|_2 \le \Delta$$

ist.

6. Gegeben sei das Trust-Region Hilfsproblem

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T B p, \quad \|p\|_2 \le \Delta,$$

wobei $f \in \mathbb{R}$, $g \in \mathbb{R}^{n\times n} \setminus \{0\}$, die symmetrische Matrix $B \in \mathbb{R}^{n\times n}$ und $\Delta > 0$ gegeben sind. Sei der sogenannte Cauchy-Punkt definiert durch

$$p^C := -\tau \frac{\Delta}{\|g\|_2} g,$$

wobei

$$\tau := \begin{cases} 1, & \text{falls} \quad g^T B g \le 0, \\ \min(\|g\|_2^3/(\Delta g^T B g), 1), & \text{sonst.} \end{cases}$$

Man zeige, dass

$$f - \phi(p^C) \ge \frac{1}{2} \|g\|_2 \min\left(\Delta, \frac{\|g\|_2}{\|B\|_2}\right).$$

Der Cauchy-Punkt p^C genügt also derselben Abschätzung wie eine globale Lösung p^* von (P), siehe Lemma 2.5. Weiter zeige man, dass p^C Lösung der Aufgabe

$$\text{Minimiere} \quad \phi(p), \quad \|p\|_2 \le \Delta, \; p \in \operatorname{span}\{g\}$$

ist

7. Seien $f \in \mathbb{R}$, $g \in \mathbb{R}^n \setminus \{0\}$, die symmetrische und positiv definite Matrix $B \in \mathbb{R}^{n\times n}$ und $\Delta > 0$ gegeben. Ferner sei $\phi\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ durch

$$\phi(p) := f + g^T p + \frac{1}{2} p^T B p$$

definiert. Sei

$$p^B := -B^{-1} g$$

(unrestringiertes Minimum von ϕ) und

$$p^U := -\frac{\|g\|_2^2}{g^T B g} g$$

(unstringiertes Minimum von ϕ in $\operatorname{span}\{g\}$). Wir nehmen an, es sei $p^B \neq p^U$. Man definiere $\tilde{p}\colon [0,2] \longrightarrow \mathbb{R}^n$ durch

$$\tilde{p}(\tau) := \begin{cases} \tau p^U, & 0 \leq \tau \leq 1, \\ p^U + (\tau - 1)(p^B - p^U), & 1 \leq \tau \leq 2. \end{cases}$$

Man zeige:

   (a) $\|\tilde{p}(\cdot)\|_2$ ist strikt monoton wachsend,

   (b) $\phi(\tilde{p}(\cdot))$ ist strikt monoton fallend.

   (c) Sei $\|p^B\|_2 > \Delta$ (andernfalls ist p^B die Lösung des Trust-Region-Hilfsproblems). Man zeige, dass es genau ein $\tau^* \in (0,2)$ mit $\|\tilde{p}(\tau^*)\|_2 = \Delta$ gibt[6].

   (d) Sei $\|p^B\|_2 > \Delta$. Man zeige, dass $\phi(\tilde{p}(\tau^*)) \leq \phi(p^C)$, wobei p^C der in Aufgabe 6 definierte Cauchy-Punkt ist. Insbesondere gilt auch für $\tilde{p}(\tau^*)$ die Abschätzung aus Lemma 2.5, also

$$f - \phi(\tilde{p}(\tau^*)) \geq \frac{1}{2}\|g\|_2 \min\left(\Delta, \frac{\|g\|_2}{\|B\|_2}\right).$$

8. Seien $f \in \mathbb{R}$, $g \in \mathbb{R}^n \setminus \{0\}$, die symmetrische und positiv definite Matrix $B \in \mathbb{R}^{n\times n}$ und $\Delta > 0$ gegeben. Man gebe ein Verfahren zur Berechnung der Lösung p^M von

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T B p \quad \text{auf} \\ M := \{p \in \mathbb{R}^n : p \in \operatorname{span}\{g, B^{-1}g\},\ \|p\|_2 \leq \Delta\} \end{cases}$$

an und zeige, dass

$$f - \phi(p^M) \geq \frac{1}{2}\|g\|_2 \min\left(\Delta, \frac{\|g\|_2}{\|B\|_2}\right).$$

Hinweis: Für den letzten Teil der Aufgabe kann man die Aussage von Aufgabe 6 verwenden.

[6] Die Bestimmung von $\tilde{p}(\tau^*)$ als Näherungslösung des Trust-Region-Hilfsproblems wird in der englischsprachigen Literatur als "dog leg method" bezeichnet.

9. In Aufgabe 1 sollte die exakte Lösung p^* von

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2} p^T B p, \quad \|p\|_2 \leq \Delta,$$

berechnet werden, wobei

$$g := \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad B := \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \Delta := \frac{1}{2}.$$

Zum Vergleich berechne man den Cauchy-Punkt p^C (siehe Aufgabe 6) und den Dog Leg Punkt $\tilde{p}(\tau^*)$ (siehe Aufgabe 7). Man vergleiche die Werte $\phi(p^*)$, $\phi(p^C)$ und $\phi(\tilde{p}(\tau^*))$. Schließlich plotte man den "Dog Leg Pfad" $\{\tilde{p}(\tau) : \tau \in [0,2]\}$ und (gestrichelt) den optimalen Pfad $\{p^*(\Delta) : \Delta \in [0,1]\}$, wobei $p^*(\Delta)$ die Lösung von (P) (mit variablem Δ) ist.

4.3 Trust-Region-Verfahren bei nichtlinearen Approximationsaufgaben

In diesem Abschnitt betrachten wir die (unrestringierte) nichtlineare Approximationsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n,$$

wobei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ als glatt, etwa stetig partiell differenzierbar, vorausgesetzt wird und $\|\cdot\|$ eine Norm auf dem $\mathbb{R}^m$ ist. Besonders beschäftigen wird uns der Fall, dass es sich bei der Norm um die euklidische Norm bzw. es sich bei (P) um ein *nichtlineares Ausgleichsproblem* (nonlinear least square problem) handelt. In (P) wird grundsätzlich $m \geq n$ angenommen, so dass man (P) auch als die Aufgabe auffassen kann, ein i. Allg. überbestimmtes nichtlineares Gleichungssystem so zu "lösen", dass der Defekt bezüglich einer gegebenen Norm minimal ist. Definiert man $g\colon \mathbb{R}^m \longrightarrow \mathbb{R}$ durch $g(y) := \|y\|$, so erkennt man, dass $f = g \circ F$, die Zielfunktion f also Komposition der konvexen, global lipschitzstetigen Abbildung g und der glatten Abbildung F ist. Viele der späteren Aussagen können auf sogenannte halbglatte Aufgaben übertragen werden, was uns hier aber nicht besonders interessieren soll. Für uns ist wichtig, dass f in jedem x_c (in dem F differenzierbar ist) eine Richtungsableitung bzw. Gateaux-Variation $f'(x_c;\cdot)\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ besitzt und diese nach der Kettenregel durch

$$f'(x_c; p) = g'(F(x_c); F'(x_c)p)$$

gegeben ist. Hierbei ist $g'(F(x);\cdot)$ die Gateaux-Variation von g in $F(x)$, die wegen der Konvexität von g existiert. Zur Erinnerung: Wir nennen ein $x^* \in \mathbb{R}^n$ eine *stationäre Lösung* von (P), wenn $f'(x^*; p) \geq 0$ für alle $p \in \mathbb{R}^n$. Wegen

$$f'(x^*; p) := \lim_{t \to 0+} \frac{f(x^* + tp) - f(x^*)}{t}$$

bedeutet dies, dass es in x^* keine Abstiegsrichtung gibt.

In dem in Abschnitt 4.1 angegebenen Modellalgorithmus für Trust-Region-Verfahren wählen wir als Modellfunktion in einer aktuellen Näherung $x_c \in \mathbb{R}^n$ die durch

$$f_c(p) := \|F(x_c) + F'(x_c)p\|$$

definierte Funktion f_c.

4.3.1 Globale Konvergenzaussagen

Unser Ziel in diesem Unterabschnitt ist es, eine (schwache) globale Konvergenzaussage für das auf die nichtlineare Approximationsaufgabe (P) angewandte Trust-Regin-Verfahren zu beweisen.

Das folgende Lemma wird als Rechtfertigung für eine STOP-Bedingung dienen. Hierbei machen wir eine geringfügige Bezeichnungsänderung.

Lemma 3.1 *Gegeben sei die nichtlineare Approximationsaufgabe (P). Die Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ sei in $x \in \mathbb{R}^n$ stetig partiell differenzierbar. Mit einem vorgegebenen $\Delta > 0$ sei $p^* \in \mathbb{R}^n$ eine Lösung von*

$$(\mathrm{P}_{x,\Delta}) \qquad \textit{Minimiere} \quad f_x(p) := \|F(x) + F'(x)p\|, \quad \|p\| \le \Delta.$$

Dann ist x genau dann eine stationäre Lösung von (P), wenn $f(x) = f_x(p^*)$.

Beweis: Sei x eine stationäre Lösung von (P). Insbesondere ist

$$\begin{aligned} 0 &\le f'(x;p^*) \\ &= g'(F(x);F'(x)p^*) \\ &\le \|F(x) + F'(x)p^*\| - \|F(x)\| \\ &= f_x(p^*) - f(x) \\ &\le 0, \end{aligned}$$

also $f(x) = f_x(p^*)$. Sei umgekehrt $f(x) = f_x(p^*)$. Dann ist auch $p^{**} := 0$ eine Lösung von $(\mathrm{P}_{x,\Delta})$ und daher auch der Aufgabe, $f_x(\cdot)$ auf dem $\mathbb{R}^n$ zu minimieren (siehe Aufgabe 1). Für alle $p \in \mathbb{R}^n$ ist daher

$$0 \le f'_x(0;p) = \lim_{t\to 0+} \frac{g(F(x) + tF'(x)p) - g(F(x))}{t} = g'(F(x);F'(x)p) = f'(x;p).$$

Also ist x eine stationäre Lösung von (P). □ □

Es ist zweckmäßig, jetzt eine Bezeichnung einzuführen. Bei gegebenen $x \in \mathbb{R}^n$ und $\Delta > 0$ sei

$$v(x,\Delta) := \min\{\|F(x) + F'(x)p\| : \|p\| \le \Delta\},$$

also $v(x,\Delta)$ der Wert von

$$(\mathrm{P}_{x,\Delta}) \qquad \text{Minimiere} \quad f_x(p) := \|F(x) + F'(x)p\|, \quad \|p\| \le \Delta.$$

Ist $x^* \in \mathbb{R}^n$ keine Lösung von (P), so ist $f(x^*) - v(x^*,\Delta) > 0$ für alle $\Delta > 0$ (siehe Lemma 3.1). Das folgende Lemma impliziert insbesondere, dass alle Punkte aus einer hinreichend kleinen Umgebung eines nichtstationären Punktes ebenfalls nichtstationär sind.

Lemma 3.2 *Gegeben sei die nichtlineare Approximationsaufgabe (P). Ein $\Delta^* > 0$ sei vorgegeben. Dann gilt:*

1. *Ist F in $x \in \mathbb{R}^n$ stetig partiell differenzierbar, so ist*

$$f(x) - v(x, \Delta) \geq \frac{\Delta}{\Delta^*}[f(x) - v(x, \Delta^*)] \qquad \text{für alle } \Delta \in (0, \Delta^*].$$

2. *Ist F in $x^* \in \mathbb{R}^n$ stetig partiell differenzierbar und x^* keine stationäre Lösung von (P), so gibt es ein $\delta > 0$ mit*

$$f(x) - v(x, \Delta^*) \geq \frac{1}{2}[f(x^*) - v(x^*, \Delta^*)] \qquad \text{für alle } x \text{ mit } \|x - x^*\| \leq \delta.$$

Beweis: Sei $\Delta \in (0, \Delta^*]$ und p^* eine Lösung von $(\mathrm{P}_{x,\Delta^*})$. Dann ist $p := \lambda p^*$ mit $\lambda := \Delta/\Delta^*$ zulässig für $(\mathrm{P}_{x,\Delta})$ und daher

$$\begin{aligned} f(x) - v(x, \Delta) &\geq \|F(x)\| - \|F(x) + \lambda F'(x)p^*\| \\ &\geq \|F(x)\| - [(1-\lambda)\|F(x)\| + \lambda\|F(x) + F'(x)p^*\|] \\ &= \lambda[\|F(x)\| - \|F(x) + F'(x)p^*\|] \\ &= \frac{\Delta}{\Delta^*}[f(x) - v(x, \Delta^*)], \end{aligned}$$

womit der erste Teil des Lemmas bewiesen ist.

Sei p^* eine Lösung von $(\mathrm{P}_{x^*,\Delta^*})$. Die Abbildung

$$x \mapsto \|F(x)\| - \|F(x) + F'(x)p^*\|$$

ist in x^* stetig, dort hat sie den positiven Wert $f(x^*) - v(x^*, \Delta^*) > 0$, da x^* nach Voraussetzung keine stationäre Lösung von (P) ist. Daher gibt es ein $\delta > 0$ derart, dass

$$f(x) - v(x, \Delta^*) \geq \|F(x)\| - \|F(x) + F'(x)p^*\| \geq \frac{f(x^*) - v(x^*, \Delta^*)}{2}$$

für alle x mit $\|x - x^*\| \leq \delta$. Damit ist auch der zweite Teil des Lemmas bewiesen. □□

In dem folgenden Satz geben wir das Madsen-Verfahren (siehe K. Madsen (1975)[7], dort ist allerdings $\|\cdot\| := \|\cdot\|_\infty$) an und beweisen (unter unnötig starken Voraussetzungen an die Zielfunktion) eine (schwache) Konvergenzaussage.

Satz 3.3 *Gegeben sei die Approximationsaufgabe*

(P) $$\text{Minimiere} \quad f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n.$$

Mit dem Startwert x_0 des gleich anzugebenden Verfahrens sei die Niveaumenge $L_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ kompakt. Die Abbildung $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ sei auf einer Umgebung von L_0 stetig partiell differenzierbar, die Funktionalmatrix $F'(\cdot)$ sei dort lipschitzstetig. Man betrachte das folgende Verfahren:

[7] Madsen, K. (1975) An algorithm for minimax solution of overdetermined systems of non-linear equations. J. Inst. Maths. Applcs. 16, 321–328.

- *Gegeben seien Konstanten* $0 < \rho_1 < \rho_2 < 1$ *und* $0 < \sigma_1 < 1 < \sigma_2$.

  (Bei K. MADSEN *(1975) wird* $\rho_1 := 0.01$, $\rho_2 := 0.25$, $\sigma_1 := 0.25$ *und* $\sigma_2 := 2$ *empfohlen.)*

- *Gegeben seien* $x_0 \in \mathbb{R}^n$ *und* $\Delta_0 > 0$. *Berechne* $F(x_0)$ *und* $F'(x_0)$.

- *Für* $k = 0, 1, \ldots$:

  - *Berechne eine Lösung* p_k *der Aufgabe*

    $$(\mathrm{P}_k) \qquad \textit{Minimiere} \quad f_k(p) := \|F(x_k) + F'(x_k)p\|, \quad \|p\| \le \Delta_k.$$

  - *Falls* $f(x_k) = f_k(p_k)$, *dann: STOP,* x_k *ist eine stationäre Lösung von (P).*
  - *Berechne*

    $$r_k := \frac{f(x_k) - f(x_k + p_k)}{f(x_k) - f_k(p_k)}.$$

  - *Falls* $r_k \ge \rho_1$, *dann setze* $x_{k+1} := x_k + p_k$ *und berechne* $F(x_{k+1})$ *und* $F'(x_{k+1})$. *Andernfalls setze* $x_{k+1} := x_k$.
  - *Falls* $r_k \le \rho_2$, *dann setze* $\Delta_{k+1} := \sigma_1 \|p_k\|$, *andernfalls wähle* $\Delta_{k+1} \in [\|p_k\|, \sigma_2 \|p_k\|]$.
    (Bei K. MADSEN *(1975) wird für* $r_k > \rho_2$ *ein weiterer Test gemacht. Ist*

    $$\|F(x_k + p_k) - F(x_k) - F'(x_k)p_k\| \le \rho_3 [f(x_k) - f(x_k + p_k)]$$

    mit einem $\rho_3 \in (0, 1)$, *etwa* $\rho_3 := 0.25$, *so wird* $\Delta_{k+1} := \sigma_2 \|p_k\|$ *gesetzt, andernfalls* $\Delta_{k+1} := \|p_k\|$.*)*

Das Verfahren breche nicht vorzeitig mit einer stationären Lösung von (P) ab. Dann liefert es eine Folge $\{x_k\}$ *mit der Eigenschaft, dass jeder Häufungspunkt von* $\{x_k\}$ *eine stationäre Lösung von (P) ist.*

Beweis: Zunächst beachten wir, dass die Folge $\{\Delta_k\} \subset \mathbb{R}_+$ beschränkt ist, da es die Folge $\{x_k\} \subset L_0$ ist, und Δ_{k+1} im k-ten Iterationsschritt nur dann vergrößert werden kann, wenn $r_k > \rho_2$ und damit $\Delta_{k+1} \le \sigma_2 \|p_k\| = \sigma_2 \|x_{k+1} - x_k\|$ gilt. Daher ist $\Delta_k \le \max(\Delta_0, \sigma_2 d)$, wobei $d := \sup_{x,y \in L_0} \|x - y\|$ den Durchmesser von L_0 bedeutet. Daher existiert ein $\Delta^* > 0$ mit $\{\Delta_k\} \subset (0, \Delta^*]$. Der Beweis des Satzes zerfällt in zwei Teile.

(1) Konvergiert die durch das Verfahren erzeugte Folge $\{x_k\}$ gegen einen Punkt x^*, so ist x^* eine stationäre Lösung von (P).

Zum Beweis dieser Zwischenbehauptung nehmen wir an, x^* sei keine stationäre Lösung. Wegen $\lim_{k\to\infty} x_k = x^*$ und Lemma 3.2 existiert eine positive Konstante c_0 (z. B. kann $c_0 := [f(x^*) - v(x^*, \Delta^*)]/(2\Delta^*)$ gesetzt werden) mit

$$(*) \qquad f(x_k) - f_k(p_k) \ge c_0 \Delta_k \qquad \text{für alle hinreichend großen } k.$$

Hieraus wollen wir schließen (siehe auch einen entsprechenden Schluss beim Beweis von Satz 2.6), dass $\sum_{k=0}^{\infty} \Delta_k < \infty$. Dies ist trivialerweise richtig, wenn es nur endlich viele erfolgreiche Iterationsschritte gibt, denn dann ist $\Delta_{k+1} = \sigma_1 \|p_k\| \leq \sigma_1 \Delta_k$ mit $\sigma_1 \in (0,1)$ für alle hinreichend großen k. Daher nehmen wir jetzt an, dass es unendlich viele erfolgreiche Iterationsschritte gibt, diese seien in der Indexmenge E zusammengefaßt. Ist $k \in E$, so ist

$$f(x_k) - f(x_{k+1}) \geq \rho_1 [f(x_k) - f_k(p_k)] \geq \rho_1 c_0 \Delta_k.$$

Da $\{f(x_k)\}$ eine monoton nicht wachsende, nach unten beschränkte Folge ist, erhalten wir hieraus

$$\rho_1 c_0 \sum_{k \in E} \Delta_k \leq \sum_{k \in E} [f(x_k) - f(x_{k+1})] \leq \sum_{k=0}^{\infty} [f(x_k) - f(x_{k+1})] < \infty$$

und damit $\sum_{k \in E} \Delta_k < \infty$. Nun betrachten wir die k, die zwischen zwei aufeinanderfolgenden erfolgreichen Iterationsschritten $i < j$ liegen. Dann ist $\Delta_{i+1} \leq \sigma_2 \|p_i\| \leq \sigma_2 \Delta_i$ (da $i \in E$) und $\Delta_{k+1} \leq \sigma_1 \Delta_k$, $k = i+1, \ldots, j-1$. Hieraus folgt durch Zurückspulen und Summieren

$$\sum_{k=i+1}^{j-1} \Delta_k \leq \frac{\sigma_2}{1 - \sigma_1} \Delta_i.$$

Insgesamt ist damit $\sum_{k=0}^{\infty} \Delta_k < \infty$ bewiesen. Insbesondere ist $\{\Delta_k\}$ und damit auch $\{\|p_k\|\}$ eine Nullfolge. Zum Beweis von (1) machen wir eine Fallunterscheidung.

(i) Es gibt eine unendliche Teilmenge $K \subset \mathbb{N}$ mit $\|p_k\| < \Delta_k$ für alle $k \in K$.

Sei $k \in K$ fest. Wir überlegen[8] uns, dass $f_k(p_k) \leq f_k(p)$ für alle $p \in \mathbb{R}^n$. Denn nach Definition von p_k ist trivialerweise $f_k(p_k) \leq f_k(p)$ für alle p mit $\|p\| \leq \Delta_k$. Ist dagegen $\|p\| > \Delta_k$, so bestimme man ein $\lambda \in (0,1)$ mit $\|(1-\lambda)p_k + \lambda p\| = \Delta_k$, nutze die Konvexität von f_k aus:

$$f_k(p_k) \leq f_k((1-\lambda)p_k + \lambda p) \leq (1-\lambda) f_k(p_k) + \lambda f_k(p)$$

und schließe hieraus auf $f_k(p_k) \leq f_k(p)$. Lässt man $k \in K$ gegen unendlich streben, so folgt wegen $x_k \to x^*$ und $p_k \to 0$, daß

$$\|F(x^*)\| = f(x^*) \leq f_{x^*}(p) = \|F(x^*) + F'(x^*)p\| \qquad \text{für alle } p \in \mathbb{R}^n.$$

Wegen Lemma 3.1 ist x^* eine stationäre Lösung von (P), ein Widerspruch zu der Annahme, dass das gerade nicht der Fall ist.

(ii) Für alle hinreichend großen k ist $\|p_k\| = \Delta_k$.

Wir zeigen, dass $\lim_{k \to \infty} r_k = 1$. Dann ist $r_k > \rho_2$ für alle hinreichend großen k und daher $\|p_{k+1}\| = \Delta_{k+1} \geq \|p_k\|$ für alle hinreichend großen k, ein Widerspruch zu $p_k \to 0$.

[8] Siehe auch Aufgabe 1.

Wegen $(*)$ und (ii) ist für alle hinreichend großen k

$$\begin{aligned} |r_k - 1| &= \frac{|\|F(x_k + p_k)\| - \|F(x_k) + F'(x_k)p_k\||}{f(x_k) - f_k(p_k)} \\ &\leq \frac{\|F(x_k + p_k) - F(x_k) - F'(x_k)p_k\|}{c_0 \|p_k\|} \\ &\leq \frac{1}{c_0} \int_0^1 \|F'(x_k + tp_k) - F'(x_k)\| \, dt. \end{aligned}$$

Wegen $p_k \to 0$ und $x_k \to x^*$ folgt $\lim_{k\to\infty} r_k = 1$. Damit ist (1) bewiesen.

(2) Jeder Häufungspunkt x^* der Folge $\{x_k\}$ ist eine stationäre Lösung von (P).

Angenommen, x^* sei ein Häufungspunkt der Folge $\{x_k\}$, der keine stationäre Lösung von (P) ist. Eine Anwendung von Lemma 3.2 zeigt die Existenz positiver Konstanten δ und c_0 mit

$$\|x_k - x^*\| \leq \delta \Longrightarrow f(x_k) - f_k(p_k) \geq c_0 \|p_k\|.$$

Wegen Beweisteil (1) wissen wir, dass nicht die gesamte Folge $\{x_k\}$ gegen x^* konvergiert, so dass angenommen werden kann, dass es unendlich viele k mit $\|x_k - x^*\| > \delta$ gibt (notfalls verkleinere man δ). Da ferner x^* ein Häufungspunkt von $\{x_k\}$ ist, gibt es in jeder Umgebung von x^*, insbesondere in der $\delta/2$-Kugel um x^*, unendlich viele Elemente von x_k. Daher existieren Folgen $\{k(i)\}_{i\in\mathbb{N}}$ und $\{l(i)\}_{i\in\mathbb{N}}$ natürlicher Zahlen mit $k(i) < l(i) < k(i+1) < l(i+1)$ und

$$\|x_{k(i)} - x^*\| \leq \frac{\delta}{2}, \qquad \|x_k - x^*\| \leq \delta \quad \text{für } k(i) < k < l(i), \qquad \|x_{l(i)} - x^*\| > \delta.$$

Hieraus folgt

$$f(x_k) - f(x_{k+1}) \geq \rho_1 c_0 \|x_{k+1} - x_k\|, \qquad k(i) \leq k < l(i),$$

und damit

$$\begin{aligned} \rho_1 c_0 \|x_{k(i)} - x_{l(i)}\| &\leq \rho_1 c_0 \sum_{k=k(i)}^{l(i)-1} \|x_{k+1} - x_k\| \\ &\leq \sum_{k=k(i)}^{l(i)-1} [f(x_k) - f(x_{k+1})] \\ &= f(x_{k(i)}) - f(x_{l(i)}). \end{aligned}$$

Da die Folge $\{f(x_k)\}$ konvergiert, insbesondere eine Cauchy-Folge ist, konvergiert die rechte Seite und damit auch $\{\|x_{k(i)} - x_{l(i)}\|\}$ gegen Null. Andererseits ist nach Konstruktion

$$\|x_{k(i)} - x_{l(i)}\| \geq \|x_{l(i)} - x^*\| - \|x_{k(i)} - x^*\| > \delta - \frac{\delta}{2} = \frac{\delta}{2},$$

ein Widerspruch. Der Satz ist damit bewiesen. □ □

Bemerkung: Aus dem letzten Satz folgt natürlich die Konvergenz der gesamten durch das Verfahren erzeugten Folge $\{x_k\}$, wenn es in der Niveaumenge L_0 nur eine stationäre Lösung von (P) gibt. □

4.3.2 Superlineare Konvergenz

Wir erinnern zunächst an die Definition 4.5 in Abschnitt 3.4, wo definiert wurde, wann eine Lösung einer nichtlinearen Approximationsaufgabe lokal stark eindeutig heißt. Um das Blättern zu vermeiden, wiederholen wir sie hier:

Definition 3.4 Gegeben sei die nichtlineare Approximationsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n.$$

Ein $x^* \in \mathbb{R}^n$ heißt *lokal stark eindeutige Lösung* von (P), wenn es positive Konstanten σ und δ mit

$$f(x) \geq f(x^*) + \sigma\|x - x^*\| \qquad \text{für alle } x \in \mathbb{R}^n \text{ mit } \|x - x^*\| \leq \delta$$

gibt.

Nun folgt eine Aussage zur superlinearen Konvergenz des Madsen-Verfahrens.

Satz 3.5 *Das in Satz 3.3 angegebene Verfahren zur Lösung der nichtlinearen Approximationsaufgabe*

$$\text{(P)} \qquad \textit{Minimiere} \quad f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n$$

liefere eine Folge $\{x_k\}$, die gegen ein $x^* \in \mathbb{R}^n$ konvergiert. Sei x^* lokal stark eindeutige Lösung von (P) und $F'(\cdot)$ auf einer Umgebung von x^* lipschitzstetig. Dann sind fast alle Iterationsschritte erfolgreich und die Folge $\{x_k\}$ konvergiert sogar quadratisch gegen x^*, d. h. es existiert eine Konstante $C > 0$ mit

$$\|x_{k+1} - x^*\| \leq C\,\|x_k - x^*\|^2 \qquad \textit{für alle hinreichend großen } k.$$

Beweis: Es kann angenommen werden, dass die wegen der lokalen starken Eindeutigkeit von x^* existierende Konstante $\delta > 0$ so klein ist, dass $F'(\cdot)$ auf der Kugel $B[x^*;\delta] := \{x \in \mathbb{R}^n : \|x - x^*\| \leq \delta\}$ lipschitzstetig mit einer Lipschitzkonstanten $L > 0$ ist. Mit positiven Konstanten σ und δ ist also

$$f(x) \geq f(x^*) + \sigma\,\|x - x^*\|, \qquad \|F'(x) - F'(y)\| \leq L\,\|x - y\|$$

für alle $x, y \in B[x^*;\delta]$. O. B. d. A. bricht das Verfahren nicht vorzeitig ab, so dass $x_k \neq x^*$ für alle k angenommen werden kann. Wir erinnern an das beim Madsen-Verfahren auftretende Hilfsproblem

$$(\text{P}_k) \qquad \text{Minimiere} \quad f_k(p) := \|F(x_k) + F'(x_k)p\|, \quad \|p\| \leq \Delta_k$$

mit Lösung p_k. Der Beweis des Satzes zerfällt in mehrere Teile.

(1) Sind $x_k, x_k + p \in B[x^*;\delta]$, so ist

$$|f(x_k + p) - f_k(p)| \leq \frac{L}{2}\,\|p\|^2.$$

Denn: Es ist

$$\begin{aligned}
|f(x_k+p)-f_k(p)| &= |\,\|F(x_k+p)\| - \||F(x_k)+F'(x_k)p\|\,| \\
&\le \|F(x_k+p)-F(x_k)-F'(x_k)p\| \\
&\le \int_0^1 \|[F'(x_k+tp)-F'(x_k)]p\|\,dt \\
&\le \frac{L}{2}\|p\|^2.
\end{aligned}$$

Damit ist (1) bewiesen.

(2) Es ist $\|p_k\| \le 2\,\|x_k - x^*\|$ für alle hinreichend großen k.

Denn: Wegen der vorausgesetzten Konvergenz der Folge $\{x_k\}$ gegen x^* existiert ein $k_0 \in \mathbb{N}$ mit

$$\|x_k - x^*\| \le \frac{\delta}{3}, \quad \sigma - \frac{5L}{2}\|x_k - x^*\| > 0 \qquad \text{für alle } k \ge k_0.$$

Wir wollen zeigen, dass $\|p_k\| \le 2\,\|x_k - x^*\|$ für alle $k \ge k_0$. Angenommen, es wäre $\|p_k\| > 2\,\|x_k - x^*\|$ für ein $k \ge k_0$. Dann ist $\|x_k - x^*\| < \frac{1}{2}\|p_k\| < \Delta_k$, also $x^* - x_k$ zulässig für (P_k). Wir zeigen, dass $f_k(p_k) > f_k(x^* - x_k)$, was ein Widerspruch dazu bedeutet, dass p_k das Hilfsproblem (P_k) löst. Hierzu bestimme man $\lambda_k \in (0,1)$ so, dass

$$\|(1-\lambda_k)p_k + \lambda_k(x^* - x_k)\| = 2\,\|x_k - x^*\|$$

und setze anschließend

$$q_k := (1-\lambda_k)p_k + \lambda_k(x^* - x_k).$$

Dann ist

$$\begin{aligned}
\underbrace{(1-\lambda_k)}_{>0}[f_k(p_k) - f_k(x^*-x_k)] &\ge f_k(q_k) - f_k(x^*-x_k) \\
&\qquad \text{(Konvexität von } f_k) \\
&= f_k(q_k) - f(x_k+q_k) + f(x_k+q_k) - f_k(x^*-x_k) \\
&\ge -\frac{L}{2}\|q_k\|^2 + f(x_k+q_k) - f_k(x^*-x_k) \\
&\qquad \text{(Anwendung von (1))} \\
&\ge -2L\,\|x_k - x^*\|^2 + \sigma\|x_k + q_k - x^*\| \\
&\qquad + f(x^*) - f_k(x^*-x_k) \\
&\qquad \text{(lokal starke Eindeutigkeit)} \\
&\ge -2L\,\|x_k - x^*\|^2 + \sigma\|x_k + q_k - x^*\| - \frac{L}{2}\|x_k - x^*\|^2 \\
&\qquad \text{(Anwendung von (1))} \\
&= -\frac{5L}{2}\|x_k - x^*\|^2 + \sigma\|x_k + q_k - x^*\| \\
&\ge -\frac{5L}{2}\|x_k - x^*\|^2 + \sigma(\|q_k\| - \|x_k - x^*\|) \\
&= \Big(\sigma - \frac{5L}{2}\|x_k - x^*\|\Big)\|x_k - x^*\| \\
&> 0,
\end{aligned}$$

also $f_k(p_k) > f_k(x^* - x_k)$, womit der gewünschte Widerspruch erreicht ist. Damit ist (2) bewiesen.

(3) Es existiert eine Konstante $c_0 > 0$ derart, dass $f(x_k) - f_k(p_k) \geq c_0 \|p_k\|$ für alle hinreichend großen k.

Denn: Sei $k_0 \in \mathbb{N}$ wie im Beweis von (2) bestimmt. Ist $\|x_k - x^*\| \leq \Delta_k$, also $x_k - x^*$ zulässig für (P_k), und $k \geq k_0$, so ist wegen (1) und (2)

$$\begin{aligned} f(x_k) - f_k(p_k) &\geq f(x_k) - f_k(x^* - x_k) \\ &= f(x_k) - f(x^*) + f(x^*) - f_k(x^* - x_k) \\ &\geq \sigma \|x_k - x^*\| - \frac{L}{2} \|x_k - x^*\|^2 \\ &\geq \frac{4\sigma}{5} \|x_k - x^*\| \\ &\geq \frac{2\sigma}{5} \|p_k\|. \end{aligned}$$

Ist dagegen $\|x_k - x^*\| > \Delta_k$ bzw. $\lambda_k := \Delta_k / \|x_k - x^*\| \in (0,1)$ und ist außerdem $k \geq k_0$, so folgt aus der Konvexität von f_k, daß

$$\begin{aligned} f(x_k) - f_k(p_k) &= f_k(0) - f_k(p_k) \\ &\geq f_k(0) - f_k(\lambda_k(x^* - x_k)) \\ &= f_k(0) - f_k((1-\lambda_k)0 + \lambda_k(x^* - x_k)) \\ &\geq f_k(0) - [(1-\lambda_k) f_k(0) + \lambda_k f_k(x^* - x_k)] \\ &= \frac{\Delta_k}{\|x_k - x^*\|} [f(x_k) - f_k(x^* - x_k)] \\ &\geq \frac{4\sigma}{5} \Delta_k \\ &\geq \frac{4\sigma}{5} \|p_k\|. \end{aligned}$$

Damit ist auch (3) bewiesen.

(4) Mit

$$r_k := \frac{f(x_k) - f(x_k + p_k)}{f(x_k) - f_k(p_k)}$$

ist $\lim_{k\to\infty} r_k = 1$. Insbesondere ist $r_k \geq \rho_1$ für alle hinreichend großen k, folglich sind fast alle Iterationsschritte erfolgreich und damit $x_{k+1} = x_k + p_k$ für alle hinreichend großen k.

Denn: Wegen (1)–(3) ist

$$|r_k - 1| = \frac{|f(x_k + p_k) - f_k(p_k)|}{f(x_k) - f_k(p_k)} \leq \frac{L}{2c_0} \|p_k\| \leq \frac{L}{c_0} \|x_k - x^*\|$$

für alle hinreichend großen k. Wegen $\lim_{k\to\infty} x_k = x^*$ folgt (4).

(5) Es ist

$$(*) \qquad f_k(p_k) \le f_k(x^* - x_k)$$

und

$$(**) \qquad \|x_{k+1} - x^*\| \le \frac{5L}{2\sigma} \|x_k - x^*\|^2$$

für alle hinreichend großen k. Insbesondere konvergiert die Folge $\{x_k\}$ von mindestens zweiter Ordnung gegen x^*.

Denn: Sei $k_0 \in \mathbb{N}$ so groß gewählt, daß

$$\|x_k - x^*\| \le \min\Big(\delta, \frac{\sigma}{5L}\Big), \quad \|p_k\| \le 2\|x_k - x^*\|, \quad r_k \ge \rho_2 \qquad \text{für alle } k \ge k_0,$$

was wegen $x_k \to x^*$ sowie (2) und (4) möglich ist. Insbesondere ist $x_{k+1} = x_k + p_k$ und $\Delta_{k+1} \ge \|p_k\|$ für alle $k \ge k_0$. Letzteres liefert zusammen mit $p_k \to 0$ die Existenz eines $l \ge k_0$ mit $\|p_l\| < \Delta_l$. Wegen der Konvexität von f_l ist p_l nicht nur Minimum von f_l auf der Kugel $B[0; \Delta_l]$, sondern auf dem gesamten $\mathbb{R}^n$ (siehe Aufgabe 1). Daher ist speziell $(*)$ für $k = l$ richtig. Hieraus folgt aber auch die Richtigkeit von $(**)$ für $k = l$. Denn es ist

$$\begin{aligned}
\|x_{l+1} - x^*\| &\le \frac{1}{\sigma}\,[f(x_l + p_l) - f(x^*)] \\
&= \frac{1}{\sigma}\,[f_l(x_l + p_l) - f_l(p_l) + f_l(p_l) - f(x^*)] \\
&\le \frac{1}{\sigma}\,[2L\,\|x_l - x^*\|^2 + f_l(x^* - x_l) - f(x^*)] \\
&\le \frac{5L}{2\sigma}\,\|x_l - x^*\|^2.
\end{aligned}$$

Damit ist $(**)$ für $k = l$ richtig. Nach Wahl von k_0 folgt

$$\|x_{l+1} - x^*\| \le \frac{1}{2}\,\|x_l - x^*\| \le \frac{1}{2}\,[\|p_l\| + \|x_{l+1} - x^*\|],$$

damit

$$\|x_{l+1} - x^*\| \le \|p_l\| \le \Delta_{l+1}$$

und hieraus schließlich $(*)$ für $k = l + 1$. In dieser Weise kann man fortfahren und erhält, dass $(*)$ und $(**)$ für alle $k \ge l$ gelten. Der Satz ist bewiesen. □ □

4.3.3 Nichtlineare Ausgleichsprobleme

In diesem Unterabschnitt betrachten wir das nichtlineare Ausgleichsproblem

$$(\mathrm{P}) \qquad \text{Minimiere} \quad f(x) := \|F(x)\|_2, \quad x \in \mathbb{R}^n,$$

wobei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ hinreichend glatt sei. Auf (P) kann das Madsen-Verfahren aus Satz 3.3 angewandt werden. Die wesentliche Arbeit besteht in der Lösung des hierbei

auftretenden in einer aktuellen Lösung linearisierten Hilfsproblems. Da wir auf eine Skalierung verzichten wollen, lautet dieses:

$$(\mathrm{P}_{x,\Delta}) \qquad \text{Minimiere} \quad f_x(p) := \|F(x) + F'(x)p\|_2, \quad \|p\|_2 \le \Delta.$$

Diese Aufgabe ist natürlich äquivalent zu

$$\text{Minimiere} \quad \frac{1}{2}\|F(x)\|_2^2 + [F'(x)^T F(x)]^T p + \frac{1}{2} p^T F'(x)^T F'(x) p, \quad \|p\|_2 \le \Delta.$$

Da $B := F'(x)^T F'(x)$ positiv semidefinit ist, braucht nicht zwischen lokaler und globaler Lösung von $(\mathrm{P}_{x,\Delta})$ unterschieden zu werden. Aus Satz 2.1 erhalten wir damit:

- *Genau dann ist ein $p^* \in \mathbb{R}^n$ mit $\|p^*\|_2 \le \Delta$ eine Lösung von $(P_{x,\Delta})$, wenn ein $\lambda^* \ge 0$ mit*

  $$[F'(x)^T F'(x) + \lambda^* I]p^* = -F'(x)^T F(x), \qquad \lambda^*(\Delta - \|p^*\|_2) = 0$$

  existiert.

Im folgenden werden wir die Existenz einer *Singulärwertzerlegung* benutzen. Genauer gilt die folgende Aussage:

- Sei $A \in \mathbb{R}^{m\times n}$ mit $m \ge n$ gegeben. Dann existieren orthogonale Matrizen $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ und eine Diagonalmatrix $\hat{\Sigma} := \mathrm{diag}\,(\sigma_1, \ldots, \sigma_n)$ mit $\sigma_1 \ge \cdots \ge \sigma_n$ derart, dass

  $$A = U \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T.$$

  Die Anzahl r der positiven sogenannten singulären Werte σ_i stimmt mit dem Rang von A überein. Ferner ist die sogenannte *Pseudoinverae* $A^+ \in \mathbb{R}^{n\times m}$ definiert durch

  $$A^+ := V(\ \hat{\Sigma}^+ \quad 0\ )U^T,$$

  wobei die Diagonalmatrix $\hat{\Sigma}^+ \in \mathbb{R}^{n\times n}$ durch

  $$\hat{\Sigma}^+ := \mathrm{diag}\,(1/\sigma_1, \ldots, 1/\sigma_r, 0, \ldots, 0)$$

  definiert ist. Natürlich ist hier r der Rang von A bzw. die Anzahl positiver singulärer Werte. Bei vorgegebenem $b \in \mathbb{R}^m$ ist A^+b unter allen Lösungen des linearen Ausgleichsproblems, $\|Ax - b\|_2$ auf dem $\mathbb{R}^n$ zu minimieren, die eindeutige Lösung mit minimaler euklidischer Norm.

Für uns ist es vorteilhaft, dass man die Singulärwertzerlegung in Matlab sehr einfach erhalten kann. Nach `help svd` erhalten wir die Information:

```
SVD    Singular value decomposition.
   [U,S,V] = SVD(X) produces a diagonal matrix S, of the same
   dimension as X and with nonnegative diagonal elements in
   decreasing order, and unitary matrices U and V so that
```

```
X = U*S*V'.

S = SVD(X) returns a vector containing the singular values.

[U,S,V] = SVD(X,0) produces the "economy size"
decomposition. If X is m-by-n with m > n, then only the
first n columns of U are computed and S is n-by-n.

See also SVDS, GSVD.
```

Für $\lambda > 0$ besitzt das lineare Gleichungssystem

$$[F'(x)^T F'(x) + \lambda I]p = -F'(x)^T F(x)$$

eine eindeutige Lösung $p(\lambda)$. Wir wollen uns überlegen, dass $p(0) := \lim_{\lambda \to 0+} p(\lambda)$ existiert. Dies ist natürlich trivial, wenn Rang $(F'(x)) = n$, der Rang von $F'(x)$ also maximal ist. Mit einer Singulärwertzerlegung

$$F'(x) = U \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T$$

von $F'(x)$ ist

$$\begin{aligned} F'(x)^T F'(x) &= V(\ \hat{\Sigma}\ \ 0\ )U^T U \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T \\ &= V(\ \hat{\Sigma}\ \ 0\ ) \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T \\ &= V\hat{\Sigma}^2 V^T. \end{aligned}$$

Mit $\lambda \to 0+$ ist folglich

$$\begin{aligned} p(\lambda) &= -[F'(x)^T F'(x) + \lambda I]^{-1} F'(x)^T F(x) \\ &= -[V\hat{\Sigma}^2 V^T + \lambda I]^{-1} V(\ \hat{\Sigma}\ \ 0\ )U^T F(x) \\ &= -V(\hat{\Sigma}^2 + \lambda I)^{-1}(\ \hat{\Sigma}\ \ 0\ )U^T F(x) \\ &\to -F'(x)^+ F(x), \end{aligned}$$

wobei $F'(x)^+ \in \mathbb{R}^{n \times m}$ die Pseudoinverse von $F'(x)$ ist. Wir definieren daher die sogenannte *Levenberg-Marquardt-Trajektorie* $p\colon [0, \infty) \longrightarrow \mathbb{R}^n$ durch

$$p(\lambda) := \begin{cases} -F'(x)^+ F(x), & \lambda = 0, \\ -[F'(x)^T F'(x) + \lambda I]^{-1} F'(x)^T F(x), & \lambda > 0. \end{cases}$$

Offenbar gibt es zwei Möglichkeiten:

(a) Es ist $\lambda^* = 0$ und $\|p(0)\|_2 \leq \Delta$. Dann ist $p^* := p(0)$ eine Lösung von $(\mathrm{P}_{x,\Delta})$, und zwar eine mit minimaler euklidischer Norm.

(b) Es ist $\lambda^* > 0$ und $\|p(\lambda^*)\|_2 = \Delta$. Dann ist $p^* := p(\lambda^*)$ die eindeutige Lösung von $(\mathrm{P}_{x,\Delta})$.

Den Vektor $p(0) = -F'(x)^+F(x)$ kann man in Matlab sehr einfach mit Hilfe der Funktion `pinv` erhalten, da `pinv(A)` die Pseudoinverse der Matrix A liefert[9]. Daher werden wir uns jetzt mit der Berechnung von $p(\lambda)$ für $\lambda > 0$ beschäftigen. Wir machen es uns einfach[10] und nehmen an, es sei eine Singulärwertzerlegung von $F'(x)$ berechnet. Es seien also orthogonale Matrizen

$$U = \begin{pmatrix} u_1 & \cdots & u_m \end{pmatrix} \in \mathbb{R}^{m\times m}, \qquad V = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix} \in \mathbb{R}^{n\times n}$$

und eine "Diagonalmatrix"

$$\Sigma = \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} \in \mathbb{R}^{m\times n}, \qquad \hat{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$$

mit

$$F'(x) = U\Sigma V^T$$

mit $\sigma_1 \geq \cdots \geq \sigma_n$ (notwendigerweise sind diese singulären Werte nichtnegativ) bekannt. Ferner sei $r := \operatorname{Rang}(F'(x))$ (der Rang stimmt mit der Anzahl positiver Singulärwerte überein[11]). Dann ist

$$\begin{aligned} p(\lambda) &= -(F'(x)^TF(x) + \lambda I)^{-1}F'(x)^TF(x) \\ &= -(V\Sigma^T \underbrace{U^TU}_{=I}\Sigma V^T + \lambda I)^{-1}V\Sigma^TU^TF(x) \\ &= -V(\Sigma^T\Sigma + \lambda I)^{-1}\Sigma^TU^TF(x) \\ &= -V(\hat{\Sigma}^2 + \lambda I)^{-1}\begin{pmatrix} \hat{\Sigma} & 0 \end{pmatrix}U^TF(x) \\ &= -\sum_{j=1}^{r} \frac{\sigma_j z_j}{\sigma_j^2 + \lambda} v_i, \end{aligned}$$

wobei $z := U^TF(x) = (u_j^TF(x))$. Daher ist

$$\|p(\lambda)\|_2 = \Big(\sum_{j=1}^{r} \frac{\sigma_j^2 z_j^2}{(\sigma_j^2 + \lambda)^2}\Big)^{1/2}.$$

Die Abbildung $\psi\colon [0,\infty) \longrightarrow \mathbb{R}$ sei durch $\psi(\lambda) := \|p(\lambda)\|_2$ definiert. Wir können annehmen, dass $\psi(0) > \Delta$, da andernfalls $\lambda = 0$ schon der gesuchte Parameter ist. Weiter ist

[9] Im Gegensatz zu `svd` ist `pinv` keine built-in function, man kann (und sollte) sie sich also ansehen.

[10] Bei

Moré, J. J. (1978) The Levenberg-Marquardt algorithm: implementation and theory. In: Lecture Notes in Mathematics 630, G. A. Watson, ed., Springer-Verlag, Berlin-Heidelberg-New York, 105–116.

wird statt mit der Singulärwertzerlegung nit der QR-Zerlegung gearbeitet, genauer mit einer QR-Zerlegung mit Spaltenpivotisierung. Hierdurch kann ebenfalls der Rang einer Matrix berechnet werden. Bis zur Version 4 wurde dies in Matlab auch so realisiert.

[11] In Matlab wird der Rang einer Matrix auf genau diese Weise berechnet.

$\lim_{\lambda\to\infty}\psi(\lambda) = 0 < \Delta$. Ferner ist

$$\psi'(\lambda) = -\frac{1}{\psi(\lambda)}\sum_{j=1}^{r}\frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^3} < 0,$$

also $\psi(\cdot)$ monoton fallend, so dass eine eindeutige Lösung $\lambda^* \in (0,\infty)$ von $\psi(\lambda) = \Delta$ existiert. Weiter ist

$$\begin{aligned}
\psi''(\lambda) &= \frac{3}{\psi(\lambda)}\sum_{j=1}^{r}\frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^4} + \frac{\psi'(\lambda)}{\psi(\lambda)^2}\sum_{j=1}^{r}\frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^3} \\
&= \frac{3}{\psi(\lambda)}\sum_{j=1}^{r}\frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^4} - \frac{1}{\psi(\lambda)^3}\Big(\sum_{j=1}^{r}\frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^3}\Big)^2 \\
&= \frac{2}{\psi(\lambda)}\sum_{j=1}^{r}\frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^4} \\
&\qquad + \frac{1}{\psi(\lambda)^3}\underbrace{\Big[\sum_{j=1}^{r}\frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^2}\sum_{j=1}^{r}\frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^4} - \Big(\sum_{j=1}^{r}\frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^3}\Big)^2\Big]}_{\geq 0} \\
&\geq \frac{2}{\psi(\lambda)}\sum_{j=1}^{r}\frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^4} \\
&\qquad \text{(Cauchy-Schwarzsche Ungleichung)},
\end{aligned}$$

also $\psi(\cdot)$ konvex. Ähnlich wie im entsprechenden Fall in Unterabschnitt 4.2.1 ist es nicht ratsam, auf $\psi(\lambda) - \Delta = 0$ das Newton-Verfahren anzuwenden. Besser ist es, das Newton-Verfahren auf

$$\chi(\lambda) := \frac{1}{\psi(\lambda)} - \frac{1}{\Delta} = 0$$

anzuwenden. Entsprechend der Aussage von Lemma 2.2 kann man auch hier zeigen, dass $\chi(\cdot)$ auf $(0,\infty)$ streng monoton wachsend und konkav ist. Ersteres folgt aus

$$\chi'(\lambda) = -\frac{\psi'(\lambda)}{\psi(\lambda)^2} > 0,$$

während die zweite Aussage aus

$$\begin{aligned}
\chi''(\lambda) &= -\frac{\psi''(\lambda)\psi(\lambda)^2 - 2\psi'(\lambda)^2\psi(\lambda)}{\psi(\lambda)^4} \\
&= -\frac{\psi''(\lambda)}{\psi(\lambda)^2} + 2\frac{\psi'(\lambda)^2}{\psi(\lambda)^3} \\
&\leq -\frac{2}{\psi(\lambda)^3}\sum_{j=1}^{r}\frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^4} + 2\frac{\psi'(\lambda)^2}{\psi(\lambda)^3} \\
&= -\frac{2}{\psi(\lambda)^3}\sum_{j=1}^{r}\frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^4} + \frac{2}{\psi(\lambda)^5}\Big(\sum_{j=1}^{r}\frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^3}\Big)^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{\psi(\lambda)^5}\underbrace{\Big[\Big(\sum_{j=1}^r \frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^3}\Big)^2 - \sum_{j=1}^r \frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^2}\sum_{j=1}^r \frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda)^4}\Big]}_{\le 0}\\
&\le 0
\end{aligned}
$$

folgt, wobei am Schluss wieder die Cauchy-Schwarzsche Ungleichung benutzt wurde.

Wir wollen nun obere und untere Schranken für die Lösung $\lambda^* \in (0,\infty)$ von $\psi(\lambda) = \Delta$ bestimmen (wir gehen natürlich weiter davon aus, dass $\psi(0) > \Delta$). Wegen

$$
\Delta = \psi(\lambda^*) = \Big(\sum_{j=1}^r \frac{\sigma_j^2 z_j^2}{(\sigma_j^2+\lambda^*)^2}\Big)^{1/2} \le \frac{1}{\lambda^*}\Big(\sum_{j=1}^r \sigma_j^2 z_j^2\Big)^{1/2}
$$

ist

$$
u_0 := \frac{1}{\Delta}\Big(\sum_{j=1}^r \sigma_j^2 z_j^2\Big)^{1/2}
$$

eine obere Schranke für λ^*. Wegen der Konvexität von $\psi(\cdot)$ auf $[0,\infty)$ ist

$$
\Delta = \psi(\lambda^*) \ge \psi(0) + \psi'(0)\lambda^*,
$$

so dass

$$
l_0 := -\frac{\psi(0)-\Delta}{\psi'(0)}
$$

eine (positive) untere Schranke für λ^* ist.

In dem folgenden Lemma formulieren wir einen Algorithmus zur Lösung der Gleichung $\psi(\lambda) = \Delta$ und beweisen anschließend seine quadratische Konvergenz. Im wesentlichen handelt es sich hierbei um das auf $\chi(\lambda) = 0$ angewandte Newton-Verfahren.

Lemma 3.6 *Gegeben sei die obige Gleichung $\psi(\lambda) = \Delta$, wobei $\psi(0) > \Delta$ vorausgesetzt wird. Man betrachte den folgenden Algorithmus:*

- *Berechne*

  $$
  l_0 := -\frac{\psi(0)-\Delta}{\psi'(0)}, \qquad u_0 := \frac{1}{\Delta}\Big(\sum_{j=1}^r \sigma_j^2 z_j^2\Big)^{1/2}
  $$

  und

  $$
  \lambda_0 := \max(10^{-4}u_0, \sqrt{l_0 u_0}).
  $$

- *Für $k = 0, 1, \ldots$:*
  - *Berechne $\psi(\lambda_k)$.*
  - *Falls $\psi(\lambda_k) = \Delta$, dann STOP.*
  - *Berechne $\psi'(\lambda_k)$ und*

  $$
  l_{k+1} := \max\Big(l_k, \lambda_k - \frac{\psi(\lambda_k)-\Delta}{\psi'(\lambda_k)}\Big), \qquad u_{k+1} := \begin{cases} \lambda_k & \text{für} \quad \psi(\lambda_k) < \Delta, \\ u_k & \text{sonst.} \end{cases}
  $$

– *Berechne*

$$\lambda_{k+1} := \lambda_k + \Big(1 - \frac{\psi(\lambda_k)}{\Delta}\Big)\frac{\psi(\lambda_k)}{\psi'(\lambda_k)}.$$

– *Falls* $\lambda_{k+1} \notin (l_{k+1}, u_{k+1})$, *dann setze* $\lambda_{k+1} := \max(10^{-4}u_{k+1}, \sqrt{l_{k+1}u_{k+1}})$.

Dann konvergiert die Folge $\{\lambda_k\}$ *quadratisch gegen die eindeutige Lösung* λ^* *von* $\psi(\lambda) = \Delta$ *in* $(0, \infty)$.

Beweis: Es ist leicht einzusehen, dass $\{l_k\}$ untere, $\{u_k\}$ obere Schranken für λ^* sind, die offenbar monoton nicht fallen bzw. monoton nicht wachsen. Denn nicht ganz offensichtlich ist nur die Beziehung

$$\lambda_k - \frac{\psi(\lambda_k) - \Delta}{\psi'(\lambda_k)} \le \lambda^*,$$

welche aber wegen der Konvexität von $\psi(\cdot)$ sofort aus

$$\psi'(\lambda_k)(\lambda^* - \lambda_k) \le \psi(\lambda^*) - \psi(\lambda_k) = \Delta - \psi(\lambda_k)$$

folgt.

Ist $\psi(\lambda_k) > \Delta$, also $l_k \le \lambda_k < \lambda^*$, so ist

$$l_{k+1} = \lambda_k \underbrace{- \frac{\psi(\lambda_k) - \Delta}{\psi'(\lambda_k)}}_{>0}$$

die neue (verbesserte) untere Schranke, $u_{k+1} = u_k$ die neue (und alte) obere Schranke und

$$\lambda_k < \tilde{\lambda}_{k+1} := \lambda_k + \Big(1 - \frac{\psi(\lambda_k)}{\Delta}\Big)\frac{\psi(\lambda_k)}{\psi'(\lambda_k)} < \lambda^*,$$

wie man mit

$$\chi(\lambda) := \frac{1}{\psi(\lambda)} - \frac{1}{\Delta}, \qquad \chi'(\lambda) = -\frac{\psi'(\lambda)}{\psi(\lambda)^2}$$

aus

$$\tilde{\lambda}_{k+1} = \lambda_k \underbrace{- \frac{\chi(\lambda_k)}{\chi'(\lambda_k)}}_{>0} < \lambda^*$$

und mit Hilfe der Konkavität von χ aus

$$0 = \chi(\lambda^*) \le \chi(\lambda_k) + \underbrace{\chi'(\lambda_k)}_{>0}(\lambda^* - \lambda_k)$$

erhält. Offenbar ist $l_{k+1} \le \tilde{\lambda}_{k+1} \le u_{k+1}$ und daher $\lambda_{k+1} = \tilde{\lambda}_{k+1}$. Liegt also ein λ_k links von λ^*, so auch alle folgenden, ab diesem Index ist die Folge außerdem monoton wachsend. Hieraus folgt die Konvergenz und dann auch die quadratische Konvergenz gegen λ^*, wenn es nur ein λ_k gibt, welches links von λ^* liegt.

Nun nehmen wir an, es sei $\psi(\lambda_k) < \Delta$ für alle k, d. h. λ_k liege für alle k rechts von λ^*. Dann ist $u_{k+1} = \lambda_k$. Wieder sei

$$\tilde{\lambda}_{k+1} := \lambda_k + \Big(1 - \frac{\psi(\lambda_k)}{\Delta}\Big) \frac{\psi(\lambda_k)}{\psi'(\lambda_k)} = \lambda_k - \frac{\chi(\lambda_k)}{\chi'(\lambda_k)}.$$

Wie oben folgt $\tilde{\lambda}_{k+1} < \lambda^* < u_{k+1}$. Wäre $l_{k+1} \leq \tilde{\lambda}_{k+1}$, so würde $\lambda_{k+1} = \tilde{\lambda}_{k+1}$ links von λ^* liegen, was wir gerade ausgeschlossen haben. Also ist

$$\tilde{\lambda}_{k+1} < l_{k+1} = \max\Big(l_k, \lambda_k - \frac{\psi(\lambda_k) - \Delta}{\psi'(\lambda_k)}\Big).$$

Wegen

$$\tilde{\lambda}_{k+1} > \lambda_k - \frac{\psi(\lambda_k) - \Delta}{\psi'(\lambda_k)}$$

folgt hieraus $l_{k+1} = l_k$. Damit haben wir erhalten: Ist $\psi(\lambda_k) < \Delta$ für alle k, so ist $l_k = l_0$, $u_{k+1} = \lambda_k$ und $\tilde{\lambda}_{k+1} \notin (l_{k+1}, u_{k+1})$ für alle k. Daher ist die Folge $\{\lambda_k\}$ durch die Vorschrift

$$\lambda_{k+1} := \max(10^{-4}\lambda_k, \sqrt{l_0\lambda_k})$$

gegeben. Wegen $l_0 < \lambda^* < \lambda_k$ ist

$$\lambda_{k+1} \leq \max(10^{-4}\lambda_k, \lambda_k) = \lambda_k.$$

Also ist $\{\lambda_k\} \subset \mathbb{R}_+$ eine konvergente Folge. Der Limes sei mit $\hat{\lambda}$ bezeichnet. Es ist $l_0 < \lambda^* \leq \hat{\lambda}$, insbesondere also $\hat{\lambda} > 0$. Durch Grenzübergang $k \to \infty$ erhält man aus $\lambda_{k+1} = \max(10^{-4}\lambda_k, \sqrt{l_0\lambda_k})$, dass

$$\hat{\lambda} = \max(10^{-4}\hat{\lambda}, \sqrt{l_0\hat{\lambda}}).$$

Hieraus folgt

$$10^{-4}\hat{\lambda} < \sqrt{l_0\hat{\lambda}} = \hat{\lambda},$$

aus der letzten Gleichung ergibt sich $\hat{\lambda} = l_0$, womit der gewünschte Widerspruch erreicht wurde. Das Lemma ist damit bewiesen. □
□

Bemerkung: Von Moré wird vorgeschlagen, obiges Iterationsverfahren zur Bestimmung der Lösung von $\psi(\lambda) = \Delta$ abzubrechen, wenn $|\psi(\lambda_k) - \Delta| \leq \sigma\Delta$ mit einem vorgegebenen $\sigma \in (0, 1)$, etwa $\sigma := 0.1$, ist. Dies entspricht der Abbruchbedingung

$$(1 - \sigma)\Delta \leq \|p(\lambda_k)\|_2 \leq (1 + \sigma)\Delta.$$

Es wird von Moré bemerkt, dass in der Praxis weniger als zwei Iterationen genügen, um diese Bedingung zu erfüllen. □

Damit sind wir ziemlich ausführlich auf die Lösung des Hilfsproblems ($P_{x,\Delta}$) eingegangen. Wir wollen hier nicht auf die Moré-Version eines Trust-Region-Verfahrens für nichtlineare Ausgleichsprobleme eingehen. Der Unterschied zum obigen Madsen-Verfahren besteht vor allem darin, dass Moré das zu (P) gleichwertige Problem

$$\text{Minimiere} \quad \tilde{f}(x) := \frac{1}{2}\|F(x)\|_2^2, \quad x \in \mathbb{R}^n$$

betrachtet und die entsprechende Modellfunktion

$$\tilde{f}_x(p) := \frac{1}{2}\|F(x) + F'(x)p\|_2^2$$

benutzt (was Vorteile bei der Berechnung von $\tilde{r}_x := (\tilde{f}(x) - \tilde{f}(x+p^*))/(\tilde{f}(x) - \tilde{f}_x(p^*))$ bringt), eine diagonale Skalierungsmatrix verwendet und eine andere Update-Strategie für die Trust-Region-Radien verfolgt.

Wir geben nun eine Implementation des auf ein nichtlineares Ausgleichsproblem angewandtes Madsen-Verfahren an.

```
function [x,iter]=TrustLeast(Fun,x_0,Delta_0,max_iter,tol);
%************************************************************
%This function solves the nonlinear least square problem
%      Minimize f(x):=||F(x)||,  x in R^n
%with the More Trust-Region method
%************************************************************
%Input parameter:
%      Fun            [F(x),F'(x)]=Fun(x)
%      x_0            initial iterate
%      Delta_0        initial radius
%      max_iter       maximal number of iterations
%      tol            tolerance
%Output parameter:
%      x              approximate solution
%      iter           number of iterations
%************************************************************
rho_1=0.01;rho_2=0.25;rho_3=0.25;sigma_1=0.25;sigma_2=2;
x_c=x_0;Delta_c=Delta_0;[F_c,J_c]=feval(Fun,x_c);iter=0;
[p,lambda]=Hebden(F_c,J_c,Delta_c);
f=norm(F_c);f_c=norm(F_c+J_c*p);
while (f-f_c>tol)&(iter<max_iter)
     iter=iter+1;
     x_plus=x_c+p;
     [F_plus,J_plus]=feval(Fun,x_plus);
     f_plus=norm(F_plus);
     r_c=(f-f_plus)/(f-f_c);
     norm_p=norm(p);
     if (r_c<=rho_2)
         Delta_c=sigma_1*norm_p;
     else
         if (norm(F_plus-F_c-J_c*p)<=rho_3*(f-f_plus))
              Delta_c=sigma_2*norm_p;
         else
              Delta_c=norm_p;
```

```
        end;
    end;
    if (r_c>=rho_1)
        x_c=x_plus;F_c=F_plus;J_c=J_plus;
    end;
    [p,lambda]=Hebden(F_c,J_c,Delta_c);
    f=norm(F_c);f_c=norm(F_c+J_c*p);
end;
x=x_c;%End TrustLeast
%*********************************************************
function [p,lambda]=Hebden(F_c,J_c,Delta_c);
%*********************************************************
%This function solves the trust region subproblem
%   Minimize f_c(p):=||F_c+J_c p||_2, ||p||_2<=Delta_c
%*********************************************************
%Input parameter:
%           F_c,J_c
%           Delta_c    current radius
%Output parameter:
%           p          (approximate) solution
%           lambda     corresponding multiplier
%*********************************************************
sigma=0.1;
%terminate if | ||p(lambda)||-Delta_c|<=sigma*Delta_c
[U,S,V]=svd(J_c,0);s=diag(S);toll=max(size(J_c))*max(s)*eps;
r=sum(s > toll);%r=rank(J_c)
z=U(:,1:r)'*F_c;s=s(1:r);
p_0=-V(:,1:r)*diag(ones(r,1)./s)*z;psi_0=norm(p_0);
if  psi_0<=Delta_c
    p=p_0;lambda=0;
else
    dpsi_0=-sum((z./s.^2).^2)/psi_0;
    l=-(psi_0-Delta_c)/dpsi_0;u=norm(s.*z)/Delta_c;
    lambda=max(0.0001*u,sqrt(l*u));
    psi=norm((s.*z)./(s.^2+lambda));
    while abs(psi-Delta_c)>sigma*Delta_c
        dpsi=-sum((s.*z).^2./((s.^2+lambda).^3))/psi;
        l=max(l,lambda-(psi-Delta_c)/dpsi);
        if psi<Delta_c
            u=lambda;
        end;
        lambda=lambda+(1-psi/Delta_c)*(psi/dpsi);
        if (lambda<l)|(lambda>u)
            lambda=max(0.0001*u,sqrt(l*u));
        end;
        psi=norm((s.*z)./(s.^2+lambda));
    end;
    p=-V(:,1:r)*diag(s./(s.^2+lambda))*z;
end;%End Hebden
%*********************************************************
```

Beispiel: Wir kehren zu einem Beispiel aus dem ersten Kapitel zurück, siehe auch die Anwendung des durch die Armijo-Schrittweite gedämpften Gauß-Newton-Verfahrens am Schluss von Unterabschnitt 3.4.2. Nach `format long` und

```
[x,iter]=TrustLeast('Schwarz',[1.75;1.2;0.8;-0.5;-2],0.5,100,1e-10);
```

erhalten wir

$$x = \begin{pmatrix} 1.75773906245383 \\ 1.42100956539402 \\ 0.67067089707437 \\ -0.55524763139943 \\ -3.38352476697719 \end{pmatrix}, \qquad \text{iter} = 8.$$

Bei der Anwendung des Gauß-Newton-Verfahrens erhalten wir nach dem Aufruf

```
[x,iter]=GauNew('Schwarz',[1.75;1.2;0.8;-0.5;-2],100,1e-10);
```

das Ergebnis

$$x = \begin{pmatrix} 1.75773868939074 \\ 1.42100338889534 \\ 0.67067735263334 \\ -0.55524516124732 \\ -3.38347366913270 \end{pmatrix}, \qquad \text{iter} = 6.$$

Hier darf man natürlich lange nicht allen Dezimalstellen trauen. □

4.3.4 Aufgaben

1. Sei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ konvex und $M \subset \mathbb{R}^n$ konvex. Ist dann x^* eine Lösung der konvexen Optimierungsaufgabe

   (P) $$\text{Minimiere} \quad f(x), \quad x \in M,$$

   welche im *Inneren* von M liegt (insbesondere sei dieses also nichtleer), so ist $f(x^*) \leq f(x)$ für alle $x \in \mathbb{R}^n$, d. h. in x^* liegt ein unrestringiertes Minimum von f.

2. Gegeben sei die nichtlineare Approximationsaufgabe

   (P) $$\text{Minimiere} \quad f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n.$$

   Hierbei sei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ in $x \in \mathbb{R}^n$ stetig partiell differenzierbar. Mit gegebenem $\Delta > 0$, einer symmetrischen und positiv semidefiniten Matrix $B \in \mathbb{R}^{n\times n}$ und einer Norm $\|\cdot\|$ auf dem $\mathbb{R}^n$ sei p^* eine (globale) Lösung von

   ($\mathrm{P}_{x,\Delta}$) $$\text{Minimiere} \quad f_x(p) := \|F(x) + F'(x)p\| + \frac{1}{2}p^T Bp, \quad \|p\| \leq \Delta.$$

   Dann gilt die folgende Verallgemeinerung von Lemma 3.1 bzw. des ersten Teils von Lemma 3.2:

   (a) Es ist x genau dann eine stationäre Lösung von (P), wenn $f(x) = f_x(p^*)$.

   (b) Bezeichnet man den Optimalwert von ($\mathrm{P}_{x,\Delta}$) mit $v(x,\Delta)$ und ist $\Delta^* > 0$, so ist

   $$f(x) - v(x,\Delta) \geq \frac{\Delta}{\Delta^*}[f(x) - v(x,\Delta^*)] \qquad \text{für alle } \Delta \in (0,\Delta^*].$$

3. Man schreibe eine Matlab-Funktion, mit welcher mit Hilfe des Madsen-Verfahrens das nichtlineare Tschebyscheffproblem

   (P) $$\text{Minimiere} \quad \|F(x)\|_\infty, \quad x \in \mathbb{R}^n,$$

   gelöst werden kann, wobei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ stetig differenzierbar ist.

   Hinweis: Das auftretende Hilfsproblem

   $$\text{Minimiere} \quad \|F(x) + F'(x)p\|_\infty, \quad \|p\|_\infty \le \Delta$$

   ist äquivalent zur linearen Optimierungsaufgabe

   $$\begin{gathered}\text{Minimiere} \quad \delta \quad \text{unter den Nebenbedingungen} \\ -\delta e \le F(x) + F'(x)p \le \delta e, \qquad -\Delta e \le p \le \Delta e.\end{gathered}$$

   Hierbei ist e der Vektor des $\mathbb{R}^m$ bzw. $\mathbb{R}^n$, dessen Komponenten alle gleich 1 sind. Steht die Optimization toolbox von Matlab zur Verfügung, so kann man die Funktion `linprog` benutzen, andernfalls ist man leider gezwungen, sich selbst einen Löser für linease Programme zu schreiben. Anschließend teste man die Funktion an den folgenden Beispielen:

   (a) Man setze $t_i := (i-11)/10$, $i = 1, \ldots, 21$. Die Abbildung $F\colon \mathbb{R}^5 \longrightarrow \mathbb{R}^{21}$ sei durch

   $$F_i(x_1, x_2, x_3, x_4, x_5) := \frac{x_1 + x_2 t_i}{1 + x_3 t_i + x_4 t_i^2 + x_5 t_i^3} - \exp(t_i), \qquad i = 1, \ldots, 21,$$

   gegeben.

   (b) Sei $F\colon \mathbb{R}^2 \longrightarrow \mathbb{R}^3$ durch

   $$F(x_1, x_2) := \begin{pmatrix} x_1^2 + x_2^2 + x_1 x_2 \\ \sin x_1 \\ \cos x_2 \end{pmatrix}$$

   gegeben.

   (c) Die Abbildung $F\colon \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ sei durch

   $$F(x_1, x_2) := \begin{pmatrix} x_1 - x_2^3 + 5x_2^2 - 2x_2 - 13 \\ x_1 + x_2^3 + x_2^2 - 14x_2 - 29 \end{pmatrix}$$

   gegeben.

4. Man löse das nichtlineare Ausgleichsproblem

   (P), $$\text{Minimiere} \quad f(x) := \|F(x)\|_2, \quad x \in \mathbb{R}^n,$$

   wobei F wie in den drei Beispielen in Aufgabe 3 gegeben ist.

Kapitel 5

Lösungen zu den Aufgaben

5.1 Aufgaben zu Kapitel 2

5.1.1 Aufgaben zu Abschnitt 2.1

1. Die Funktion $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$ sei definiert durch

$$f(x) := (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2.$$

Für $(x_1, x_2) \in [-5, 5] \times [-5, 5]$ gebe man einen Flächen- und einen Höhenlinienplot an. Anschließend berechne man wenigstens einen stationären Punkt von f.

Lösung: In Abbildung 5.1 geben wir einen Flächen- und einen Höhenlinienplot von f an. Letzteren haben wir z. B. durch

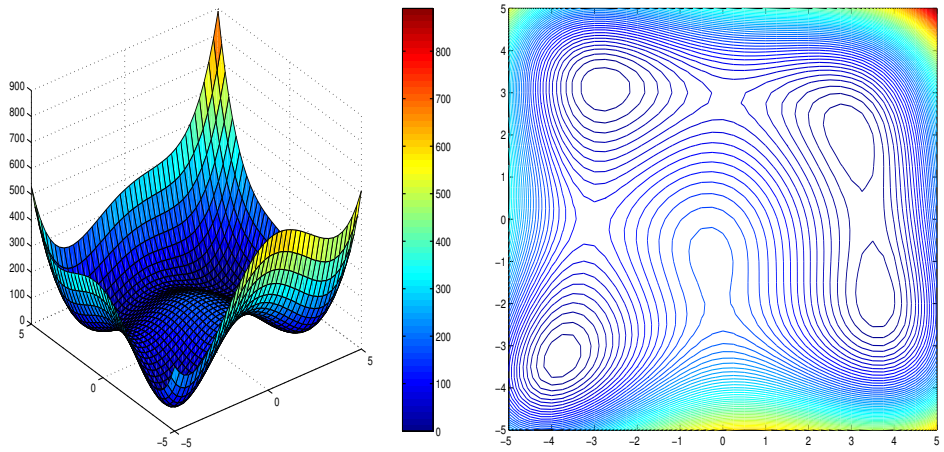

Abbildung 5.1: Flächenplot, Höhenlinienplot

```
x_1=-5:0.25:5;x_2=x_1;
[X_1,X_2]=meshgrid(x_1,x_2);
F=(X_1.^2+X_2-11).^2+(X_1+X_2.^2-7).^2;
contour(X_1,X_2,F,80);
```

erhalten. Als Gradienten von f berechnen wir

$$\nabla f(x) = \begin{pmatrix} 4x_1(x_1^2 + x_2 - 11) + 2(x_1 + x_2^2 - 7) \\ 2(x_1^2 + x_2 - 11) + 4x_2(x_1 + x_2^2 - 7) \end{pmatrix}.$$

Offenbar ist $x^* = (3, 2)$ ein stationärer Punkt von f. Wir haben diesen mit Maple gefunden, zur Berechnung der anderen stationären Punkte ist man auf numerische Verfahren angewiesen. Mit dem Maple-Befehl `fsolve` erhält man z. B. die weitere stationäre Lösung $x^* = (3.584428340, -1.848126527)$.

2. Man berechne die Gateaux-Variation der durch

$$f(x) := \max_{j=1,\ldots,n} x_j, \qquad f(x) := \|x\|_1 = \sum_{j=1}^{n} |x_j|$$

definierten konvexen Funktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$.

Lösung: Für den ersten Teil definieren wir

$$J(x) := \{j \in \{1, \ldots, n\} : x_j = f(x)\},$$

ferner sei $p \in \mathbb{R}^n$ gegeben. Für alle $j \notin J(x)$ ist $x_j + tp_j < f(x + tp)$ und daher

$$\frac{f(x + tp) - f(x)}{t} = \max_{j \in J(x)} \frac{x_j + tp_j - x_j}{t} = \max_{j \in J(x)} p_j.$$

Also ist

$$f'(x; p) = \max_{j \in J(x)} p_j$$

die zugehörige Richtungsableitung.

Im zweiten Fall definieren wir

$$J(x) := \{j \in \{1, \ldots, n\} : x_j = 0\}.$$

Gegeben seien wieder $x, p \in \mathbb{R}^n$. Für alle $j \notin J(x)$ ist dann $x_j + tp_j \neq 0$ für alle hinreichend kleinen $t > 0$, genauer hat $x_j + tp_j$ dasselbe Vorzeichen wie x_j. Für diese t ist dann

$$\begin{aligned} \frac{f(x + tp) - f(x)}{t} &= \sum_{j=1}^{n} \frac{|x_j + tp_j| - |x_j|}{t} \\ &= \sum_{j \in J(x)} |p_j| + \sum_{j \notin J)x)} \frac{|x_j + tp_j| - |x_j|}{t} \\ &= \sum_{j \in J(x)} |p_j| + \sum_{j \notin J)x)} \operatorname{sign}(x_j) p_j. \end{aligned}$$

Also ist

$$f'(x; p) = \sum_{j \in J(x)} |p_j| + \sum_{j \notin J)x)} \operatorname{sign}(x_j) p_j$$

die zugehörige Richtungsableitung.

3. Ist $f(x) := \max_{i=1,\dots,m} F_i(x)$ mit in $x^* \in \mathbb{R}^n$ stetig differenzierbaren $F_i\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \dots, m$, so ist x^* genau dann stationärer Punkt von f bzw.

$$(*) \qquad f'(x^*; p) = \max_{i \in I(x^*)} \nabla F_i(x^*)^T p \geq 0 \qquad \text{für alle } p \in \mathbb{R}^n,$$

wobei

$$I(x^*) := \{i \in \{1, \dots, m\} : F_i(x^*) = f(x^*)\},$$

wenn reelle Zahlen λ_i^*, $i \in I(x^*)$, existieren mit

$$(**) \qquad \lambda_i^* \geq 0 \quad (i \in I(x^*)), \qquad \sum_{i \in I(x^*)} \lambda_i^* = 1, \qquad \sum_{i \in I(x^*)} \lambda_i^* \nabla F_i(x^*) = 0.$$

Lösung: Dass die Gateaux-Variation $f'(x^*; \cdot)$ die angegebene Form hat, kann mit Hilfe von Aufgabe 2 analog zum Beweis von Satz 1.8 (Kettenregel!) bewiesen werden, hierauf wollen wir nicht eingehen. Zunächst nehmen wir an, dass es λ_i^*, $i \in I(x^*)$, mit $(**)$ gibt. Für jedes $p \in \mathbb{R}^n$ ist dann

$$0 = \sum_{i \in I(x^*)} \lambda_i^* \nabla F_i(x^*)^T p \leq \max_{i \in I(x^*)} \nabla F_i(x^*)^T p,$$

es gilt also $(*)$. Umgekehrt nehmen wir an, dass $(*)$ gilt. Durch Widerspruch zeigen wir, dass λ_i^*, $i \in I(x^*)$, mit $(**)$ existieren. Ist dies nicht der Fall, so besitzt

$$\begin{pmatrix} B \\ e^T \end{pmatrix} \lambda = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \lambda \geq 0$$

keine Lösung, wobei die $n \times q$-Matrix B gerade die Spalten $\nabla F_i(x^*)$, $i \in I(x^*)$, besitzt und $q := \#(I(x^*))$ die Anzahl der Elemente von $I(x^*)$ ist. Das Farkas-Lemma liefert genau wie im Beweis von Satz 1.9 einen Widerspruch.

4. Gegeben sei die Min-Max-Aufgabe

$$(\mathrm{P}) \qquad \text{Minimiere} \quad f(x) := \max_{t=1,2,3} F_i(x), \quad x \in \mathbb{R}^2,$$

wobei

$$F_1(x) := x_1^4 + x_2^2, \qquad F_2(x) := (2 - x_1)^2 + (2 - x_2)^2, \qquad F_3(x) := 2\exp(-x_1 + x_2).$$

Man zeige, dass $x^* = (1, 1)$ eine stationäre Lösung von (P) ist. Ferner mache man einen Flächen- und einen Höhenlinienplot auf $[0, 2] \times [0, 2]$.

Lösung: Zunächst geben wir in Abbildung 5.2 den Flächen- bzw. Höhenlinienplot an. Wegen

$$\nabla F_1(x^*) = \begin{pmatrix} 4 \\ 2 \end{pmatrix}, \qquad \nabla F_2(x^*) = \begin{pmatrix} -2 \\ -2 \end{pmatrix}, \qquad \nabla F_3(x^*) = \begin{pmatrix} -2 \\ 2 \end{pmatrix}$$

und $I(x^*) = \{1, 2, 3\}$ hat man wegen Aufgabe 3 die Existenz von $(\lambda_1^*, \lambda_2^*, \lambda_3^*)$ mit

$$\begin{pmatrix} 1 & 1 & 1 \\ 4 & -2 & -2 \\ 2 & -2 & 2 \end{pmatrix} \begin{pmatrix} \lambda_1^* \\ \lambda_2^* \\ \lambda_3^* \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} \lambda_1^* \\ \lambda_2^* \\ \lambda_3^* \end{pmatrix} \geq 0$$

nachzuweisen. Offenbar ist $(\lambda_1^*, \lambda_2^*, \lambda_3^*) = (\frac{1}{3}, \frac{1}{2}, \frac{1}{6})$ die gesuchte Lösung.

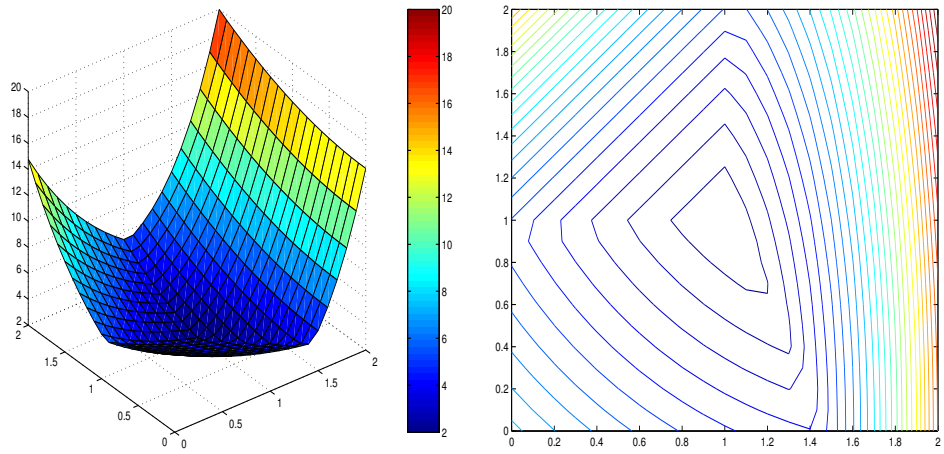

Abbildung 5.2: Flächenplot, Höhenlinienplot zur Min-Max-Aufgabe

5. Ist $f(x) := \sum_{i=1}^{m} |F_i(x)|$ mit in $x^* \in \mathbb{R}^n$ stetig differenzierbaren $F_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$, so ist x^* genau dann stationärer Punkt von f bzw.

$$(*) \qquad \begin{cases} f'(x^*; p) = \displaystyle\sum_{i \in I(x^*)} |\nabla F_i(x^*)^T p| + \sum_{i \notin I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)^T p \geq 0 \\ \qquad\qquad \text{für alle } p \in \mathbb{R}^n, \end{cases}$$

wobei

$$I(x^*) := \{i \in \{1, \ldots, m\} : F_i(x^*) = 0\},$$

wenn reelle Zahlen λ_i^*, $i \in I(x^*)$, existieren mit

$$(**) \quad \lambda_i^* \in [-1, 1] \quad (i \in I(x^*)), \qquad \sum_{i \in I(x^*)} \lambda_i^* \nabla F_i(x^*) + \sum_{i \notin I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*) = 0.$$

Lösung: Dass die Gateaux-Variation $f'(x^*; \cdot)$ die angegebene Form hat, kann mit Hilfe von Aufgabe 2 analog zum Beweis von Satz 1.8 (Kettenregel!) bewiesen werden, hierauf wollen wir nicht eingehen. Dass $(**)$ wieder $(*)$ impliziert, sieht man durch genaueres Hinsehen. Umgekehrt nehmen wir an, dass $(*)$ gilt. Sei $q := \#(I(x^*))$, o. B. d. A. ist $q \geq 1$ (andernfalls ist die Implikation $(*) \Rightarrow (**)$ trivial). Mit $B \in \mathbb{R}^{n \times q}$ bezeichne man die Matrix, deren Spalten durch $\nabla F_i(x^*)$, $i \in I(x^*)$, gegeben sind. Schließlich sei wieder $e := (1, \ldots, 1)^T \in \mathbb{R}^q$ und zur Abkürzung

$$c := \sum_{i \notin I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*).$$

Die Annahme, $(**)$ würde nicht gelten, bedeutet dann, dass das System $B\lambda = -c$, $-e \leq \lambda \leq e$ keine Lösung $\lambda \in \mathbb{R}^q$ besitzt. Um wie beim Beweis von Satz 1.9 das Farkas-Lemma anwenden zu können, muss dieses System sozusagen auf Simplex-Normalform gebracht werden. Tut man dies (Einführung von nichtnegativen Schlupfvariablen y und z sowie Darstellung von als Differenz nichnegativer μ und ν), so erhält man, dass das System

$$\begin{pmatrix} B & -B & 0 & 0 \\ I & -I & I & 0 \\ -I & I & 0 & I \end{pmatrix} \begin{pmatrix} \mu \\ \nu \\ y \\ z \end{pmatrix} = \begin{pmatrix} -c \\ e \\ e \end{pmatrix}, \qquad \begin{pmatrix} \mu \\ \nu \\ y \\ z \end{pmatrix} \geq 0$$

nicht lösbar ist. Nun ist das Farkas-Lemma anwendbar, es zeigt die Lösbarkeit von

$$\begin{pmatrix} B^T & I & -I \\ -B^T & -I & I \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} p \\ u \\ v \end{pmatrix} \leq 0, \qquad \begin{pmatrix} -c \\ e \\ e \end{pmatrix}^T \begin{pmatrix} p \\ u \\ v \end{pmatrix} > 0.$$

Daher existieren $p \in \mathbb{R}^n$ sowie $u_i, v_i \leq 0$, $i \in I(x^*)$, mit

$$\nabla F_i(x^*)^T p + u_i - v_i = 0 \quad (i \in I(x^*))$$

sowie

$$\sum_{i \notin I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)^T p < \sum_{i \in I(x^*)} (u_i + v_i).$$

Dann ist aber

$$\begin{aligned} f'(x^*; p) &= \sum_{i \in I(x^*)} |\nabla F_i(x^*)^T p| + \sum_{i \notin I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)^T p \\ &< \sum_{i \in I(x^*)} \underbrace{[|u_i - v_i| + (u_i + v_i)]}_{\leq 0} \\ &\leq 0, \end{aligned}$$

ein Widerspruch zu $(*)$.

6. Mit Hilfe von Aufgabe 5 zeige man: Ist x_2^* die reelle Lösung von $x_2^3 + x_2 - 9 = 0$ und $x_1^* := \sqrt{10 - x_2^*}$, so ist $x^* = (x_1^*, x_2^*)$ eine stationäre Lösung der Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \sum_{i=1}^{3} |F_i(x)|, \quad x \in \mathbb{R}^2,$$

wobei

$$F_1(x) := x_1^2 + x_2 - 10, \qquad F_2(x) = x_1 + x_2^2 - 7, \qquad F_3(x) := x_1^2 - x_2^3 - 1.$$

Auf $[0, 5] \times [0, 4]$ mache man einen Höhenlinienplot.

Lösung: Es ist

$$x^* \approx (2.84250327681, 1.92017512134),$$

ferner $F_1(x^*) = F_3(x^*) = 0$ und $F_2(x^*) \approx -0.47042422658$, also $F_2(x^*) < 0$ und $I(x^*) = \{1, 3\}$. Wegen Aufgabe 5 ist die Existenz von $\lambda_1^*, \lambda_3^* \in [-1, 1]$ mit

$$\lambda_1^* \nabla F_1(x^*) + \lambda_3^* \nabla F_3(x^*) - \nabla F_2(x^*) = 0$$

nachzuweisen. Letzteres führt auf das lineare Gleichungssystem

$$\begin{pmatrix} 2x_1^* & 2x_1^* \\ 1 & -3(x_2^*)^2 \end{pmatrix} \begin{pmatrix} \lambda_1^* \\ \lambda_3^* \end{pmatrix} = \begin{pmatrix} 1 \\ 2x_2^* \end{pmatrix}$$

bzw.

$$\begin{pmatrix} 5.68500655362 & 5.68500655362 \\ 1 & -11.0612174898 \end{pmatrix} \begin{pmatrix} \lambda_1^* \\ \lambda_2^* \end{pmatrix} = \begin{pmatrix} 1 \\ 3.84035024268 \end{pmatrix}$$

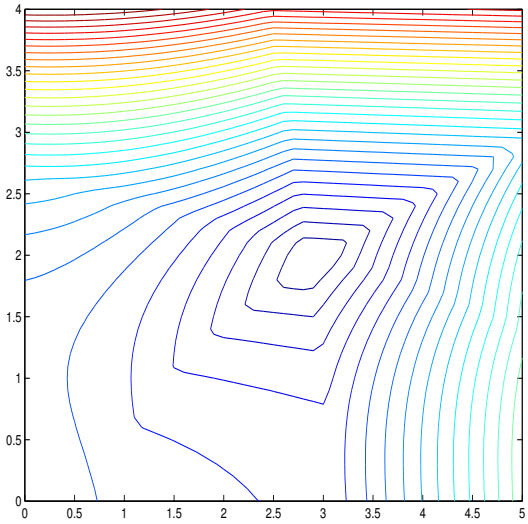

Abbildung 5.3: Höhenlinienplot zum diskreten L_1-Problem

mit der Lösung

$$\begin{pmatrix} \lambda_1^* \\ \lambda_2^* \end{pmatrix} = \begin{pmatrix} 0.47972210925660 \\ -0.30382081687141 \end{pmatrix}.$$

Da die Komponenten in $[-1, 1]$ enthalten sind, ist x^* stationäre Lösung von (P). In Abbildung 5.3 haben wir einen Höhenlinienplot angegeben. Dieses haben wir durch

```
x_1=0:0.1:5;x_2=0:0.1:4;
[X_1,X_2]=meshgrid(x_1,x_2);
F_1=X_1.^2+X_2-10;
F_2=X_1+X_2.^2-7;
F_3=X_1.^2-X_2.^3-1;
F=abs(F_1)+abs(F_2)+abs(F_3);
contour(X_1,X_2,F,30);
```

erhalten.

7. Man[1] bestimme alle stationären Punkte der durch

$$f(x) := x_1^2 x_2^2 + (x_2^2 - 1)^2$$

definierten Funktion $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$. Welche dieser stationären Punkte sind lokale Minima, welche lokale Maxima?

Lösung: Es ist

$$\nabla f(x) = \begin{pmatrix} 2x_1 x_2^2 \\ 2x_1^2 x_2 + 4x_2(x_2^2 - 1) \end{pmatrix}$$

und

$$\nabla^2 f(x) = \begin{pmatrix} 2x_2^2 & 4x_1 x_2 \\ 4x_1 x_2 & -4 + 2x_1^2 + 12x_2^2 \end{pmatrix}.$$

[1] Siehe C. Geiger, C. Kanzow (1999, S. 10).

Stationäre Punkte sind $(0, \pm 1)$ (lokale Minima) sowie $(\alpha, 0)$ mit beliebigem $\alpha \in \mathbb{R}$. Für $|\alpha| \geq 2$ ist die notwendige Bedingung für ein lokales Minimum erfüllt, für $|\alpha| \leq 2$ die notwendige Bedingung für ein lokales Maximum. Dabei ist $f(\alpha, 0) = 1$. In Abbildung

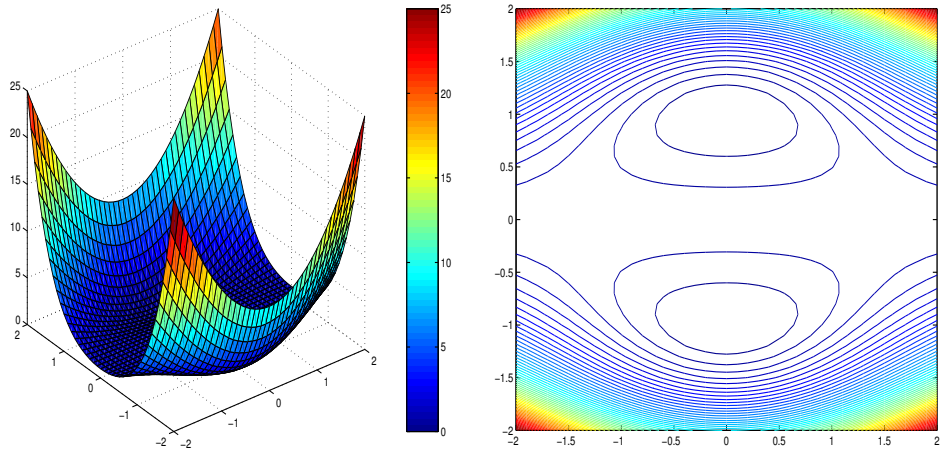

Abbildung 5.4: Flächenplot, Höhenlinienplot

5.4 geben wir einen Flächen- und einen Höhenlinienplot an.

8. Zeige, dass die durch $f(x) := 8x_1 + 12x_2 + x_1^2 - 2x_2^2$ definierte Funktion nur einen stationären Punkt besitzt, der weder ein lokales Minimum noch ein lokales Maximum ist. Man mache einen Höhenlinienplot von f in der Nähe des stationären Punktes.

   Lösung: Es ist

$$\nabla f(x) = \begin{pmatrix} 8 + 2x_1 \\ 12 - 4x_2 \end{pmatrix},$$

   so dass $x^* = (-4, 3)$ der einzige stationäre Punkt von f ist. Da

$$\nabla^2 f(x^*) = \begin{pmatrix} 2 & 0 \\ 0 & -4 \end{pmatrix}$$

   indefinit ist, ist bei x^* weder ein Minimum noch ein Maximum. In Abbildung 5.5 geben wir einen Höhenlinienplot an. In diesem sieht man deutlich, dass x^* ein *Sattelpunkt* ist.

9. Gegeben sei die unrestringierte Min-Max-Optimierungsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \max_{i=1,\ldots,m} F_i(x), \quad x \in \mathbb{R}^n.$$

   Die Funktionen $F_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$, seien auf einer offenen Umgebung von $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbar. Sei

$$I(x^*) := \{i \in \{1, \ldots, m\} : F_i(x^*) = f(x^*)\}.$$

   Es wird vorausgesetzt, dass reelle Zahlen λ_i^*, $i \in I(x^*)$, existieren mit:

   (a) Es ist

$$\lambda_i^* \geq 0 \quad (i \in I(x^*)), \qquad \sum_{i \in I(x^*)} \lambda_i^* = 1, \qquad \sum_{i \in I(x^*)} \lambda_i^* \nabla F_i(x^*) = 0.$$

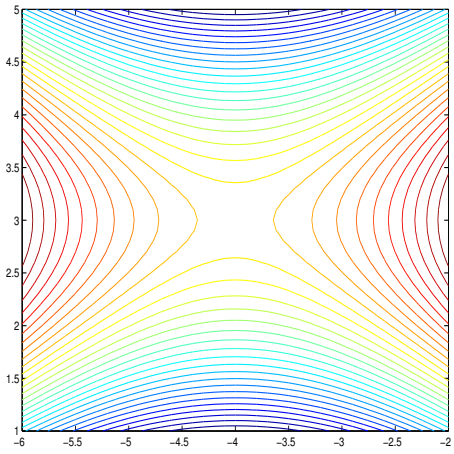

Abbildung 5.5: Ein Sattelpunkt

(b) Mit

$$T^* := \{p \in \mathbb{R}^n : \nabla F_i(x^*)^T p = 0 \text{ für alle } i \in I(x^*) \text{ mit } \lambda_i^* > 0\}$$

ist

$$p^T \left\{ \sum_{i \in I(x^*)} \lambda_i^* \nabla^2 F_i(x^*) \right\} p > 0 \qquad \text{für alle } p \in T^* \setminus \{0\}.$$

Man zeige, dass x^* eine isolierte lokale Lösung von (P) ist.

Lösung: Wir folgen sehr genau dem Beweis von Satz 1.12. Wieder nehmen wir an, die Behauptung sei nicht richtig, so dass eine Folge gegen x^* konvergente Folge $\{x_k\}$ mit $x_k \neq x^*$ und $f(x_k) \leq f(x^*)$ für alle k existiert. Es ist $x_k = x^* + t_k p_k$ mit

$$t_k := \|x_k - x^*\|, \qquad p_k := \frac{x_k - x^*}{\|x_k - x^*\|}.$$

O. B. d. A. existiert $p = \lim_{k\to\infty}(x_k - x^*)/\|x_k - x^*\|$, ferner ist

$$f'(x^*; p) = \max_{i \in I(x^*)} \nabla F_i(x^*)^T p \leq 0.$$

Aus der ersten Voraussetzung erhalten wir

$$\sum_{i \in I(x^*)} \lambda_i^* \underbrace{\nabla F_i(x^*)^T p}_{\leq 0} = 0$$

und hieraus $p \in T^*$. Für $i \in I(x^*)$ ist

$$F_i(x_k) \leq f(x_k) \leq f(x^*) = F_i(x^*)$$

und daher

$$\begin{aligned} 0 &\geq F_i(x_k) - F_i(x^*) \\ &= F_i(x^* + t_k p_k) - F_i(x^*) \\ &= t_k \nabla F_i(x^*)^T p_k + \frac{1}{2} t_k^2 p_k^T \nabla^2 F_i(z_{ik}) p_k \end{aligned}$$

mit $z_{ik} = x^* + \theta_{ik} t_k p_k$ und $\theta_{ik} \in (0,1)$. Eine Multiplikation dieser Ungleichung mit λ_i^*, $i \in I(x^*)$, und anschließendes Aufsummieren liefert unter erneuter Benutzung der ersten Voraussetzung, dass

$$0 \geq t_k \underbrace{\Big\{ \sum_{i \in I(x^*)} \lambda_i^* \nabla F_i(x^*) \Big\}^T}_{=0} p_k + \frac{1}{2} t_k^2 p_k^T \Big\{ \sum_{i \in I(x^*)} \lambda_i^* \nabla^2 F_i(z_{ik}) \Big\} p_k$$

bzw.

$$p_k^T \Big\{ \sum_{i \in I(x^*)} \lambda_i^* \nabla^2 F_i(z_{ik}) \Big\} p_k \leq 0$$

für alle k. Mit $k \to \infty$ folgt wegen $p_k \to p$ und $z_{ik} \to x^*$, dass

$$p^T \Big\{ \sum_{i \in I(x^*)} \lambda_i^* \nabla^2 F_i(x^*) \Big\} p \leq 0,$$

was wegen $p^* \in T^* \setminus \{0\}$ ein Widerspruch zur zweiten Voraussetzung ist.

10. Gegeben sei die Min-Max-Optimierungsaufgabe aus Aufgabe 4, also

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \max_{t=1,2,3} F_i(x), \quad x \in \mathbb{R}^2,$$

wobei

$$F_1(x) := x_1^4 + x_2^2, \qquad F_2(x) := (2 - x_1)^2 + (2 - x_2)^2, \qquad F_3(x) := 2 \exp(-x_1 + x_2).$$

Man zeige, dass $x^* = (1,1)$ eine isolierte lokale Lösung von (P) ist.

Lösung: Es ist $I(x^*) = \{1,2,3\}$ und Teil (a) der hinreichenden Optimalitätsbedingung zweiter Ordnung in Aufgabe 9 ist mit $\lambda^* = (\frac{1}{3}, \frac{1}{2}, \frac{1}{6})$ erfüllt. Man rechnet leicht nach, dass $T^* = \{0\}$ (mit den Bezeichnungen von Aufgabe 9), so dass Teil (b) der hinreichenden Optimalitätsbedingung zweiter Ordnung automatisch erfüllt ist.

11. Gegeben sei die diskrete L_1-Approximationsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \sum_{i=1}^{m} |F_i(x)|, \quad x \in \mathbb{R}^n.$$

Die Funktionen $F_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$, seien auf einer offenen Umgebung von $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbar. Sei

$$I(x^*) := \{i \in \{1, \ldots, m\} : F_i(x^*) = 0\}.$$

Es wird vorausgesetzt, dass reelle Zahlen λ_i^*, $i \in I(x^*)$, existieren mit:

(a) Es ist $\lambda_i^* \in [-1,1]$, $i \in I(x^*)$, und

$$\sum_{i \in I(x^*)} \lambda_i^* \nabla F_i(x^*) + \sum_{i \notin I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*) = 0.$$

(b) Mit

$$T^* := \left\{ p \in \mathbb{R}^n : \nabla F_i(x^*)^T p \left\{ \begin{array}{ll} = 0, & i \in I(x^*) \text{ mit } |\lambda_i^*| < 1, \\ \geq 0, & i \in I(x^*) \text{ mit } \lambda_i^* = 1, \\ \leq 0, & i \in I(x^*) \text{ mit } \lambda_i^* = -1 \end{array} \right\} \right\}$$

ist

$$p^T \Big\{ \sum_{i \in I(x^*)} \lambda_i^* \nabla^2 F_i(x^*) + \sum_{i \notin I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla^2 F_i(x^*) \Big\} p > 0 \text{ für alle } p \in T^* \setminus \{0\}.$$

Man zeige, dass x^* eine isolierte lokale Lösung von (P) ist.

Lösung: Wie im Beweis von Satz 1.12 (und der Lösung von Aufgabe 9) erhält man unter der Annahme, die Aussage sei falsch, die Existenz einer Richtung $p \neq 0$ mit $f'(x^*; p) \leq 0$. Also ist

$$\begin{aligned} 0 &\geq f'(x^*; p) \\ &= \sum_{i \in I(x^*)} |\nabla F_i(x^*)^T p| + \sum_{i \in I(x^*)} \operatorname{sign}(F_i(x^*)) \nabla F_i(x^*)^T p \\ &= \sum_{i \in I(x^*)} \underbrace{[|nablaF_i(x^*)^T p| - \lambda_i^* \nabla F_i(x^*)^T p]}_{\geq 0}, \end{aligned}$$

woraus wir $\lambda_i^* \nabla F_i(x^*)^T p = |\nabla F_i(x^*)^T p|$ für alle $i \in I(x^*)$ schließen. Hieraus folgt ganz offensichtlich $p \in T^*$. Der Rest folgt analog zum Beweis von Satz 1.12 bzw. Aufgabe 9, wir wollen diese Schlüsse diesmal nicht wiederholen.

12. Wir betrachten noch einmal die dikrete L_1-Approximationsaufgabe aus Aufgabe 6, also

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \sum_{i=1}^{3} |F_i(x)|, \quad x \in \mathbb{R}^2,$$

wobei

$$F_1(x) := x_1^2 + x_2 - 10, \qquad F_2(x) = x_1 + x_2^2 - 7, \qquad F_3(x) := x_1^2 - x_2^3 - 1.$$

Wir wissen: Ist x_2^* die reelle Lösung von $x_2^3 + x_2 - 9 = 0$ und $x_1^* := \sqrt{10 - x_2^*}$, so ist $x^* = (x_1^*, x_2^*)$ eine stationäre Lösung. Man zeige, dass x^* eine isolierte lokale Lösung von (P) ist.

Lösung: Es ist $I(x^*) = \{1, 3\}$. Teil (a) von Aufgabe 11 ist mit gewissen λ_1^*, λ_3^*, die betragsmäßig kleiner als 1 sind, erfüllt (siehe Lösung von Aufgabe 6). Daher ist

$$T^* = \{p \in \mathbb{R}^2 : \nabla F_1(x^*)^T p = \nabla F_3(x^*)^T p = 0\} = \{0\},$$

so dass Teil (b) trivialerweise erfüllt ist.

5.1.2 Aufgaben zu Abschnitt 2.2

1. Sei $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ affin linear und $f\colon \mathbb{R}^m \longrightarrow \mathbb{R}$ konvex. Dann ist auch $h := f \circ g$ konvex.

   Lösung: Seien $x, y \in \mathbb{R}^n$ und $t \in [0,1]$. Dann ist

   $$\begin{aligned} h((1-t)x+ty) &= f(g((1-t)x+ty)) \\ &= f((1-t)g(x)+tg(y)) \\ &\qquad (g \text{ affin linear}) \\ &\leq (1-t)f(g(x))+tf(g(y)) \\ &\qquad (f \text{ konvex}) \\ &= (1-t)h(x)+th(y). \end{aligned}$$

   Damit ist die Konvexität von h bewiese

2. Sei $D \subset \mathbb{R}^n$ konvex, $g\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ auf D konvex und $f\colon \mathbb{R} \longrightarrow \mathbb{R}$ konvex und monoton nicht fallend. Dann ist $h := f \circ g$ auf D konvex.

   Lösung: Auch dieser Beweis ist völlig elementar und benutzt nur die Definition der Konvexität von Funktionen. Seien $x, y \in D$ und $t \in [0,1]$. Dann ist

   $$\begin{aligned} h((1-t)x+ty) &= f(g((1-t)x+ty)) \\ &\leq f((1-t)g(x)+tg(y)) \\ &\qquad (g \text{ konvex und } f \text{ monoton nicht fallend}) \\ &\leq (1-t)f(g(x))+tf(g(y)) \\ &\qquad (f \text{ konvex}) \\ &= (1-t)h(x)+th(y). \end{aligned}$$

   Damit ist die Konvexität von h bewiesen.

3. Seien $F_i\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$, konvex auf dem $\mathbb{R}^n$. Dann ist auch die durch

   $$f(x) := \max_{i=1,\ldots,m} F_i(x)$$

   definierte Funktion $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ konvex (auf dem $\mathbb{R}^n$).

   Lösung: Seien $x, y \in \mathbb{R}^n$ und $t \in [0,1]$. Dann ist

   $$\begin{aligned} f((1-t)x+ty) &= \max_{i=1,\ldots,m} F_i((1-t)x+ty) \\ &= F_k((1-t)x+ty) \qquad \text{mit einem gewissen } k \in \{1,\ldots,m\}. \\ &\leq (1-t)F_k(x)+tF_k(y) \\ &\qquad (F_k \text{ konvex}) \\ &\leq (1-t)\max_{i=1,\ldots,m} F_i(x) + t\max_{i=1,\ldots,m} F_i(y) \\ &= (1-t)f(x)+tf(y), \end{aligned}$$

   also f konvex.

4. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit. Die Abbildung $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ sei definiert durch

$$f(x) := \frac{1}{2}(x^T Ax)^2.$$

Man berechne den Gradienten und die Hessesche von f und weise nach, dass f auf dem $\mathbb{R}^n$ konvex ist.

Lösung: Es ist

$$\begin{aligned} f(x+h) &= \frac{1}{2}[(x+h)^T A(x+h)]^2 \\ &= \frac{1}{2}[x^T Ax + 2(Ax)^T h + h^T Ah]^2 \\ &= \frac{1}{2}[(x^T Ax + 2(Ax)^T h)^2 + 2(x^T Ax)(h^T Ah)] + O(\|h\|_2^3) \\ &= \frac{1}{2}(x^T Ax)^2 + 2x^T Ax\,(Ax)^T h + 2((Ax)^T h)^2 + (x^T Ax)(h^T Ah) + O(\|h\|_2^3) \\ &= \frac{1}{2}(x^T Ax)^2 + [2x^T Ax\,Ax]^T h + \frac{1}{2}h^T[4Ax(Ax)^T + 2(x^T Ax)A]h + O(\|h\|_2^3) \\ &= f(x) + \nabla f(x)^T h + \frac{1}{2}h^T \nabla^2 f(x) h + O(\|h\|_2^3). \end{aligned}$$

Hieraus erhält man

$$\nabla f(x) = 2x^T Ax\,Ax, \qquad \nabla^2 f(x) = 4Ax(Ax)^T + 2(x^T Ax)A.$$

Als Summe von zwei positiv semidefiniten Matrizen ist die Hessesche $\nabla^2 f(x)$ für alle $x \in \mathbb{R}^n$ positiv semidefinit und daher f auf dem $\mathbb{R}^n$ konvex. Natürlich hätte man die Konvexität von f aber aber auch direkt beweisen können.

5. Man zeige: Ist $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ gleichmäßig konvex mit der Konstanten $c > 0$, so ist

$$\frac{c}{2}\|p\|_2^2 + f'(x;p) \le f(x+p) - f(x) \qquad \text{für alle } x, p \in \mathbb{R}^n.$$

Lösung: Bei vorgegebenen $x, p \in \mathbb{R}^n$ und alle $t \in [0,1]$ ist (Definition der gleichmäßigen Konvexität mit $y := x + p$)

$$\frac{c}{2}t(1-t)\|p\|_2^2 + f(x+tp) - f(x) \le t[f(x+p) - f(x)].$$

Nach Division dieser Ungleichung durch $t > 0$ und anschließendem Grenzübergang $t \to 0+$ folgt die Behauptung.

6. Sei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ gleichmäßig konstant mit der Konstanten $c > 0$. Man zeige:

   (a) Es ist

$$\frac{c}{2}\|p\|_2^2 + v^T p \le f(x+p) - f(x)$$

   für beliebige $x \in \mathbb{R}^n$, $v \in \partial f(x)$ und $p \in \mathbb{R}^n$.

   (b) Für beliebiges $x_0 \in \mathbb{R}^n$ ist die Niveaumenge

$$L_0 := \{x \in \mathbb{R}^n : f(x) \le f(x_0)\}$$

   kompakt.

Lösung: Seien $x \in \mathbb{R}^n$, $v \in \partial f(x)$ und $p \in \mathbb{R}^n$ beliebig gegeben. Wegen Aufgabe 5 ist

$$\frac{c}{2}\|p\|_2^2 + f'(x;p) \le f(x+p) - f(x).$$

Wegen $v \in \partial f(x)$ ist für alle $t > 0$ ferner

$$v^T p = \frac{v^T(tp)}{t} = \frac{v^T(x+tp-x)}{t} \le \frac{f(x+tp)-f(x)}{t},$$

woraus mit $t \to 0+$ gerade $v^T p \le f'(x;p)$ folgt (das ist die einfache Richtung in Teil 5b von Satz 2.4).

Für beliebiges $x_0 \in \mathbb{R}^n$ ist die Niveaumenge

$$L_0 := \{x \in \mathbb{R}^n : f(x) \le f(x_0)\}$$

wegen der Stetigkeit von f (siehe Teil 1 von Satz 2.4) abgeschlossen. Zu zeigen bleibt daher die Beschränktheit von L_0. Sei hierzu $x \in L_0$ beliebig. Mit $v \in \partial f(x_0)$ erhält man aus dem ersten Teil dieser Aufgabe, dass

$$\frac{c}{2}\|x - x_0\|_2^2 + v^T(x - x_0) \le f(x) - f(x_0) \le 0.$$

Daher ist

$$\frac{c}{2}\|x - x_0\|_2^2 \le -v^T(x - x_0) \le \|v\|_2 \, \|x - x_0\|_2,$$

so dass L_0 in der euklidischen Kugel um x_0 mit dem Radius $2\|v\|_2/c$ enthalten ist. Also ist L_0 auch beschränkt und die Aussage bewiesen.

7. Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n \times n$-Matrizen und $\mathcal{S}_+^{n\times n}$ die konvexe, offene Teilmenge aller positiv definiten $n \times n$-Matrizen. Man definiere $f\colon \mathcal{S}_+^{n\times n} \subset \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ durch

$$f(A) := \frac{\operatorname{tr}(A)/n}{\det(A)^{1/n}}.$$

Man zeige:

(a) Es ist $f(A) \ge 1$ für alle $A \in \mathcal{S}_+^{n\times n}$, ferner ist $f(A) = 1$ genau dann, wenn A ein positives Vielfaches der Identität ist.

(b) Für jedes $A \in \mathcal{S}_+^{n\times n}$ und jedes $P \in \mathcal{S}^{n\times n}$ existiert

$$f'(A;P) := \lim_{t\to 0+} \frac{f(A+tp)-f(A)}{t}.$$

Man berechne $f'(A;P)$ und zeige, dass die Abbildung $f'(A;\cdot)\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ linear ist.

(c) Sind $A, B \in \mathcal{S}_+^{n\times n}$ und $0 \le f'(A; B-A)$, so ist $f(A) \le f(B)$.

Lösung: Für $A \in \mathcal{S}_+^{n\times n}$ mögen $\lambda_i(A)$, $i = 1, \ldots, n$, die (positiven) Eigenwerte von A bezeichnen. Dann ist

$$f(A) = \frac{\operatorname{tr}(A)/n}{\det(A)^{1/n}} = \frac{(1/n)\sum_{i=1}^n \lambda_i(A)}{[\prod_{i=1}^n \lambda_i(A)]^{1/n}}.$$

Wegen der Ungleichung vom geometrisch-arithmetischen Mittel ist $f(A) \geq 1$. Gleichheit gilt hier genau dann, wenn alle Eigenwerte von A gleich sind, was für eine symmetrische, positiv definite Matrix genau dann der Fall ist, wenn sie ein positives Vielfaches der Identität ist.

Nun seien $A \in \mathcal{S}_+^{n\times n}$ und $P \in \mathcal{S}^{n\times n}$ gegeben. Dann ist $A + tP \in \mathcal{S}_+^{n\times n}$ für alle hinreichend kleinen $t > 0$ und

$$\begin{aligned}
\frac{f(A+tP)-f(A)}{t} &= \frac{1}{t}\left(\frac{\operatorname{tr}(A+tP)/n}{\det(A+tP)^{1/n}} - \frac{\operatorname{tr}(A)/n}{\det(A)^{1/n}}\right)\\
&= \frac{1}{t}\left(\frac{\operatorname{tr}(A)/n + t\operatorname{tr}(P)/n}{\det(A)^{1/n}\det(I+tA^{-1/2}PA^{-1/2})^{1/n}} - \frac{\operatorname{tr}(A)/n}{\det(A)^{1/n}}\right)\\
&= \frac{1}{t\det(A)^{1/n}}\left(\frac{\operatorname{tr}(A)/n + t\operatorname{tr}(P)/n}{[\prod_{i=1}^n(1+t\lambda_i(A^{-1/2}PA^{-1/2}))]^{1/n}} - \operatorname{tr}(A)/n\right)\\
&= \frac{1}{t\det(A)^{1/n}}\left(\frac{\operatorname{tr}(A)/n + t\operatorname{tr}(P)/n}{[1+t\operatorname{tr}(A^{-1/2}PA^{-1/2})+O(t^2)]^{1/n}} - \operatorname{tr}(A)/n\right)\\
&= \frac{1}{t\det(A)^{1/n}}\left(\frac{\operatorname{tr}(A)/n + t\operatorname{tr}(P)/n}{1+t\operatorname{tr}(A^{-1/2}PA^{-1/2})/n+O(t^2)} - \operatorname{tr}(A)/n\right)\\
&= \frac{1}{n\det(A)^{1/n}}\left(\operatorname{tr}(P) - \frac{\operatorname{tr}(A)}{n}\operatorname{tr}(A^{-1/2}PA^{-1/2})\right) + O(t).
\end{aligned}$$

Daher ist

$$f'(A;P) = \frac{1}{n\det(A)^{1/n}}\left(\operatorname{tr}(P) - \frac{\operatorname{tr}(A)}{n}\operatorname{tr}(A^{-1/2}PA^{-1/2})\right).$$

Offensichtlich ist $f'(A;\cdot)\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ linear. Damit ist auch der zweite Teil der Aufgabe gelöst.

Für $A, B \in \mathcal{S}_+^{n\times n}$ ist

$$\begin{aligned}
f'(A;B-A) &= \frac{1}{n\det(A)^{1/n}}\left(\operatorname{tr}(B) - \operatorname{tr}(A) - \frac{\operatorname{tr}(A)}{n}\left[\operatorname{tr}(A^{-1/2}BA^{-1/2}) - n\right]\right)\\
&= \frac{1}{n\det(A)^{1/n}}\left(\operatorname{tr}(B) - \operatorname{tr}(A)\,\frac{\operatorname{tr}(A^{-1/2}BA^{-1/2})}{n}\right)\\
&\leq \frac{1}{n\det(A)^{1/n}}[\operatorname{tr}(B) - \operatorname{tr}(A)\det(A^{-1/2}BA^{-1/2})^{1/n}]\\
&\qquad \text{(Ungleichung vom geometrisch-arithmetischen Mittel)}\\
&= \frac{\det(B)^{1/n}}{\det(A)^{1/n}}\left(\frac{\operatorname{tr}(B)/n}{\det(B)^{1/n}} - \frac{\operatorname{tr}(A)}{\det(A)^{1/n}}\right).
\end{aligned}$$

Hieraus liest man ab, dass $f(A) \leq f(B)$ aus $A, B \in \mathcal{S}_+^{n\times n}$ und $0 \leq f'(A;B-A)$ folgt. Damit ist die Aufgabe gelöst.

8. Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n\times n$-Matrizen. Wir definieren die Abbildung $\lambda_{\max}\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ dadurch, dass $\lambda_{\max}(A)$ den maximalen Eigenwert von $A \in \mathcal{S}^{n\times n}$ bedeutet. Man zeige:

   (a) Die Abbildung $\lambda_{\max}\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ ist konvex.

(b) Die Abbildung $\lambda_{\max}\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ besitzt auf $\mathcal{S}^{n\times n}$ eine Gateaux-Variation, d. h. für alle $A, P \in \mathcal{S}^{n\times n}$ existiert

$$\lambda'_{\max}(A;P) := \lim_{t\to 0+} \frac{\lambda_{\max}(A+tP) - \lambda_{\max}(A)}{t}.$$

(c) Die Gateaux-Variation von $\lambda_{\max}$ ist für $A, P \in \mathcal{S}^{n\times n}$ gegeben durch

$$f'(A;P) = \max_{q\in Q_{\max}(A)\setminus\{0\}} \frac{q^T P q}{q^T q}.$$

Hierbei bezeichne $Q_{\max}(A)$ den Eigenraum zu $\lambda_{\max}(A)$, also die lineare Hülle aller Eigenvektoren zu $\lambda_{\max}(A)$.

Lösung: Seien $A, B \in \mathcal{S}^{n\times n}$ und $t \in [0,1]$. Wir benutzen im folgenden, dass der größte Eigenwert einer symmetrischen Matrix das Maximum des Rayleigh-Quotien über alle von Null verschiedenen Vektoren ist. Daher ist

$$\begin{aligned}
\lambda_{\max}((1-t)A + tB) &= \max_{x\neq 0} \frac{x^T((1-t)A+tB)x}{x^T x} \\
&\leq (1-t)\max_{x\neq 0}\frac{x^T A x}{x^T x} + t\max_{x\neq 0}\frac{x^T B x}{x^T x} \\
&= (1-t)\lambda_{\max}(A) + t\lambda_{\max}(B),
\end{aligned}$$

das ist die Konvexität von $\lambda_{\max}\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$.

Die Existenz der Gateaux-Variation folgt aus Lemma 1.6, denn natürlich gilt die Aussage dort nicht nur für konvexe Funktionen auf dem $\mathbb{R}^n$ sondern auch für solche auf einem beliebigen linearen Raum. Damit ist auch der zweite Teil der Aufgabe bewiesen.

Gegeben seien $A, P \in \mathcal{S}^{n\times n}$. Sei $\{t_k\} \subset \mathbb{R}_+$ eine Nullfolge. Hierzu existiert eine Folge $\{q_k\} \subset \mathbb{R}^n \setminus \{0\}$ mit

$$\lambda_{\max}(A+tP) = \frac{q_k^T(A+t_kP)q_k}{q_k^T q_k} = \max_{q\in Q_{\max}(A)\setminus\{0\}} \frac{q^T(A+t_kP)q}{q^T q}.$$

O. B. d. A. können wir $\|q_k\|_2 = 1$ annehmen. Dann kann aus $\{q_k\}$ eine konvergente Teilfolge ausgewählt werden, o. B. d. A. ist schon $\{q_k\}$ konvergent gegen ein $q \in \mathbb{R}^n$ mit $\|q\|_2 = 1$. Wir überlegen uns zunächst, dass $q \in Q_{\max}(A)$. Hierzu beachten wir, dass $\{\lambda_{\max}(A+t_kP)\}$ einerseits gegen $\lambda_{\max}(A)$ und anderseits gegen q^TAq konvergiert. Da $\lambda_{\max}(A)I - A$ symmetrisch und positiv semidefinit ist, folgt aus $q^T(\lambda_{\max}(A)I - A)q = 0$, dass $Aq = \lambda_{\max}(A)q$ bzw. $q \in Q_{\max}(A)$. Wegen

$$\frac{\lambda_{\max}(A+t_kP) - \lambda_{\max}(A)}{t_k} \leq \frac{q_k^T(A+t_kP)q_k - q_k^T A q_k}{t_k} = q_k^T P q_k \to q^T P q$$

folgt

$$\lambda'_{\max}(A;P) \leq \max_{q\in Q_{\max}(A)\setminus\{0\}} \frac{q^T P q}{q^T q},$$

womit eine Richtung bewiesen ist. Zum Nachweis der anderen gebe man sich ein beliebiges $q \in Q_{\max}(A)$ mit $q \neq 0$ vor. O. B. d. A. ist $\|q\|_2 = 1$. Wegen $\lambda_{\max}(A) = q^TAq$ ist

$$\frac{\lambda_{\max}(A+t_kP) - \lambda_{\max}(A)}{t_k} \geq \frac{q^T(A+t_kP)q - q^TAq}{t_k} = q^TPq$$

und folglich $\lambda'_{\max}(A;P) \geq q^T P q$ für jedes $q \in Q_{\max}(A)$ mit $\|q\|_2 = 1$. Insgesamt ist damit

$$\lambda'_{\max}(A;P) = \max_{q \in Q_{\max}(A)\setminus\{0\}} \frac{q^T P q}{q^T q}$$

bewiesen, also auch der dritte Teil der Aufgabe gelöst.

5.2 Aufgaben zu Kapitel 3

5.2.1 Aufgaben zu Abschnitt 3.1

1. Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Bedingungen (V) (a)–(c) in Unterabschnitt 3.1.1. Sei $x_c \in L_0$ keine stationäre Lösung von (P) und $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x_c, d. h. $\nabla f(x_c)^T p < 0$. Seien $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ gegeben. Hiermit definiere man

$$T_{SW}(x_c, p) := \left\{ t > 0 : \begin{array}{rcl} f(x_c + tp) & \leq & f(x_c) + \alpha t \nabla f(x_c)^T p, \\ |\nabla f(x_c + tp)^T p| & \leq & -\beta \nabla f(x_c)^T p \end{array} \right\},$$

die Menge der *strengen Wolfe-Schrittweiten*. Man zeige:

   (a) Es ist $T_{SW}(x_c, p) \neq \emptyset$.

   (b) Es existiert eine von x_c und p unabhängige Konstante $\theta > 0$ mit

$$f(x_c) - f(x_c + tp) \geq \theta \left(\frac{\nabla f(x_c)^T p}{\|p\|_2} \right)^2 \qquad \text{für alle } t \in T_{SW}(x_c, p).$$

Lösung: Jede strenge Wolfe-Schrittweite ist auch eine Wolde-Schrittweite, so dass es wegen Satz 1.3 genügt, die Existenz strenger Wolfe-Schrittweiten zu zeigen. Die Argumentation kann aber nun genau wie im Beweis zu Satz 1.3 verlaufen. Wir wiederholen sie hier. Zur Abkürzung setze man

$$\Phi(t) := f(x_c) + \alpha t \nabla f(x_c)^T p - f(x_c + tp).$$

Ist t^* die erste positive Nullstelle $\nabla f(x_c + tp)^T p$, so ist

$$\Phi'(0) = -(1 - \alpha) \nabla f(x_c)^T p > 0, \qquad \Phi'(t^*) = \alpha \nabla f(x_c)^T p < 0.$$

Wegen $\Phi(0) = 0$ existiert daher ein $t \in (0, t^*)$ mit $\Phi(t) > 0$ und $\Phi'(t) = 0$. Wegen $\Phi(t) > 0$ ist $f(x_c + tp) < f(x_c) + \alpha t \nabla f(x_c)^T p$, die erste zu erfüllende Ungleichung ist also strikt erfüllt. Wegen $\Phi'(t) = 0$ ist $\nabla f(x_c + tp)^T p = \alpha \nabla f(x_c)^T p$ und daher

$$|\nabla f(x_c + tp)^T p| = -\alpha \nabla f(x_c)^T p < -\beta \nabla f(x_c)^T p,$$

also ist auch die zweite zu erfüllende Ungleichung strikt erfüllt und die Aufgabe gelöst.

2. Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Bedingungen (V) (a)–(c) in Unterabschnitt 3.1.1. Sei $x_c \in L_0$ keine stationäre Lösung von (P) und $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x_c, d. h. $\nabla f(x_c)^T p < 0$. Seien $\alpha \in (0, 1)$, $\sigma > 0$ und $\rho \in (0, 1)$ gegeben. Folgendermaßen bestimme man eine Schrittweite $t = t(x_c, p)$:

- Wähle $\tau \geq -\sigma \nabla f(x_c)^T p / \|p\|_2^2$, bestimme die kleinste nichtnegative ganze Zahl j mit
$$f(x_c + \tau\rho^j p) \leq f(x_c) + \alpha\tau\rho^j \nabla f(x_c)^T p$$
und setze $t := \tau\rho^j$.

Man zeige, dass eine von (x_c, p) unabhängige Konstante $\theta > 0$ mit

$$f(x_c) - f(x_c + tp) \geq \theta \left(\frac{\nabla f(x_c)^T p}{\|p\|_2} \right)^2$$

existiert.

Lösung: Wir gehen praktisch genau wie im Beweis von Satz 1.4 vor. Ist $j = 0$, so ist $t = \tau$ und

$$f(x_c) - f(x_c + tp) \geq -\alpha\tau \nabla f(x_c)^T p \geq \alpha\sigma \left(\frac{\nabla f(x_c)^T p}{\|p\|_2} \right)^2.$$

Ist dagegen $j > 0$, so gelten mit $t := \tau\rho^j$ und $s := \tau\rho^{j-1}$ zwei Ungleichungen, nämlich

$$f(x_c + tp) \leq f(x_c) + \alpha t \nabla f(x_c)^T p, \qquad f(x_c + sp) > f(x_c) + \alpha s \nabla f(x_c)^T p.$$

Ferner ist $\rho s = t$. Der Rest des Beweises ist praktisch genau wie im Beweis von Satz 1.4. Wir wiederholen ihn trotzdem hier. Mit $\hat{t}$ wie in Lemma 1.1 machen wir eine Fallunterscheidung. Für $s \leq \hat{t}$ ist

$$f(x_c) + \alpha s \nabla f(x_c)^T p < f(x_c + sp) \leq f(x_c) + s \nabla f(x_c)^T p + s^2 \frac{\gamma}{2} \|p\|_2^2,$$

daher

$$-\frac{2\rho(1-\alpha)}{\gamma} \frac{\nabla f(x_c)^T p}{\|p\|_2^2} \leq \rho s = t$$

und folglich

$$f(x_c) - f(x_c + tp) \geq -\alpha t \nabla f(x_c)^T p \geq \frac{2\alpha\rho(1-\alpha)}{\gamma} \left(\frac{\nabla f(x_c)^T p}{\|p\|_2} \right)^2.$$

Ist dagegen $s > \hat{t}$, so ist wiederum wegen Lemma 1.1

$$-\frac{2\rho \nabla f(x_c)^T p}{\gamma \|p\|_2^2} \leq \rho \hat{t} < \rho s = t$$

und daher

$$f(x_c) - f(x_c + tp) \geq -\alpha t \nabla f(x_c)^T p \geq \frac{2\alpha\rho}{\gamma} \left(\frac{\nabla f(x_c)^T p}{\|p\|_2} \right)^2.$$

Mit

$$\theta := \alpha \min(1, 2\rho(1-\alpha)/\gamma)$$

ist die Aussage der Aufgabe bewiesen.

3. Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Bedingungen (V) (a)–(c) in Unterabschnitt 3.1.1. Sei $x_c \in L_0$ keine stationäre Lösung von (P) und $p \in \mathbb{R}^n$ eine Abstiegsrichtung für f in x_c, d. h. $\nabla f(x_c)^T p < 0$. Bei vorgegebenem $\alpha \in (0, \frac{1}{2})$ definiere man

$$T_G(x_c, p) := \{t > 0 : f(x_c) + (1-\alpha)t\nabla f(x_c)^T p \le f(x_c + tp) \le f(x_c) + \alpha t \nabla f(x_c)^T p\}$$

(Menge der *Goldstein-Schrittweiten*). Analog zu Satz 1.3 zeige man:

(a) Es ist $T_G(x_c, p) \neq \emptyset$.

(b) Es existiert eine von x_c und p unabhängige Konstante $\theta > 0$ mit

$$f(x_c) - f(x_c + tp) \ge \theta \left(\frac{\nabla f(x_c)^T p}{\|p\|_2} \right)^2 \qquad \text{für alle } t \in T_G(x_c, p).$$

Lösung: Zu Abkürzung definiere man

$$\Phi(t) := f(x_c) + \alpha t \nabla f(x_c)^T p - f(x_c + tp).$$

Dann ist $\Phi(0) = 0$ und $\Phi'(0) > 0$. Sei $\tilde{t}$ die erste positive Nullstelle von $\Phi(\cdot)$ (eine solche muss existieren, denn andernfalls hätte man einen Widerspruch zur Kompaktheit der Niveaumenge). Dann ist

$$\begin{aligned} f(x_c + \tilde{t}p) &= f(x_c) + \alpha \tilde{t} \nabla f(x_c)^T p \\ &= f(x_c) + (1-\alpha)\tilde{t}\nabla f(x_c)^T p + \underbrace{(2\alpha - 1)\tilde{t}\nabla f(x_c)^T p}_{>0} \\ &> f(x_c) + (1-\alpha)\tilde{t}\nabla f(x_c)^T p \end{aligned}$$

und folglich $\tilde{t} \in T_G(x_c, p)$, womit der erste Teil der Aufgabe bewiesen ist. Im zweiten Teil geben wir uns ein $t \in T_G(x_c, p)$ vor und machen eine Fallunterscheidung. Im ersten Fall sei $t \le \hat{t}$, wobei $\hat{t}$ wie in Lemma 1.1 die erste positive Nullstelle von $\psi(t) := f(x_c) - f(x_c + tp)$. Dann ist

$$f(x_c) + (1-\alpha)t\nabla f(x_c)^T p \le f(x_c + tp) \le f(x_c) + t\nabla f(x_c)^T p + t^2 \frac{\gamma}{2}\|p\|_2^2,$$

woraus wir

$$t \ge -\frac{2\alpha \nabla f(x_c)^T p}{\gamma \|p\|_2^2}$$

folgern. In diesem ersten Fall ist daher

$$f(x_c) - f(x_c + tp) \ge -\alpha t \nabla f(x_c)^T p \ge \frac{2\alpha^2}{\gamma} \left(\frac{\nabla f(x_c)^T p}{\|p\|_2} \right)^2.$$

Im zweiten Fall ist $\hat{t} < t$, wegen Lemma 1.1 folglich

$$t > -\frac{2\nabla f(x_c)^T p}{\gamma \|p\|_2^2}.$$

Dann ist aber

$$f(x_c) - f(x_c + tp) \ge -\alpha t \nabla f(x_c)^T p \ge \frac{2\alpha}{\gamma} \left(\frac{\nabla f(x_c)^T p}{\|p\|_2} \right)^2.$$

Der zweite Teil der Aufgabe ist daher mit $\theta := 2\alpha^2/\gamma$ bewiesen.

4. Eine Funktion $\phi\colon [a,b] \longrightarrow \mathbb{R}$ heißt unimodal, wenn es genau ein $t^* \in (a,b)$ gibt mit $\phi(t^*) = \min_{t\in[a,b]} \phi(t)$, und wenn ϕ auf $[a,t^*]$ monoton fallend und auf $[t^*,b]$ monoton wachsend ist. Zur Lokalisierung des Minimums t^* der auf $[a,b]$ unimodularen Funktion ϕ betrachte man die Methode vom goldenen Schnitt:

   - Sei $\epsilon > 0$ (gewünschte Genauigkeit) gegeben, setze $F := (\sqrt{5}-1)/2$.
   - Berechne $\left\{ \begin{array}{lcllcl} s & := & a+(1-F)(b-a), & \phi_s & := & \phi(s), \\ t & := & a+F(b-a), & \phi_t & := & \phi(t). \end{array} \right.$
   - Solange $b-a > \epsilon$:
     - Falls $\phi_s > \phi_t$, dann:

       $$a := s, \quad s := t, \quad t := a+F(b-a), \quad \phi_s := \phi_t, \quad \phi_t := \phi(t)$$

     - Andernfalls:

       $$b := t, \quad t := s, \quad s := a+(1-F)(b-a), \quad \phi_t := \phi_s, \quad \phi_s := \phi(s).$$

   - $t^* \approx (a+b)/2$.

   Man beweise, dass dieser Algorithmus nach endlich vielen Schritten mit einem Intervall $[a,b]$ abbricht, das t^* enthält.

   Lösung: Es genügt den ersten Schritt zu betrachten. Durch s und t wird das Intervall $[a,b]$ im goldenen Schnitt geteilt. Ist $\phi(s) > \phi(t)$, so kann t^* nicht in $[a,s]$ liegen, so dass man sich bei der Minimumsuche auf das Intervall $[s,b]$ beschränken kann. Der Witz besteht nun darin, dass t einer der beiden Punkte ist, welche dieses Intervall im goldenen Schnitt teilt, in diesem ist der Funktionswert von ϕ schon bekannt. Dies motiviert die Setzungen $a := s$, $s := t$, $\phi_s := \phi_t$. Ist im zweiten Fall $\phi(s) \leq \phi(t)$, so kann man sich bei der Suche nach dem Minimum t^* auf das Intervall $[a,t]$ beschränken. In jedem Schritt wird die Intervalllänge mit F multipliziert, diese geht daher gegen Null und die Aussage ist bewiesen.

5. Man gebe eine Matlab-Implementation der Methode des goldenen Schnittes an und erprobe sie an den Funktionen[2]

   (a) $\phi(t) := -t/(t^2+c)$ mit $c := 2$,

   (b) $\phi(t) := (t+c)^5 - 2(t+c)^4$ mit $c := 0.004$.

   Hierbei veranschauliche man sich die Funktionen durch einen Plot.

   Lösung: Zunächst geben wir Plots der beiden Funktionen an, siehe Abbildung 5.6. Nun eine Matlab-Implementation der oben angegebenen Methode.

```
function t_stern=GoldenSection(fun,a,b,epsi);
%Input-Parameter:
%              fun    real valued function
%              a,b    interval [a,b], fun should be
%                     unimodal on [a,b]
%              epsi   desired accuracy
%*********************************************************
```

[2] Siehe C. Geiger, C. Kanzow (1999, S. 52).

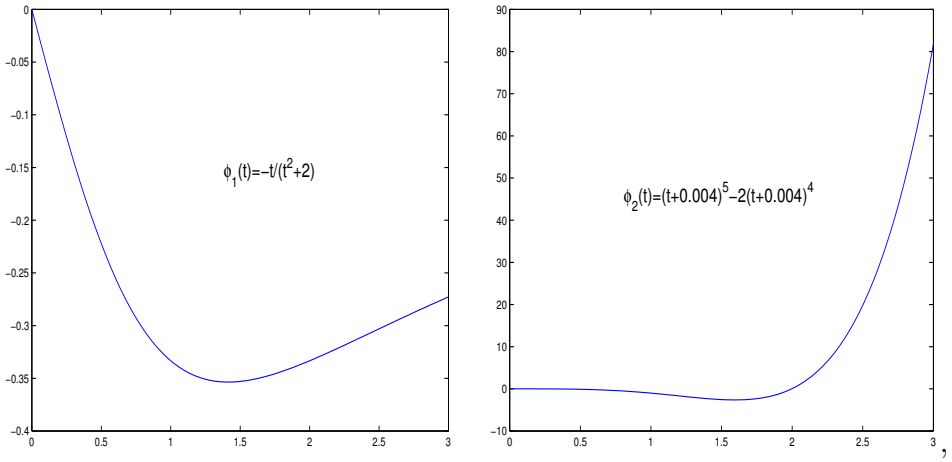

Abbildung 5.6: Zwei unimodale Funktionen

```
F=0.5*(sqrt(5)-1);G=1-F;
s=a+G*(b-a);t=a+F*(b-a);phi_s=feval(fun,s);phi_t=feval(fun,t);
while (b-a>epsi)
   if (phi_s>phi_t)
       a=s;s=t;t=a+F*(b-a);phi_s=phi_t;phi_t=feval(fun,t);
   else
       b=t;t=s;s=a+G*(b-a);phi_t=phi_s;phi_s=feval(fun,s);
   end;
end;
t_stern=0.5*(a+b);
%*************************************************************
function out=phi_1(t);
out=-t/(t^2+2);
%*************************************************************
function out=phi_2(t);
out=(t+0.004)^5-2*(t+0.004);
```

Der Aufruf

```
t_stern=GoldenSection('phi_1',0,3,1e-10);
```

liefert (nach `format long g`) $t^* = 1.41421355853631$, mit der Genauigkeit `3*eps` (hierbei ist `eps` die Maschinengenauigkeit) erhält man $t^* = 1.4142135585311$. Bei der Minimierung von ϕ_2 erhält man $t^* = 0.791270726350009$ bzw. (mit der Genauigkeit `3*eps`) $t^* = 0.791270726342191$. In Matlab gibt es die Funktion `fminbnd` (diese ersetzt die frühere Funktion `fmin` (jeweils ist nicht die Optimization Toolbox nötig). Wir können die zu minimierende Funktion auch `inline` übergeben (das hätten wir auch bei obiger Funktion `GoldenSection` machen können, es hat bei einfachen Funktionen den Vorteil, dass man nicht extra ein M-file zu schreiben braucht):

```
>> phi=inline('-t/(t^2+2)');
>>t_stern=fminbnd(phi,0,3);
```

Als Ergebnis erhalten wir $t^* = 1.41421729307232$, wobei wir darauf hingewiesen werden, dass hier nur mit einer Toleranz von `1e-4` gerechnet wurde. Diese kann man erhöhen, indem man z. B.

```
>> options=optimset('TolX',1e-10);
>> t_stern=fminbnd(phi,0,3,options);
```

eingibt, das Ergebnis ist $t^* = 1.41421356247266$. Man erkennt, dass man lange nicht allen Dezimalen trauen darf.

6. Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Bedingungen (V) (a)–(c) in Unterabschnitt 3.1.1. Zur Lösung von (P) wende man den Modellalgorithmus mit $p_k := -g_k$ (hierbei sei wieder $g_k := \nabla f(x_k)$) und der konstanten Schrittweite $t_k := 1/\gamma$ an. Dann ist jeder Häufungspunkt der durch das Verfahren erzeugten Folge $\{x_k\}$ eine stationäre Lösung von (P).

   Hinweis: Man wende Lemma 1.1 an, um

   $$f(x_k) - f(x_{k+1}) \geq \frac{1}{2\gamma}\|g_k\|^2, \qquad k = 0, 1, \ldots,$$

   zu zeigen und beweise hiermit die Behauptung.

   Lösung: Mit der Spezialisierung $p_k = -g_k$ haben wir in Lemma 1.1 bewiesen, dass

   $$\frac{2}{\gamma} \leq \hat{t}_k$$

   und

   $$f(x_k - tg_k) \leq f(x_k) - t\|g_k\|_2^2 + t^2\frac{\gamma}{2}\|g_k\|_2^2 \qquad \text{für alle } t \in [0, \hat{t}_k],$$

   wobei $\hat{t}_k$ die erste positive Nullstelle von $\psi_k(t) := f(x_k) - f(x_k - tg_k)$. Folglich ist $t_k = 1/\gamma \in [0, \hat{t}_k]$ und daher

   $$f(x_k) - f(x_{k+1}) \geq \frac{1}{2\gamma}\|g_k\|^2, \qquad k = 0, 1, \ldots,$$

   woraus $\lim_{k\to\infty} \|g_k\|_2 = 0$ und dann die Behauptung folgt.

7. Die Voraussetzungen von Satz 1.5 seien erfüllt. Die Zielfunktion f besitze in der Niveaumenge L_0 nur endlich viele stationäre Punkte. Der Modellalgorithmus erzeuge eine Folge $\{x_k\}$ mit $\lim_{k\to\infty}(x_{k+1} - x_k) = 0$. Dann konvergiert die gesamte Folge $\{x_k\}$ gegen einen der stationären Punkte von f.

   Hinweis: Siehe J. M. Ortega, W. C. Rheinboldt (1970, S. 476)[3].

   Lösung: Sei $H \subset L_0$ die Menge der Häufungspunkte von $\{x_k\}$. Da jeder Häufungspunkt von $\{x_k\}$ wegen Satz 1.5 eine stationäre Lösung ist und es von diesen nach Voraussetzung nur endlich viele gibt, ist H ebenfalls endlich. Sei etwa $H = \{z_1, \ldots, z_m\}$. Ist $m = 1$, so konvergiert die gesamte Folge $\{x_k\}$ trivialerweise gegen den einzigen Häufungspunkt. Wir nehmen daher an, es sei $m > 1$, und definieren die positive Zahl

   $$\delta := \min\{\|z_i - z_j\|_2 : i \neq j,\ i, j = 1, \ldots, m\},$$

   also den kleinsten Abstand zwischen zwei Häufungspunkten. Es existiert ein $k_0 \in \mathbb{N}$ mit

   $$x_k \in \bigcup_{i=1}^{m} B[z_i, \delta/4], \qquad \|x_{k+1} - x_k\|_2 \leq \delta/4 \qquad \text{für alle } k \geq k_0,$$

[3] J. M. Ortega, W. C. Rheinboldt (1970) *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York-London.

wobei $B[z_i;\delta/4]$ die (abgeschlossene euklidische) Kugel um z_i mit Radius $\delta/4$ bedeutet. Ist nun $x_{k_1} \in B[z_1;\delta/4]$ für ein $k_1 \geq k_0$, so ist

$$\begin{aligned} \|z_i - x_{k_1+1}\|_2 &\geq \|z_i - z_1\|_2 - (\|z_1 - x_{k_1}\|_2 + \|x_{k_1} - x_{k_1+1}\|_2) \\ &\geq \delta - 2\delta/4 \\ &= \delta/2, \qquad i = 2, \ldots, m. \end{aligned}$$

Folglich ist notwendigerweise $x_{k_1+1} \in B[z_1;\delta/4]$. Durch vollständige Induktion folgt $x_k \in B[z_1;\delta/4]$ für alle $k \geq k_1$, was der Annahme widerspricht, dass $z_2, \ldots, z_m$ Häufungspunkte von $\{x_k\}$ sind. Also ist $m = 1$ und die Aussage bewiesen.

8. Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische, positiv definite Matrix mit kleinstem Eigenwert $\lambda_{\min}$ und größtem Eigenwert $\lambda_{\max}$ Dann gilt die *Ungleichung von Kantorowitsch*:

$$(x^TAx)(x^TA^{-1}x) \leq \frac{(\lambda_{\min}+\lambda_{\max})^2}{4\lambda_{\min}\lambda_{\max}}(x^Tx)^2 \qquad \text{für alle } x \in \mathbb{R}^n.$$

Hinweis: Durch eine orthogonale Ähnlichkeitstranformation kann man erreichen, dass A o. B. d. A. eine Diagonalmatrix ist, siehe auch D. G. Luenberger (1973, S. 151)[4] und C. Geiger, C. Kanzow (1999, S. 71).

Lösung: O. B. d. A. ist $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ mit $\lambda_1 \geq \cdots \geq \lambda_n$. Sei $x \in \mathbb{R}^n \setminus \{0\}$. Dann ist

$$\frac{(x^Tx)^2}{(x^TAx)(x^TA^{-1}x)} = \frac{(\sum_{j=1}^n x_j^2)^2}{(\sum_{j=1}^n \lambda_j x_j^2)(\sum_{j=1}^n (1/\lambda_j)x_j^2)} = \frac{1}{(\sum_{j=1}^n y_j\lambda_j)(\sum_{j=1}^n y_j(1/\lambda_j))},$$

wobei wir

$$y_j := x_j^2 \Big/ \sum_{j=1}^n x_j^2, \qquad j = 1, \ldots, m,$$

gesetzt haben. Sei $\bar\lambda := \sum_{j=1}^n y_j\lambda_j$. Wegen $y_j \geq 0$, $j = 1, \ldots, n$, und $\sum_{j=1}^n y_j = 1$ ist $\bar\lambda$ eine Konvexkombination der λ_j und daher $\bar\lambda \in [\lambda_n, \lambda_1]$. Wir betrachten im $\mathbb{R}^2$ die Punkte

$$P_j := (\lambda_j, (1/\lambda_j)), \qquad j = 1, \ldots, n, \qquad Q := (\bar\lambda, \textstyle\sum_{j=1}^n y_j(1/\lambda_j)).$$

Die Punkte $P_1, P_2, \ldots, P_n$ liegen auf dem Graphen der Funktion $r\colon \mathbb{R}_+ \longrightarrow \mathbb{R}$, definiert durch $r(\lambda) := 1/\lambda$. Da $r(\cdot)$ auf $\mathbb{R}_+$ konvex ist, liegen $P_2, \ldots, P_{n-1}$ unterhalb der Geraden durch P_1 und P_n. Der Punkt Q liegt in der konvexen Hülle von $\{P_1, \ldots, P_n\}$ und somit jedenfalls nicht oberhalb der Geraden $g(\lambda) := (\lambda_1 + \lambda_n - \lambda)/(\lambda_1\lambda_n)$ durch P_1 und P_n, d. h. es ist

$$\sum_{j=1}^n y_j(1/\lambda_j) \leq \frac{\lambda_1 + \lambda_n - \bar\lambda}{\lambda_1\lambda_n}.$$

Hieraus erhält man schließlich

$$\frac{(x^Tx)^2}{(x^TAx)(x^TA^{-1}x)} = \frac{1}{\bar\lambda\sum_{j=1}^n y_j(1/\lambda_j)}$$

[4] D. G. Luenberger (1973) *Introduction to Linear and Nonlinear Programmin*. Addison-Wesley, Reading.

$$\begin{aligned} &\geq \frac{\lambda_1\lambda_n}{\bar{\lambda}(\lambda_1+\lambda_n-\bar{\lambda})} \\ &\geq \min_{\lambda\in[\lambda_n,\lambda_1]} \frac{\lambda_1\lambda_n}{\lambda(\lambda_1+\lambda_n-\lambda)} \\ &= \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}, \end{aligned}$$

was zu zeigen war.

9. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit mit kleinstem Eigenwert $\lambda_{\min}$ und größtem Eigenwert $\lambda_{\max}$. Für alle $x \in \mathbb{R}^n \setminus \{0\}$ ist dann

$$\left(\frac{x^TAx}{\|x\|_2\,\|Ax\|_2}\right)^2 \geq \frac{4\lambda_{\min}\lambda_{\max}}{(\lambda_{\min}+\lambda_{\max})^2} = \frac{4\kappa_2(A)}{(1+\kappa_2(A))^2},$$

wobei $\kappa_2(A)$ natürlich die Kondition von A bezüglich der Spektralnorm bedeutet.

Lösung: Sei $x \in \mathbb{R}^n \setminus \{0\}$ beliebig gegeben, setze $z := A^{1/2}x$. Dann ist

$$\left(\frac{x^TAx}{\|x\|_2\,\|Ax\|_2}\right)^2 = \frac{(z^Tz)^2}{(z^TA^{-1}z)(z^TAz)} \geq \frac{4\lambda_{\min}\lambda_{\max}}{(\lambda_{\min}+\lambda_{\max})^2}$$

wegen der Ungleichung von Kantorowitsch und die Behauptung ist bewiesen.

10. Die Zielfunktion f der unrestringierten Optimierungsaufgabe (P) genüge den Voraussetzungen (K) (a)–(c) aus Lemma 1.6. Auf (P) wende man das Gradientenverfahren (d. h. $p_k := -g_k$, wobei $g_k := \nabla f(x_k)$) mit exakter Schrittweite (d. h. $t_k = t^*(x_k, p_k)$) an. Mit x^* werde die globale Lösung von (P) bezeichnet.

   (a) Mit Hilfe von Satz 1.2 und Lemma 1.6 zeige man, dass

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{c}{\gamma}\right)[f(x_k) - f(x^*)], \qquad k = 0, 1, \ldots.$$

   (b) Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische, positiv definite Matrix mit kleinstem Eigenwert $\lambda_{\min}$ und größtem Eigenwert $\lambda_{\max}$, ferner sei $f(x) := \frac{1}{2}x^TAx - b^Tx$. Dann sind die Voraussetzungen (K) (a)–(c) mit $c := \lambda_{\min}$ und $\gamma := \lambda_{\max}$ erfüllt. Man zeige, dass sich die aus dem vorigen Teil der Aufgabe resultierende Abschätzung zu

$$f(x_{k+1}) - f(x^*) \leq \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right)^2 [f(x_k) - f(x^*)], \qquad k = 0, 1, \ldots$$

   verbessern lässt.

   Hinweis: Man zeige

$$\frac{f(x_k) - f(x^*)}{f(x_k) - f(x_{k+1})} = \frac{(g_k^TAg_k)(g_k^TA^{-1}g_k)}{\|g_k\|_2^4}$$

   und wende die Ungleichung von Kantorowitsch an.

Lösung: Satz 1.2 und Lemma 1.6 liefern die Abschätzungen

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2\gamma}\|g_k\|_2^2 \geq \frac{c}{\gamma}[f(x_k) - f(x^*)], \qquad k = 0, 1, \ldots,$$

woraus die Aussage des ersten Teils der Aufgabe folgt.

Es ist

$$\begin{aligned} f(x_k) - f(x_{k+1}) &= \underbrace{\nabla f(x_{k+1})^T(x_k - x_{k+1})}_{=0} + \frac{1}{2}(x_{k+1} - x_k)^T A(x_{k+1} - x_k) \\ &= t_k^2 g_k^2 A g_k \\ &= \frac{1}{2}\,\frac{\|g_k\|_2^4}{g_k^T A g_k}. \end{aligned}$$

Hierbei haben wir benutzt, dass

$$0 = g_{k+1}^T g_k = (g_k - t_k A g_k)^T g_k = \|g_k\|_2^2 - t)k g_k^T A g_k.$$

Weiter ist

$$\begin{aligned} f(x_k) - f(x^*) &= \underbrace{\nabla f(x^*)}_{=0}{}^T(x_k - x^*) + \frac{1}{2}(x_k - x^*)^T A(x_k - x^*) \\ &= \frac{1}{2}(x_k - A^{-1}b)^T A(x_k - A^{-1}b) \\ &= \frac{1}{2}(Ax_k - b)^T A^{-1}(Ax_k - b) \\ &= \frac{1}{2} g_k^T A^{-1} g_k. \end{aligned}$$

Wendet man die Ungleicung von Kantorowitdch an, so erhält man

$$\frac{f(x_k) - f(x^*)}{f(x_k) - f(x_{k+1})} = \frac{(g_k^T A g_k)(g_k^T A^{-1} g_k)}{\|g_k\|_2^4} \leq \frac{(\lambda_{\min} + \lambda_{\max})^2}{4\lambda_{\min}\lambda_{\max}}.$$

Hieraus folgt nach leichter Rechnung die Behauptung.

11. Sei[5] $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ stetig differenzierbar und gleichmäßig konvex sowie $\nabla f(\cdot)$ global lipschitzstetig auf dem gesamten $\mathbb{R}^n$. Es bezeichne $\gamma > 0$ die zugehörige Lipschitz-Konstante (bezüglich der euklidischen Norm) sowie $c > 0$ die Konstante aus der Definition der gleichmäßigen Konvexität. Dann konvergiert das Gradientenverfahren mit konstanter Schrittweite

$$x_{k+1} := x_k - \alpha \nabla f(x_k), \qquad k = 0, 1, \ldots,$$

für jeden Startvektor $x_0 \in \mathbb{R}^n$, sofern $\alpha \in (0, 2c/\gamma^2)$.

Hinweis: Man wende den Banachschen Fizpunktsatz an.

Lösung: Mit einem $\alpha \in (0, 2c/\gamma^2)$ definieren wir $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ durch

$$F(x) := x - \alpha \nabla f(x).$$

[5]Diese Aufgabe haben wir C. Geiger, C. Kanzow (1999, S. 80) entnommen.

Wir zeigen, dass F bezüglich der euklidischen Norm auf dem $\mathbb{R}^n$ kontrahierend ist, so dass der Banachsche Fixpunktsatz die Behauptung liefert. Hierzu berücksichtigen wir, dass nach Voraussetzung einerseits

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \gamma \|x - y\|_2$$

und anderseits nach Satz 2.2 in Abschnitt 2.2

$$\frac{c}{2}\|y - x\|_2^2 + \nabla f(x)^T(y - x) \leq f(y) - f(x)$$

für alle $x, y \in \mathbb{R}^n$. Vertauscht man in der letzten Ungleichung die Rollen von x und y und addiert die beiden Ungleichungen, so erhält man

$$c\|x - y\|_2^2 \leq [\nabla f(x) - \nabla f(y)]^T(x - y) \qquad \text{für alle } x, y \in \mathbb{R}^n.$$

Für beliebige $x, y \in \mathbb{R}^n$ ist daher

$$\begin{aligned} \|F(x) - F(y)\|_2^2 &= \|x - y\|_2^2 - 2\alpha[\nabla f(x) - \nabla f(y)]^T(x - y) + \alpha^2\|\nabla f(x) - \nabla f(y)\|_2^2 \\ &\leq (1 - 2\alpha c + \alpha^2\gamma^2)\|x - y\|_2^2. \end{aligned}$$

Für $\alpha \in (0, 2c/\gamma^2)$ ist $1 - 2\alpha c + \alpha^2\gamma^2 \in (0, 1)$, womit die Behauptung bewiesen ist.

5.2.2 Aufgaben zu Abschnitt 3.2

1. Mit dem ungedämpften Newton-Verfahren und dem durch die Wolfe- bzw. die Armijo-Schrittweite gedämpften Newton-Verfahren löse man die Aufgabe:

   (P) Minimiere $f(x) := 1.1x_1^2 + 1.2x_2^2 - 2x_1x_2 + \sqrt{1 + x_1^2 + x_2^2} - 7x_1 - 3x_2,$

   wobei man den Startwert $x_0 := (0, 0)^T$ nehme[6].

   Lösung: Wir benutzen

```
function [f,g,B]=Myfun(x);
R=sqrt(1+x(1)^2+x(2)^2);
f=1.1*x(1)^2+1.2*x(2)^2-2*x(1)*x(2)+R-7*x(1)-3*x(2);
if nargout>1
   g=[2.2*x(1)-2*x(2)-7+x(1)/R;-2*x(1)+2.4*x(2)-3+x(2)/R];
end;
if nargout>2
   B=[2.2+(1+x(2)^2)/R^3,-2-x(1)*x(2)/R^3;
          -2-x(1)*x(2)/R^3,2.4+(1+x(1)^2)/R^3];
end;
```

   Die Konvergenz stellt sich in allen Fällen als sehr gut heraus. Im ungedämpften wie im gedämpften Fall (es wird von Anfang an die Schrittweite $t = 1$ gewählt) erhält man

k	x_1	x_2
0	0.00000000000000	0.00000000000000
1	4.33139534883721	3.43023255813954
2	15.19443611974342	13.56594263673561
3	15.37624365606965	13.78570724409425
4	15.37624818227211	13.78572059212680
5	15.37624818227225	13.78572059212699

[6]Siehe P. Spellucci (1993, S. 117).

2. Mit dem durch die Wolfe- bzw. die Armijo-Schrittweite gedämpften Newton-Verfahren löse man die unrestringierte

   (P) $$\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^4,$$

   wobei

   $$\begin{aligned} f(x) \;:=\; & 100(x_1^2 - x_2)^2 + (1 - x_1)^2 + 90(x_3^2 - x_4)^2 + (1 - x_3)^2 \\ & + 10.1[(1 - x_2)^2 + (1 - x_4)^2] + 19.8(1 - x_2)(1 - x_4) \end{aligned}$$

   (die sogenannte Wood-Funktion). Man nehme $x_0 = (-1.5, -1, -3, -1)^T$ und $x_0 = (-3.1, 8.2, 5.5, -3.5)^T$ als Startwerte.

   Lösung: Wir schreiben das File `Myfun.m` mit dem Inhalt:

```
function [f,g,B]=Myfun(x);
f=100*(x(1)^2-x(2))^2+(1-x(1))^2+90*(x(3)^2-x(4))^2+(1-x(3))^2;
f=f+10.1*((1-x(2))^2+(1-x(4))^2)+19.8*(1-x(2))*(1-x(4));
if nargout>1
      g=[400*x(1)*(x(1)^2-x(2))-2*(1-x(1));
      -200*(x(1)^2-x(2))-20.2*(1-x(2))-19.8*(1-x(4));
      360*x(3)*(x(3)^2-x(4))-2*(1-x(3));
      -180*(x(3)^2-x(4))-20.2*(1-x(4))-19.8*(1-x(2))];
end;
if nargout>2
      B=[1200*x(1)^2-400*x(2)+2,-400*x(1),0,0;
      -400*x(1),220.2,0,19.8;
      0,0,1080*x(3)^2-360*x(4)+2,-360*x(3);
      0,19.8,-360*x(3),200.2];
end;
```

   Nach 35 bzw. 18 Iterationen haben wir erreicht, dass die euklidische Norm des Gradienten kleiner als 10^{-12} ist.

3. Seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ und die symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n \times n}$ gegeben. Sei $B_\phi \in \mathbb{R}^{n \times n}$ der Broyden-Update zum Parameter $\phi \in \mathbb{R}$. Man berechne die Eigenwerte von $B_c^{-1/2} B_\phi B_c^{-1/2}$. Hierbei benutze man die Abkürzungen

   $$a := y^T B_c^{-1} y, \qquad b := y^T s, \qquad c := s^T B_c s.$$

   Lösung: Wir setzen $\tilde{s} := B_c^{1/2} s$ und $\tilde{y} := B_c^{-1/2} y$. Dann ist

   $$B_c^{-1/2} B_\phi B_c^{-1/2} = I - \frac{\tilde{s}\tilde{s}^T}{\|\tilde{y}\|_2^2} + \frac{\tilde{y}\tilde{y}^T}{\tilde{y}^T\tilde{s}} + \phi \|\tilde{s}\|_2^2 \left[\frac{\tilde{y}}{\tilde{y}^T\tilde{s}} - \frac{\tilde{s}}{\|\tilde{s}\|_2^2}\right]\left[\frac{\tilde{y}}{\tilde{y}^T\tilde{s}} - \frac{\tilde{s}}{\|\tilde{s}\|_2^2}\right]^T.$$

   Daher hat $B_c^{-1/2} B_\phi B_c^{-1/2}$ den Eigenwert 1 der Vielfachheit $\geq n - 2$ mit Eigenvektoren aus dem orthogonalen Komplement von $\operatorname{span}\{\tilde{s}, \tilde{y}\}$. Weiter ist

   $$B_c^{-1/2} B_\phi B_c^{-1/2} \tilde{s} = \tilde{y}$$

   und (jetzt benutzen wir die in der Aufgabenstellung angegebenen Abkürzungen)

   $$\begin{aligned} B_c^{-1/2} B_\phi B_c^{-1/2} \tilde{y} &= \tilde{y} - \frac{b}{c}\tilde{s} + \frac{a}{b}\tilde{y} + \phi c \left(\frac{a}{b} - \frac{b}{c}\right)\left[\frac{\tilde{y}}{b} - \frac{\tilde{s}}{c}\right] \\ &= -\left[(1-\phi)\frac{b}{c} + \phi\frac{a}{b}\right]\tilde{s} + \left[1 + \frac{a}{b} + \phi\frac{c}{b}\left(\frac{a}{b} - \frac{b}{c}\right)\right]\tilde{y}. \end{aligned}$$

Neben dem $(n-2)$-fachen Eigenwert 1 hat daher $B_c^{-1/2} B_\phi B_c^{-1/2}$ noch die Eigenwerte der 2×2-Matrix

$$R_\phi := \begin{pmatrix} 0 & -f_2(\phi) \\ 1 & 2f_1(\phi) \end{pmatrix},$$

wobei

$$f_1(\phi) := \frac{1}{2}\left[1 + \frac{a}{b} + \phi\frac{c}{b}\left(\frac{a}{b} - \frac{b}{c}\right)\right], \qquad f_2(\phi) := \left[(1-\phi)\frac{b}{c} + \phi\frac{a}{b}\right].$$

Die beiden restlichen Eigenwerte sind daher

$$\lambda_\pm(\phi) := f_1(\phi) \pm [f_1(\phi)^2 - f_2(\phi)]^{1/2}.$$

Übrigens erkennt man hieran, dass

$$\det(B_c^{-1/2} B_\phi B_c^{-1/2}) = \lambda_+(\phi)\lambda_-(\phi) = f_2(\phi) = \left[(1-\phi)\frac{b}{c} + \phi\frac{a}{b}\right]$$

und folglich

$$\det(B_\phi) = \left[(1-\phi)\frac{b}{c} + \phi\frac{a}{b}\right] \det(B_c).$$

Ferner ist B_ϕ genau dann positiv definit, wenn $f_2(\phi) > 0$ bzw. $\phi > -b^2/(ac-b^2)$.

4. Seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ und eine symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n\times n}$ gegeben. Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n \times n$-Matrizen und $\mathcal{S}_+^{n\times n} \subset \mathcal{S}^{n\times n}$ die konvexe Teilmenge der positiv definiten Matrizen. Man zeige, dass der BFGS-Update

$$B_+ := B_c - \frac{(B_c s)(B_c s)^T}{s^T B_c s} + \frac{yy^T}{y^T s}$$

Lösung der Aufgabe

$$\text{(P)} \quad \begin{cases} \text{Minimiere} & \phi(B) := \operatorname{tr}(B_c^{-1/2} B B_c^{-1/2}) - \ln\det(B_c^{-1/2} B B_c^{-1/2}) \quad \text{auf} \\ & M := \{B \in \mathcal{S}^{n\times n} : B \in \mathcal{S}_+^{n\times n},\ Bs = y\} \end{cases}$$

ist.

Hinweis: Diese Aussage findet man bei R. Fletcher (1991)[7]. Man gebe einen alternativen Beweis an, der die Gateaux-Variation und die Konvexität von ϕ benutzt.

Lösung: Die Gateaux-Variation von ϕ existiert in jedem $B \in \mathcal{S}_+^{n\times n}$ und ist eine lineare Abbildung, sie ist mit $P \in \mathcal{S}^{n\times n}$ gegeben durch (siehe das Beispiel im Anschluss an Satz 2.2 in Abschnitt 2.2)

$$\begin{aligned} \phi'(B)P &= \lim_{t\to 0+} \frac{\phi(B+tP) - \phi(B)}{t} \\ &= \operatorname{tr}(B_c^{-1/2} P B_c^{-1/2}) \\ &\qquad - \lim_{t\to 0+} \frac{\ln\det(B_c^{-1/2}(B+tP)B_c^{-1/2}) - \ln\det(B_c^{-1/2} B B_c^{-1/2})}{t} \\ &= \operatorname{tr}(B_c^{-1/2} P B_c^{-1/2}) - \operatorname{tr}(B^{-1/2} P B^{-1/2}) \\ &= \operatorname{tr}((B_c^{-1} - B^{-1})P). \end{aligned}$$

[7] R. Fletcher (1991) "A new variational result for quasi-Newton formulae." SIAM J. Opt. 1, 18–21.

Sei $B \in M$ beliebig. Dann ist

$$\begin{aligned}
\phi(B) - \phi(B_+) &\geq \phi'(B_+)(B - B_+) \\
&\qquad \text{(da } \phi \text{ auf } \mathcal{S}_+^{n\times n} \text{ konvex)} \\
&= \operatorname{tr}((B_c^{-1} - B_+^{-1})(B - B_+)) \\
&= -\operatorname{tr}\left(\left(\left(1 + \frac{y^T B_c^{-1} y}{y^T s}\right)\frac{ss^T}{y^T s} - \frac{s(B_c^{-1}y)^T + (B_c^{-1}y)s^T}{y^T s}\right)(B - B_+)\right) \\
&= 0 \\
&\qquad \text{(da } (B - B_+)s = 0 \text{ und } \operatorname{tr}(AB) = \operatorname{tr}(BA).
\end{aligned}$$

Damit ist die Behauptung bewiesen.

5. Seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ und eine symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n\times n}$ gegeben. Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n \times n$-Matrizen und $\mathcal{S}_+^{n\times n} \subset \mathcal{S}^{n\times n}$ die konvexe Teilmenge der positiv definiten Matrizen, ferner bezeichne $\|\cdot\|_F$ die Frobenius-Norm. Sei schließlich $V \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit mit $Vs = y$ und $H_c := B_c^{-1}$. Man zeige, dass die Inverse H_+ des BFGS-Updates B_+, also

$$H_+ := H_c + \left(1 + \frac{y^T H_c y}{y^T s}\right)\frac{ss^T}{y^T s} - \frac{s(H_c y)^T + (H_c y)s^T}{y^T s},$$

die Lösung der Aufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad \phi(H) := \dfrac{1}{2}\|V^{1/2}(H - H_c)V^{1/2}\|_F^2 \quad \text{auf} \\ \qquad M := \{H \in \mathcal{S}^{n\times n} : H \in \mathcal{S}_+^{n\times n},\ Hy = s\} \end{cases}$$

ist.

Lösung: Wir gehen im Prinzip wie bei der Lösung von Aufgabe 4 vor. Zunächst berechnen wir die Gateaux-Variation von ϕ in einem $H \in \mathcal{S}^{n\times n}$ und zeigen anschließend, dass ϕ sogar auf ganz $\mathcal{S}^{n\times n}$ konvex ist. Für $H, P \in \mathcal{S}^{n\times n}$ existiert

$$\begin{aligned}
\phi'(H)P &= \lim_{t\to 0+} \frac{\phi(H + tP) - \phi(H)}{t} \\
&= \operatorname{tr}(V^{1/2}(H - H_c)V^{1/2}V^{1/2}PV^{1/2}) \\
&= \operatorname{tr}(V(H - H_c)VP)
\end{aligned}$$

wie man unter Benutzung von $\|A\|_F^2 = \operatorname{tr}(A^2)$ für $A \in \mathcal{S}^{n\times n}$ und der Tatsache, dass die Spur unter Ähnlichkeitstransformationen invariant ist, leicht nachweist. Für beliebige $H, K \in \mathcal{S}^{n\times n}$ ist dann

$$\begin{aligned}
\phi'(H)(K - H) &= \operatorname{tr}(V(H - H_c)V(K - H)) \\
&= \operatorname{tr}(V^{1/2}(H - H_c)V^{1/2}V^{1/2}(K - H)V^{1/2}) \\
&= \operatorname{tr}(V^{1/2}(H - H_c)V^{1/2}V^{1/2}(K - H_c)V^{1/2}) \\
&\qquad - \|V^{1/2}(H - H_c)V^{1/2}\|_F^2 \\
&\leq \|V^{1/2}(H - H_c)V^{1/2}\|_F \, \|V^{1/2}(K - H_c)V^{1/2}\|_F \\
&\qquad - \|V^{1/2}(H - H_c)V^{1/2}\|_F^2 \\
&\qquad \text{(wegen } \operatorname{tr}(AB) \leq \|A\|_F \, \|B\|_F \text{ für } A, B \in \mathcal{S}^{n\times n})
\end{aligned}$$

$$\begin{aligned} &\leq \frac{1}{2}[\|V^{1/2}(H-H_c)V^{1/2}\|_F^2 + \|V^{1/2}(K-H_c)V^{1/2}\|_F^2] \\ &\qquad - \|V^{1/2}(H-H_c)V^{1/2}\|_F^2 \\ &\qquad (\text{wegen } ab \leq \tfrac{1}{2}(a^2+b^2)) \\ &= \frac{1}{2}\|V^{1/2}(K-H_c)V^{1/2}\|_F^2 - \frac{1}{2}\|V^{1/2}(H-H_c)V^{1/2}\|_F^2 \\ &= \phi(K) - \phi(H). \end{aligned}$$

Damit ist die Konvexität von ϕ auf $\mathcal{S}^{n\times n}$ bewiesen. Zum Beweis der eigentlichen Behauptung gebe man sich $H \in M$ beliebig vor. Dann ist

$$\begin{aligned} \phi(H) - \phi(H_+) &\geq \phi'(H_+)(H-H_+) \\ &= \operatorname{tr}(V(H_+ - H_c)V(H-H_+)) \\ &= 0 \\ &\qquad (\text{wegen } Vs = y,\ (H-H_+)y = 0 \text{ und } \operatorname{tr}(AB) = \operatorname{tr}(BA)). \end{aligned}$$

Damit ist die Aufgabe gelöst.

6. Seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ und eine symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n\times n}$ gegeben. Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n \times n$-Matrizen und $\mathcal{S}_+^{n\times n} \subset \mathcal{S}^{n\times n}$ die konvexe Teilmenge der positiv definiten Matrizen, ferner bezeichne $\|\cdot\|_F$ die Frobenius-Norm. Man zeige, dass

$$B_+ := B_c - \frac{(B_c s - y)^T s}{(s^T B_c s)^2}(B_c s)(B_c s)^T$$

eine Lösung der Aufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad \phi(B) := \dfrac{1}{2}\|B_c^{-1/2}(B-B_c)B_c^{-1/2}\|_F^2 \quad \text{auf} \\ \qquad M := \{B \in \mathcal{S}^{n\times n} : B \in \mathcal{S}_+^{n\times n},\ s^T B s = y^T s\} \end{cases}$$

ist.

Lösung: Der Beweis ist dem von Aufgabe 5 sehr ähnlich, so dass wir hier kurz sein dürfen. Zunächst aber haben wir zu zeigen, dass die in der Aufgabenstellung angegebene Matrix B_+ überhaupt zu der Restriktionenmenge M von (P) gehört. Offensichtlich gilt die "schwache Quasi-Newton-Gleichung" $s^T B_+ s = y^T s$, so dass die positive Definitheit von B_+ zu zeigen bleibt. Es ist

$$B_c^{-1/2} B_+ B_c^{-1/2} = I - \frac{(B_c s - y)^T s}{(s^T B_c s)^2}(B_c^{1/2} s)(B_c^{1/2} s)^T.$$

Hieraus liest man ab, dass $B_c^{-1/2} B_+ B_c^{-1/2}$ den $(n-1)$-fachen Eigenwert 1 mit Eigenvektoren aus dem orthogonalen Komplement von $\operatorname{span}\{B_c^{1/2} s\}$ besitzt, dass der andere Eigenwert durch $y^T s / s^T B_c s > 0$ mit dem Eigenvektor $B_c^{1/2} s$ gegeben ist. Folglich ist $B_c^{-1/2} B_+ B_c^{-1/2}$ und damit auch B_+ positiv definit. Für ein beliebiges $B \in M$ ist

$$\begin{aligned} \phi(B) - \phi(B_+) &\geq \phi'(B_+)(B-B_+) \\ &= \operatorname{tr}(B_c^{-1}(B_+ - B_c)B_c^{-1}(B-B_+)) \end{aligned}$$

$$
\begin{aligned}
&= -\frac{(B_c s - y)^T s}{(s^T B_c s)^2} \operatorname{tr}(B_c^{-1}(B_c s)(B_c s)^T B_c^{-1}(B - B_+)) \\
&= -\frac{(B_c s - y)^T s}{(s^T B_c s)^2} \operatorname{tr}(s((B - B_+)s)^T) \\
&= -\frac{(B_c s - y)^T s}{(s^T B_c s)^2} \underbrace{s^T (B - B_+) s}_{=0} \\
&= 0.
\end{aligned}
$$

Damit ist die Aufgabe schließlich gelöst.

7. Seien $y, s \in \mathbb{R}^n$ und eine symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n\times n}$ gegeben. Es sei $(y - B_c s)^T s \neq 0$. Man bestimme $\gamma \in \mathbb{R}$ und $u \in \mathbb{R}^n$ so, dass die Matrix $B_+ = B_c + \gamma u u^T$ der Quasi-Newton-Gleichung $B_+ s = y$ genügt. Unter welchen Voraussetzungen ist die so bestimmte Matrix positiv definit?

   Lösung: Eine Matrix $B_+ = B_c + \gamma u u^T$ genügt der Quasi-Newton-Gleichung $B_+ s = y$, wenn $B_c s + \gamma(u^T s)u = y$. Es liegt daher nahe, $u := y - B_c s$ zu wählen und γ aus $\gamma(u^T s) = 1$ zu bestimmen. Insgesamt kommen wir auf die sogenannte SR1-Update-Formel

$$B_+ := B_c + \frac{(y - B_c s)(y - B_c s)^T}{(y - B_c s)^T s}.$$

   Da wir $(y - B_c s)^T s \neq 0$ vorausgesetzt haben, ist B_+ wohldefiniert. Jetzt untersuchen wir, unter welchen Voraussetzungen B_+ positiv definit ist . Es ist

$$B_c^{-1/2} B_+ B_c^{-1/2} = I + \frac{(B_c^{-1/2} y - B_c^{1/2} s)(B_c^{-1/2} y - B_c^{1/2} s)}{(y - B_c s)^T s}.$$

   Hieraus liest man ab, dass 1 ein $(n-1)$-facher Eigenwert mit Eigenvektoren aus dem orthogonalen Komplement zu $\operatorname{span}\{B_c^{-1/2} y - B_c^{1/2}\}$ ist, während der andere Eigenwert

$$\lambda := 1 + \frac{\|B_c^{-1/2} y - B_c^{1/2} s\|_2^2}{(y - B_c s)^T s} = \frac{y^T B_c^{-1} y - y^T s}{y^T s - s^T B_c s}$$

   mit dem Eigenvektor $B_c^{-1/2} y - B_c^{1/2} s$ ist. Daher ist B_+ genau dann positiv definit, wenn

$$\frac{y^T B_c^{-1} y - y^T s}{y^T s - s^T B_c s} > 0.$$

8. Seien $y, s \in \mathbb{R}^n$ mit $y^T s > 0$ und die symmetrische, positiv definite Matrix $B_c \in \mathbb{R}^{n\times n}$ gegeben. Sei $\mathcal{S}^{n\times n}$ der lineare Raum der symmetrischen $n \times n$-Matrizen und $\mathcal{S}_+^{n\times n} \subset \mathcal{S}^{n\times n}$ die konvexe Teilmenge der positiv definiten Matrizen. Man zeige[8], dass der BFGS-Update der skalierten Matrix

$$\hat{B}_c := \frac{y^T B_c^{-1} y}{y^T s} B_c,$$

[8] Siehe

J. E. Dennis, H. Wolkowicz (1993) "Sizing and least-change secant algorithm." SIAM J. Numer. Anal. 30, 1291–1314.

also

$$B_+ := \hat{B}_c - \frac{(\hat{B}_c s)(\hat{B}_c s)^T}{s^T \hat{B}_c s} + \frac{yy^T}{y^T s},$$

Lösung der Aufgabe

$$\text{(P)} \qquad \left\{ \begin{array}{c} \text{Minimiere} \quad \phi(B) := \dfrac{\mathrm{tr}(B_c^{-1/2} B B_c^{-1/2})/n}{\det(B_c^{-1/2} B B_c^{-1/2})^{1/n}} \quad \text{auf} \\ M := \{B \in \mathcal{S}^{n\times n} : B \in \mathcal{S}_+^{n\times n},\ Bs = y\} \end{array} \right.$$

ist.

Hinweis: Zunächst zeige man, dass $\phi(\cdot)$ in jedem $B \in \mathcal{S}_+^{n\times n}$ in jede Richtung $P \in \mathcal{S}^{n\times n}$ richtungsdifferenzierbar ist und die Gateaux-Variation $\phi'(B;\cdot)\colon \mathcal{S}^{n\times n} \longrightarrow \mathbb{R}$ gegeben ist durch (siehe auch Aufgabe 7 in Abschnitt 2.2)

$$\phi'(B;P) = \frac{1}{n \det(B_c^{-1}B)^{1/n}} \left(\mathrm{tr}(B_c^{-1}P) - \frac{\mathrm{tr}(B_c^{-1}B)}{n} \mathrm{tr}(B^{-1}P) \right).$$

Anschließend überlege man sich, dass für beliebige $A, B \in \mathcal{S}_+^{n\times n}$ die Implikation

$$\phi'(B; A-B) \geq 0 \Longrightarrow \phi(B) \leq \phi(A)$$

gilt. Im letzten Schritt zeige man, dass $B_+ \in M$ und $\phi'(B_+; B - B_+) \geq 0$ für alle $B \in M$.

Lösung: Bei der Lösung von Aufgabe 7 in Abschnitt 2.2 haben wir nachgewiesen, dass die durch

$$f(B) := \frac{\mathrm{tr}(B)/n}{\det(B)^{1/n}}$$

definierte Abbildung die (lineare) Gateaux-Variation

$$f'(B;P) = \frac{1}{n \det(B)^{1/n}} \left(\mathrm{tr}(P) - \frac{\mathrm{tr}(B)}{n} \mathrm{tr}(B^{-1/2} P B^{-1/2}) \right)$$

besitzt. Daher ist die Gateaux-Variation von $\phi(\cdot)$ gegeben durch

$$\begin{aligned} \phi'(B;P) &= f'(B_c^{-1/2} B B_c^{-1/2}; B_c^{-1/2} P B_c^{-1/2}) \\ &= \frac{1}{n \det(B_c^{-1/2} B B_c^{-1/2})^{1/n}} \left(\mathrm{tr}(B_c^{-1/2} P B_c^{-1/2}) \right. \\ &\qquad \left. - \frac{\mathrm{tr}(B_c^{-1/2} B B_c^{-1/2})}{n} \mathrm{tr}(B^{-1/2} P B^{-1/2}) \right) \\ &= \frac{1}{n \det(B_c^{-1}B)^{1/n}} \left(\mathrm{tr}(B_c^{-1}P) - \frac{\mathrm{tr}(B_c^{-1}B)}{n} \mathrm{tr}(B^{-1}P) \right). \end{aligned}$$

Für $A, B \in \mathcal{S}_+^{n\times n}$ erhält man wieder mit Hilfe der Ungleichung vom geometrisch-arithmetischen Mittel, dass

$$\phi'(B; A-B) = \frac{1}{n \det(B_c^{-1}B)^{1/n}} \left(\mathrm{tr}(B_c^{-1}(A-B)) - \frac{\mathrm{tr}(B_c^{-1}B)}{n} \mathrm{tr}(B^{-1}(A-B)) \right)$$

$$
\begin{aligned}
&= \frac{1}{n\det(B_c^{-1}B)^{1/n}}\left(\operatorname{tr}(B_c^{-1}A) - \frac{\operatorname{tr}(B_c^{-1}B)}{n}\operatorname{tr}(B^{-1}A)\right)\\
&\le \frac{1}{n\det(B_c^{-1}B)^{1/n}}\left(\operatorname{tr}(B_c^{-1}A) - \operatorname{tr}(B_c^{-1}B)\det(B^{-1}A)^{1/n}\right)\\
&= \frac{\det(B_c^{-1}A)^{1/n}}{\det(B_c^{-1}B)^{1/n}}\left[\frac{\operatorname{tr}(B_c^{-1}A)/n}{\det(B_c^{-1}A)^{1/n}} - \frac{\operatorname{tr}(B_c^{-1}B)/n}{\det(B_c^{-1}B)^{1/n}}\right]\\
&= \frac{\det(B_c^{-1}A)^{1/n}}{\det(B_c^{-1}B)^{1/n}}[\phi(A) - \phi(B)].
\end{aligned}
$$

Damit ist gezeigt, dass es für die eigentliche Behauptung genügt nachzuweisen, dass $\phi'(B_+; B - B_+) \ge 0$ für alle $B \in M$. Sei also $B \in M$ beliebig. Es ist

$$
\begin{aligned}
B_+ &= \hat{B}_c - \frac{(\hat{B}_c s)(\hat{B}_c s)^T}{s^T\hat{B}_c s} + \frac{yy^T}{y^T s}\\
&= \frac{y^T B_c^{-1} y}{y^T s}B_c - \frac{y^T B_c^{-1} y}{y^T s}\,\frac{(B_c s)(B_c s)^T}{s^T B_c s} + \frac{yy^T}{y^T s}
\end{aligned}
$$

und daher

$$
\begin{aligned}
B_+^{-1} &= \hat{B}_c^{-1} + \left(1 + \frac{y^T\hat{B}_c^{-1}y}{y^T s}\right)\frac{ss^T}{y^T s} - \frac{s(\hat{B}_c^{-1}y)^T + (\hat{B}_c^{-1}y)s^T}{y^T s}\\
&= \frac{y^T s}{y^T B_c^{-1} y}B_c^{-1} + 2\frac{ss^T}{y^T s} - \frac{s(B_c^{-1}y)^T + (B_c^{-1}y)s^T}{y^T B_c^{-1} y}.
\end{aligned}
$$

Folglich ist

$$
\operatorname{tr}(B_+^{-1}B) = \frac{y^T s}{y^T B_c^{-1} y}\operatorname{tr}(B_c^{-1}B), \qquad \operatorname{tr}(B_c^{-1}B_+) = \frac{y^T B_c^{-1} y}{y^T s}n,
$$

wobei wir $Bs = y$ berücksichtigt haben. Daher ist schließlich

$$
\begin{aligned}
&\phi'(B_+; B - B_+)\\
&= \frac{1}{n\det(B_c^{-1}B_+)^{1/n}}\left(\operatorname{tr}(B_c^{-1}(B - B_+)) - \frac{\operatorname{tr}(B_c^{-1}B_+)}{n}\operatorname{tr}(B_+^{-1}(B - B_+))\right)\\
&= \frac{1}{n\det(B_c^{-1}B_+)^{1/n}}\Bigg(\underbrace{\operatorname{tr}(B_c^{-1}B) - \frac{\operatorname{tr}(B_c^{-1}B_+)}{n}\operatorname{tr}(B_+^{-1}B)}_{=0}\Bigg)\\
&= 0
\end{aligned}
$$

und daher $\phi(B) \ge \phi(B_+)$. Die Aufgabe ist damit gelöst.

9. Man wende das BFGS-Verfahren mit Wolfe-Schrittweite auf die Minimierung der Funktion

$$
\begin{aligned}
f(x) := {} & 100(x_1^2 - x_2)^2 + (1 - x_1)^2 + 90(x_3^2 - x_4)^2 + (1 - x_3)^2\\
& + 10.1[(1 - x_2)^2 + (1 - x_4)^2] + 19.8(1 - x_2)(1 - x_4)
\end{aligned}
$$

an. Seien $x_0 = (-1.5, -1, -3, -1)^T$ und $x_0 = (-3.1, 8.2, 5.5, -3.5)^T$ die Startwerte.

Lösung: Wir benutzen die Funktion `Myfun` aus der Lösung von Aufgabe 2. Der Aufruf

```
[x,iter]=BFGS('Myfun',[-1.5;-1;-3;-1],100,1e-8);
```

liefert

$$x = \begin{pmatrix} 0.99999999999668 \\ 0.99999999999255 \\ 1.00000000000314 \\ 1.00000000000611 \end{pmatrix}, \qquad \text{iter} = 44.$$

Für den anderen angegebenen Startwert benötigt man mehr als 100 Iterationen. Nach

```
[x,iter]=BFGS('Myfun',[-3.1;8.2;5.5;-3.5],150,1e-8);
```

erhält man

$$x = \begin{pmatrix} 1.00000000000001 \\ 0.99999999999999 \\ 1.00000000000000 \\ 0.99999999999998 \end{pmatrix}, \qquad \text{iter} = 107.$$

10. Auf die unrestringierte Optimierungsaufgabe aus Aufgabe 9 wende man das L-BFGS-Verfahren mit $m = 1, 2, 3, 4$ an, wobei man den Startwert $x_0 = (-1.5, -1, -3, -1)^T$ nehme.

    Lösung: Wir benutzen wieder die Funktion `Myfun` aus der Lösung von Aufgabe 2. Der Aufruf

    ```
    [x,iter]=LBFGS('Myfun',[-1.5;-1;-3;-1],1,500,1e-8);
    ```

    liefert

    $$x = \begin{pmatrix} 0.99999999723471 \\ 0.99999999445741 \\ 1.00000000283756 \\ 1.00000000568391 \end{pmatrix}, \qquad \text{iter} = 254.$$

    Das L-BFGS-Verfahren mit $m = 2$ ergibt:

    $$x = \begin{pmatrix} 1.00000000050082 \\ 1.00000000100238 \\ 0.99999999959062 \\ 0.99999999917253 \end{pmatrix}, \qquad \text{iter} = 179.$$

    Für $m = 3$ ist das entsprechende Ergebnis

    $$x = \begin{pmatrix} 0.99999999975472 \\ 0.99999999950173 \\ 1.00000000017441 \\ 1.00000000035774 \end{pmatrix}, \qquad \text{iter} = 133.$$

    Schließlich ist für $m = 4$ das Ergebnis

    $$x = \begin{pmatrix} 0.99999999999691 \\ 0.99999999999679 \\ 1.00000000000112 \\ 1.00000000000316 \end{pmatrix}, \qquad \text{iter} = 91.$$

5.2.3 Aufgaben zu Abschnitt 3.3

1. Mit dem Verfahren der konjugierten Gradienten von Hestenes-Stiefel löse man die Aufgabe, die Funktion

$$f(x) := x_2^2 + 0.3x_1x_2 + 0.975x_2^2 + 0.01x_1x_3 + x_3^2 + 3x_1 - 4x_2 + x_3$$

auf dem $\mathbb{R}^3$ zu minimieren[9].

Lösung: Wir geben ein:

```
>>A=[2,0.3,0.01;0.3,1.95,0;0.01,0,2];
>>b=[-3;4;-1];L=eye(3);x_0=zeros(3,1);
>>[x,error,iter]=ConGrad(A,b,x_0,L,5,1e-8);
```

Nach `format long` erhalten wir

$$x = \begin{pmatrix} -1.84788193398650 \\ 2.33557157958767 \\ -0.49076059033007 \end{pmatrix}, \qquad \text{error} = 5.303306391588063 \cdot 10^{-18}, \qquad \text{iter} = 3.$$

2. Gegeben sei die quadratische Zielfunktion $f(x) := \frac{1}{2}x^TAx - b^Tx$ mit einer symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$. Seien $p_0, \ldots, p_{n-1} \in \mathbb{R}^n$ konjugiert bezüglich A. Man betrachte das folgende Verfahren zur Minimierung von $f(x)$ auf dem $\mathbb{R}^n$:

   - Wähle $x_0 \in \mathbb{R}^n$, berechne $g_0 := Ax_0 - b$.
   - Für $k = 0, 1, \ldots$:
     - Falls $g_k = 0$, dann: $m := k$, f nimmt in x_m das Minimum an. STOP.
     - Andernfalls berechne

$$t_k := -\frac{g_k^Tp_k}{p_k^TAp_k}, \qquad x_{k+1} := x_k + t_kp_k, \qquad g_{k+1} := g_k + t_kAp_k.$$

Durch vollständige Induktion nach k zeige man: Sind $g_0, \ldots, g_k \neq 0$, ist das Verfahren im k-ten Schritt also noch nicht abgebrochen, so ist x_{k+1} die Lösung der Aufgabe

$$(\mathrm{P}_k) \qquad \text{Minimiere} \quad f(x), \quad x \in x_0 + \operatorname{span}\{p_0, \ldots, p_k\}.$$

Wegen $x_0 + \operatorname{span}\{p_0, \ldots, p_{n-1}\} = \mathbb{R}^n$ bricht das Verfahren also nach $m \leq n$ Schritten mit dem Minimum von f ab.

Hinweis: Nach Konstruktion ist klar, dass $x_{k+1} \in x_0 + \operatorname{span}\{p_0, \ldots, p_k\}$. Man zeige, dass $g_{k+1}^Tp_i = 0$, $i = 0, \ldots, k$, und überlege sich, dass dies die Behauptung impliziert.

Lösung: Durch vollständige Induktion nach k zeigen wir: Sind $g_0, \ldots, g_k \neq 0$, so ist $x_{k+1} \in x_0 + \operatorname{span}\{p_0, \ldots, p_k\}$ und $g_{k+1}^Tp_i = 0$, $i = 0, \ldots, k$. Der Induktionsanfang liegt bei $k = 0$. Es ist

$$x_1 = x_0 + t_0p_0 \in x_0 + \operatorname{span}\{p_0\},$$

ferner

$$g_1^Tp_0 = g_0^Tp_0 + t_0p_0^TAp_0 = g_0^Tp_0 - \frac{g_0^Tp_0}{p_0^TAp_0}p_0^TAp_0 = 0.$$

[9] Siehe P. Spellucci (1993, S. 164).

Der Induktionsanfang ist also gelegt. Nun nehmen wir an, die Behauptung sei für $k-1$ richtig. Es seien also $g_0, \ldots, g_k \neq 0$ und $x_k \in x_0 + \operatorname{span}\{p_0, \ldots, p_{k-1}\}$ und $g_k^T p_i = 0$, $i = 0, \ldots, k-1$. Dann ist

$$x_{k+1} = x_k + t_k p_k \in x_0 + \operatorname{span}\{p_0, \ldots, p_k\}$$

und

$$g_{k+1}^T p_i = (g_k + t_k A p_k)^T p_i = g_k^T p_i + t_k p_k^T A p_i.$$

Wegen der A-Konjugiertheit der Richtungen und der Induktionsannahme ist $g_{k+1}^T p_i = 0$, $i = 0, \ldots, k-1$. Weiter ist

$$g_{k+1}^T p_k = g_k^T p_k - \frac{g_k^T p_k}{p_k^T A p_k} p_k^T A p_k = 0,$$

so dass der Induktionsbeweis abgeschlossen ist. Sei nun $x \in x_0 + \operatorname{span}\{p_0, \ldots, p_k\}$ beliebig. Dann ist

$$\begin{aligned} f(x) - f(x_{k+1}) &= \underbrace{\nabla f(x_{k+1})}_{=g_{k+1}}{}^T (x - x_{k+1}) + \underbrace{\frac{1}{2}(x - x_{k+1})^T A (x - x_{k+1})}_{\geq 0} \\ &\geq g_{k+1}^T (x - x_{k+1}) \\ &= 0, \end{aligned}$$

da $x - x_{k+1} \in \operatorname{span}\{p_0, \ldots, p_k\}$ und $g_{k+1}^T p_i = 0$, $i = 0, \ldots, k$. Folglich ist x_{k+1} Lösung von (P_k). Es ist die eindeutige Lösung, wie man aus obiger Gleichungs-Ungleichungskette abliest. Damit ist die Aufgabe gelöst.

3. Gegeben sei die quadratische Zielfunktion $f(x) := \frac{1}{2} x^T A x - b^T x$ mit einer symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{n \times n}$. Zur Lösung der zugehörigen unrestringierten Optimierungsaufgabe bzw. des linearen Gleichungssystems $Ax = b$ betrachte man die folgende Modifikation des Verfahrens der konjugierten Gradienten, welche sich von diesem dadurch unterscheidet, dass nicht notwendig am Anfang $p_0 := -g_0$ gewählt wird.

   - Wähle $x_0 \in \mathbb{R}^n$, berechne $g_0 := Ax_0 - b$. O. B. d. A. sei $g_0 \neq 0$. Wähle $p_0 \in \mathbb{R}^n$ mit $g_0^T p_0 < 0$.
   - Für $k = 0, 1, \ldots$:
     - Berechne

       $$t_k := -\frac{g_k^T p_k}{p_k^T A p_k}, \qquad x_{k+1} := x_k + t_k p_k, \qquad q_{k+1} := g_k + t_k A p_k.$$

     - Falls $g_{k+1} = 0$, dann: $m := k+1$, STOP. x_m ist die Lösung.
     - Andernfalls:
       * Berechne

         $$p_{k+1} := \begin{cases} -g_1 - \dfrac{g_1^T (g_1 - g_0)}{g_0^T p_0} p_0, & k = 0, \\ -g_{k+1} + \dfrac{g_{k+1}^T g_0}{g_0^T p_0} p_0 + \dfrac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2} p_k, & k = 1, 2, \ldots. \end{cases}$$

Durch vollständige Induktion nach k zeige man: Ist $k \geq 1$ und sind $g_1, \ldots, g_k \neq 0$, so gilt:

(a) Es ist $p_i^T g_k = 0$, $i = 0, \ldots, k-1$,

(b) Es ist $g_k^T p_k = -\|g_k\|_2^2$,

(c) Es ist $g_i^T g_k = 0$, $i = 1, \ldots, k-1$,

(d) Es ist $p_i^T A p_k = 0$, $i = 0, \ldots, k-1$, d. h. die Richtungen $p_0, \ldots, p_k$ sind A-konjugiert.

Insbesondere bricht das Verfahren nach $m \leq n$ Schritten ab.

Lösung: Der Induktionsanfang liegt bei $k = 1$. Es ist

$$p_0^T g_1 = p_0^T (g_0 + t_0 A p_0) = p_0^T g_0 + \left(-\frac{g_0^T p_0}{p_0^T A p_0} \right) p_0^T A p_0 = 0,$$

damit stimmt (a) für $k = 1$. Es ist $g_1^T p_1 = -\|g_1\|_2^2$, denn $g_1^T p_0 = 0$ wurde gerade eben bewiesen. Für $k = 1$ ist (c) leer bzw. trivial, ferner ist

$$p_0^T A p_1 = (A p_0)^T p_1 = \frac{1}{t_0} (g_1 - g_0)^T \left(-g_1 - \frac{g_1^T (g_1 - g_0)}{g_0^T p_0} p_0 \right) = 0,$$

da $g_1^T p_0 = 0$. Damit ist der Induktionsanfang gesichert.

Wir nehmen nun an, die Aussagen seien für k richtig. Es ist $p_k^T g_{k+1} = 0$, da t_k die exakte Schrittweite ist. Für $i = 0, \ldots, k-1$ ist ferner

$$p_i^T g_{k+1} = p_i^T (g_k + t_k A p_k) = p_i^T g_k + t_k p_i^T A p_k = 0$$

wegen der Induktionsannahme für (a) und (d). Damit ist (a) für $k+1$ richtig.

Es ist $g_{k+1}^T p_{k+1} = -\|g_{k+1}\|_2^2$, da $g_{k+1}^T p_0 = 0$ und $g_{k+1}^T p_k = 0$. Damit ist (b) für $k+1$ bewiesen.

Es ist $g_k^T g_{k+1} = 0$, da g_k eine Linearkombination von p_k, p_0 und (für $k \geq 2$ von p_{k-1} ist, so dass die Aussage aus der für $k+1$ richtigen Aussage (a) folgt. Für $i = 1, \ldots, k-1$ ist ferner

$$g_i^T g_{k+1} = g_i^T (g_k + t_k A p_k) = t_k g_i^T A p_k,$$

wobei wir die Induktionsannahme für (c) benutzen. Ferner ist $g_i^T A p_k = 0$, da g_i eine Linearkombination aus p_i, p_0 und (für $i \geq 2$) p_{i-1} ist, so dass die Aussage aus der Induktionsannahme für (d) folgt. Insgesamt ist damit auch (c) für $k+1$ bewiesen.

Es ist

$$\begin{aligned}
p_k^T A p_{k+1} &= (A p_k)^T \left(-g_{k+1} + \frac{g_{k+1}^T g_0}{g_0^T p_0} p_0 + \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2} p_k \right) \\
&= \frac{1}{t_k} (g_{k+1} - g_k)^T \left(-g_{k+1} + \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2} p_k \right) \\
&= \frac{1}{t_k} \left(-\|g_{k+1}\|_2^2 - \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2} g_k^T p_k \right) \\
&= 0,
\end{aligned}$$

wobei wir $(Ap_k)^T p_0 = 0$, $g_{k+1}^T g_k = 0$ und $g_k^T p_k = -\|g_k\|_2^2$ benutzt haben. Für $i = 0, \ldots, k-1$ ist schließlich

$$p_i^T A p_{k+1} = p_i^T A(-g_{k+1}) = (Ap_i)^T(-g_{k+1}) = \frac{1}{t_i}(g_{i+1} - g_i)^T(-g_{k+1}) = 0,$$

so dass schließlich der gesamte Induktionsbeweis abgeschlossen ist.

4. Man zeige: Hat die symmetrische, positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ nur r verschiedene Eigenwerte, so bricht das Verfahren der konjugierte Gradienten, angewandt auf das lineare Gleichungssystem $Ax = b$ bzw. die äquivalente unrestringierte Optimierungsaufgabe, nach spätestens r Iterationsschritten ab.

   Hinweis: Man benutze die im Beweis von Satz 3.3 bewiesene Aussage:

   - Ist x_k die durch das Verfahren der konjugierten Gradienten gewonnene k-te Iterierte, sind $\lambda_1, \ldots, \lambda_n$ die Eigenwerte der symmetrischen, positiv definiten Matrix A und ist $p \in \Pi_k$ ein beliebiges Polynom mit $p(0) = 1$, so ist

     $$\|x_k - x^*\|_A \le \max_{i=1,\ldots,n} |p(\lambda_i)| \, \|x_0 - x^*\|_A.$$

     Hierbei ist die Norm $\|\cdot\|_A$ durch $\|x\|_A := \sqrt{x^T A x}$ auf dem $\mathbb{R}^n$ definiert.

   Lösung: Seien $\mu_1, \ldots, \mu_r$ die r verschiedenen Eigenwerte von A. Man definiere $p \in \Pi_r$ durch

   $$p(\lambda) := \frac{(-1)^r}{\mu_1 \cdots \mu_r}(\lambda - \mu_1) \cdots (\lambda - \mu_r).$$

   Offensichtlich ist $p(0) = 1$, ferner ist $p(\lambda_i) = 0$, $i = 1, \ldots, n$. Aus der oben zitierten Abschätzung folgt $x_r = x^*$ und die Aufgabe ist gelöst.

5. Das Verfahren der konjugierten Gradienten von Polak-Ribière unterscheidet sich von dem Fletcher-Reeves-Verfahren nur darin, dass $\beta_k := g_{k+1}^T(g_{k+1} - g_k)/\|g_k\|_2^2$ (statt $\beta_k := \|g_{k+1}\|_2^2/\|g_k\|_2^2$) gesetzt wird. Man betrachte das dann definierte Polak-Ribière-Verfahren mit exakter Schrittweitenstrategie zur Lösung von

   (P) $$\text{Minimiere} \quad f(x), \quad x \in \mathbb{R}^n.$$

   Man zeige:

   (a) Sind die Voraussetzungen (K) (a)–(c) erfüllt, so liefert das Verfahren, wenn es nicht vorzeitig mit der Lösung x^* von (P) abbricht, eine Folge $\{x_k\}$, die R-linear gegen x^* konvergiert.

   (a) Die Voraussetzungen (V) (a)–(c) seien erfüllt, das Verfahren breche nicht vorzeitig mit einer stationären Lösung von (P) ab. Für die durch das Verfahren erzeugte Folge $\{x_k\}$ gelte $\lim_{k\to\infty} \|x_{k+1} - x_k\|_2 = 0$. Dann ist $\liminf_{k\to\infty} \|g_k\|_2 = 0$, so dass die Folge $\{x_k\}$ wenigstens eine stationäre Lösung von (P) als Häufungspunkt besitzt.

   Hinweis: Man setze

   $$\delta_k := \left(\frac{g_k^T p_k}{\|g_k\|_2 \, \|p_k\|_2}\right)^2 = \frac{\|g_k\|_2^2}{\|p_k\|_2^2}.$$

Für den ersten Teil der Aufgabe zeige man die Existenz einer Konstanten $\delta > 0$ mit $\delta_k \geq \delta$, $k = 0, 1, \ldots$, und wende Satz 1.7 an. Für den zweiten Teil der Aufgabe mache man einen Widerspruchsbeweis. Zum einen ist $\sum_{k=0}^{\infty} \delta_k < \infty$ unter Benutzung von Satz 1.2, zum anderen $1/\delta_{k+1} \leq 1 + 1/\delta_k$ für alle hinreichend großen k und daher $\sum_{k=0}^{\infty} \delta_k = \infty$.

Lösung: Wie angegeben definiere man δ_k. Es ist

$$\begin{aligned}
\|p_{k+1}\|_2 &= \| - g_{k+1} + \beta_k p_k\|_2 \\
&\leq \|g_{k+1}\|_2 + \frac{|g_{k+1}^T(g_{k+1} - g_k)|}{\|g_k\|_2^2}\|p_k\|_2 \\
&\leq \|g_{k+1}\|_2 + \frac{\|g_{k+1}\|_2\, \gamma\, \|x_{k+1} - x_k\|_2}{\|g_k\|_2^2}\|p_k\|_2 \\
&= \left(1 + \gamma t_k \frac{\|p_k\|_2^2}{\|g_k\|_2^2}\right)\|g_{k+1}\|_2 \\
&\leq \left(1 + \frac{\gamma}{c}\right)\|g_{k+1}\|_2.
\end{aligned}$$

Hierbei haben wir am Schluss benutzt, dass

$$0 = g_{k+1}^T p_k = g_k^T p_k + (g_{k+1} - g_k)^T p_k \geq g_k^T p_k + c t_k \|p_k\|_2^2 = -\|g_k\|_2^2 + c t_k \|p_k\|_2^2.$$

Damit ist (a) bewiesen.

Für den zweiten Teil nehmen wir an, es sei $\liminf_{k\to\infty} \|g_k\|_2 > 0$. Dann existiert ein $\epsilon > 0$ mit $\|g_k\|_2 \geq \epsilon$ für alle k. Wegen Satz 1.2 existiert eine Konstante $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \geq \theta\left(\frac{g_k^T p_k}{\|p_k\|_2}\right)^2 = \theta\|g_k\|_2^2\delta_k \geq \theta\epsilon^2\delta_k,$$

woraus $\sum_{k=0}^{\infty} \delta_k < \infty$ folgt. Andererseits erhält man mit $s_k := x_{k+1} - x_k$ die Abschätzung

$$\|p_{k+1}\|_2^2 = \|g_{k+1}\|_2^2 + \beta_k^2\|p_k\|_2^2 \leq \left(1 + \frac{\gamma^2\|s_k\|_2^2\,\|p_k\|_2^2}{\|g_k\|_2^4}\right)\|g_{k+1}\|_2^2$$

und hieraus (wegen $\|g_k\|_2 \geq \epsilon$ und $\lim_{k\to\infty} \|s_k\|_2 = 0$)

$$\frac{1}{\delta_{k+1}} = \frac{\|p_{k+1}\|_2^2}{\|g_{k+1}\|_2^2} \leq 1 + \frac{\gamma^2}{\epsilon^2}\|s_k\|_2^2\frac{\|p_k\|_2^2}{\|g_k\|_2^2} \leq 1 + \frac{\|p_k\|_2^2}{\|g_k\|_2^2} = 1 + \frac{1}{\delta_k}$$

für alle hinreichend großen k. Also existiert ein $k_0 \in \mathbb{N}$ mit

$$\frac{1}{\delta_{k_0+j}} \leq j + \frac{1}{\delta_{k_0}}, \qquad j = 0, 1, \ldots.$$

Da die harmonische Reihe divergiert, erhalten wir hieraus $\sum_{k=0}^{\infty} \delta_k = \infty$ und insgesamt den erwünschten Widerspruch.

6. Unter den Voraussetzungen (V) (a)–(c) aus Unterabschnitt 3.1.1 betrachte man zur Lösung der unrestringierten Optimierungsaufgabe (P) das Fletcher-Reeves-Verfahren mit der strengen Wolfe-Schrittweite (siehe Aufgabe 1 in Abschnitt 3.1):

   - Für die Schrittweitenstrategie seien α und β mit $0 < \alpha < \beta < \frac{1}{2}$ gegeben.

- Gegeben $x_0 \in \mathbb{R}^n$, berechne $g_0 := \nabla f(x_0)$ und setze $p_0 := -g_0$.
- Für $k = 0, 1, \ldots$:
  - Falls $g_k = 0$, dann: STOP. x_k ist stationäre Lösung von (P).
  - Andernfalls:
    * Bestimme eine strenge Wolfe-Schrittweite, also ein $t_k > 0$ mit
    $$f(x_k + t_k p_k) \le f(x_k) + \alpha t_k g_k^T p_k, \qquad |\nabla f(x_k + t_k p_k)^T p_k| \le -\beta g_k^T p_k.$$
    * Setze bzw. berechne
    $$x_{k+1} := x_k + t_k p_k, \qquad g_{k+1} := \nabla f(x_{k+1})$$
    sowie
    $$\beta_k := \frac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2}, \qquad p_{k+1} := -g_{k+1} + \beta_k p_k.$$

Man zeige:

(a) Ist im k-ten Schritt noch kein Abbruch erfolgt, ist also $g_0, \ldots, g_k \neq 0$, so ist
$$-\frac{1}{1-\beta} \le -\sum_{j=0}^{k} \beta^j \le \frac{g_k^T p_k}{\|g_k\|_2^2} \le -2 + \sum_{j=0}^{k} \beta^j < -\frac{1-2\beta}{1-\beta}.$$
Wegen $\beta \in (0, \frac{1}{2})$ ist daher p_k eine Abstiegsrichtung in x_k. Mit Hilfe von Aufgabe 1 in Abschnitt 3.1 folgt die Existenz einer Schrittweite $t_k > 0$ mit den geforderten Eigenschaften.

(b) Bricht das Verfahren nicht vorzeitig mit einer stationären Lösung ab, so erzeugt es eine Folge $\{x_k\}$ mit $\liminf_{k\to\infty} \|g_k\|_2 = 0$. Sind sogar die Voraussetzungen (K) (a)–(c) erfüllt, so konvergiert die gesamte Folge $\{x_k\}$ gegen die dann eindeutige Lösung x^* von (P).

Lösung: Den ersten Teil (a) der Aufgabe zeigen wir durch vollständige Induktion nach k, wobei die außenstehenden Ungleichungen trivialerweise richtig sind. Für $k = 0$ lautet die behauptete Ungleichungskette
$$-1 \le \frac{g_0^T p_0}{\|g_0\|_2^2} \le -1,$$
die wegen $p_0 = -g_0$ offensichtlich richtig ist. Nun nehmen wir an, die Ungleichungskette gelte für k. Dann ist
$$\begin{aligned}
\frac{g_{k+1}^T p_{k+1}}{\|g_{k+1}\|_2^2} &= \frac{-\|g_{k+1}\|_2^2 + \beta_k g_{k+1}^T p_k}{\|g_{k+1}\|_2^2} \\
&= -1 + \frac{g_{k+1}^T p_k}{\|g_k\|_2^2} \\
&\le -1 - \beta \frac{g_k^T p_k}{\|g_k\|_2^2} \\
&\le -1 + \beta \sum_{j=0}^{k} \beta^j \\
&= -2 + \sum_{j=0}^{k+1} \beta^j.
\end{aligned}$$

Ferner ist

$$\frac{g_{k+1}^T p_{k+1}}{\|g_{k+1}\|_2^2} = -1 + \frac{g_{k+1}^T p_k}{\|g_k\|_2^2} \geq -1 + \beta \frac{g_k^T p_k}{\|g_k\|_2^2} \geq -1 - \beta \sum_{j=0}^{k} \beta^j = -\sum_{j=0}^{k+1} \beta^j.$$

Damit ist (a) bewiesen. Im zweiten Teil argumentieren wir wie im Beweis von Satz 3.4 und nehmen im Widerspruch zur Behauptung an,es gäbe ein $\epsilon > 0$ mit $\|g_k\|_2 \geq \epsilon$ für alle k. Eine Anwendung von Aufgabe 1 in Abschnitt 3.1 liefert die Existenz einer von k unabhängigen Konstanten $\theta > 0$ mit

$$f(x_k) - f(x_{k+1}) \geq \theta \left(\frac{g_k^T p_k}{\|p_k\|_2} \right)^2, \qquad k = 0, 1, \ldots.$$

Wegen der rechten Ungleichung in (a) ist

$$(*) \qquad f(x_k) - f(x_{k+1}) \geq \theta \left(\frac{1 - 2\beta}{1 - \beta} \right)^2 \frac{1}{\alpha_k} \qquad \text{mit} \quad \alpha_k := \frac{\|p_k\|_2^2}{\|g_k\|_2^4}.$$

Für $k = 1, 2, \ldots$ ist dann mit der linken Ungleichung in (a)

$$\begin{aligned}
\alpha_k &= \frac{\|p_k\|_2^2}{\|g_k\|_2^4} \\
&= \frac{\| - g_k + \beta_{k-1} p_{k-1}\|_2^2}{\|g_k\|_2^2} \\
&= \frac{\|g_k\|_2^2 - 2\beta_{k-1} g_k^T p_{k-1} + \beta_{k-1}^2 \|p_{k-1}\|_2^2}{\|g_k\|_2^4} \\
&\leq \frac{1}{\|g_k\|_2^2} + \frac{2\beta_{k-1} |\nabla g_k^T p_{k-1}| + \beta_{k-1}^2 \|p_{k-1}\|_2^2}{\|g_k\|_2^4} \\
&= \left(1 + 2 \frac{|\nabla g_k^T p_{k-1}|}{\|g_{k-1}\|_2^2} \right) \frac{1}{\|g_k\|_2^2} + \alpha_{k-1} \\
&\leq \left(1 - 2\beta \frac{g_{k-1}^T p_{k-1}}{\|g_{k-1}\|_2^2} \right) \frac{1}{\|g_k\|_2^2} + \alpha_{k-1} \\
&\leq \left(1 + \frac{2\beta}{1 - \beta} \right) \frac{1}{\|g_k\|_2^2} + \alpha_{k-1} \\
&= \left(\frac{1 + \beta}{1 - \beta} \right) \frac{1}{\|g_k\|_2^2} + \alpha_{k-1} \\
&\leq \left(\frac{1 + \beta}{1 - \beta} \right) \sum_{j=0}^{k} \frac{1}{\|g_j\|_2^2} \leq \frac{1}{\epsilon^2} \left(\frac{1 + \beta}{1 - \beta} \right) (k + 1)
\end{aligned}$$

und daher wegen $(*)$

$$f(x_k) - f(x_{k+1}) \geq \theta \epsilon^2 \left(\frac{1 - 2\beta}{1 - \beta} \right)^2 \left(\frac{1 - \beta}{1 + \beta} \right) \frac{1}{k + 1}, \qquad k = 0, 1, \ldots.$$

Wegen der Divergenz der harmonischen Reihe erhält man hieraus $\lim_{k \to \infty} f(x_k) = -\infty$, ein Widerspruch zu den Voaraussetzungen (V).

Sind sogar die Voraussetzungen (V) (a)–(c) erfüllt, so folgt genau wie im Beweis zu Satz 3.4 die Konvergenz der Folge $\{x_k\}$ gegen die dann eindeutige Lösung x^*.

5.2.4 Aufgaben zu Abschnitt 3.4

1. Man beweise Satz 4.3, also die folgende Aussage: Gegeben sei die diskrete, nichtlineare Approximationsaufgabe

   (P) $$\text{Minimiere} \quad f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n,$$

   die Zielfunktion f genüge den Voraussetzungen (V) (a)–(c). Sei $x_c \in L_0$ und $p \in \mathbb{R}^n$ eine Richtung mit $f_c(p) := \|F(x_c) + F'(x_c)p\| < f(x_c)$. Seien $\alpha \in (0, \frac{1}{2})$, $0 < l \le u < 1$ gegeben und $t := \rho_j$ eine zugehörige Armijo-Schrittweite. Dann existiert eine Konstante $\theta > 0$, die nur von α, γ sowie l und u, nicht aber von x_c oder p abhängt, mit

   $$f(x_c) - f(x_c + tp) \ge \theta\left[f(x_c) - f_c(p), \left(\frac{f(x_c) - f_c(p)}{\|p\|}\right)^2\right].$$

   Lösung: Wir folgen, wie schon angedeutet, fast wörtlich dem Beweis von Satz 1.4. Ist $j = 0$ bzw. $t = \rho_0 = 1$, so ist

   $$f(x_c + tp) \le f(x_c) + \alpha[f_c(p) - f(x_c)]$$

   bzw.

   $$f(x_c) - f(x_c + tp) \ge \alpha[f(x_c) - f_c(p)].$$

   Ist dagegen $j > 0$, so gelten mit $s := \rho_{j-1}$ zwei Ungleichungen:

   $$f(x_c + tp) \le f(x_c) + \alpha t[f_c(p) - f(x_c)], \qquad f(x_c + sp) > f(x_c) + \alpha s[f_c(p) - f(x_c)].$$

   Ferner ist $ls \le t$. Sei t^* wie in Lemma 4.2 definiert und $\hat{t} := \min(1, t^*)$. Wir machen eine Fallunterscheidung. Für $s \le \hat{t}$ liefert Lemma 4.2, dass

   $$f(x_c) + \alpha s[f_c(p) - f(x_c)] < f(x_c + sp) \le f(x_c) + s[f_c(p) - f(x_c)] + t^2\frac{\gamma}{2}\|p\|^2,$$

   daher

   $$\frac{2l(1-\alpha)[f(x_c) - f_c(p)]}{\gamma\|p\|^2} \le ls \le t$$

   und folglich

   $$f(x_c) - f(x_c + tp) \ge \alpha t[f(c_c) - f_c(p)] \ge \frac{2\alpha l(1-\alpha)}{\gamma}\left(\frac{f(x_c) - f_c(p)}{\|p\|}\right)^2.$$

   Ist dagegen $s > \hat{t}$, so ist notwendig $\hat{t} = t^*$ (wäre $\hat{t} = 1$, so wäre $s > 1$, was nicht möglich ist), daher

   $$\frac{2l[f(x_c) - f_c(p)]}{\gamma\|p\|^2} \le l\hat{t} < ls \le t$$

   und folglich

   $$f(x_c) - f(x_c + tp) \ge \alpha t[f(x_c) - f_c(p)] \ge \frac{2\alpha l}{\gamma}\left(\frac{f(x_c) - f_c(p)}{\|p\|}\right)^2.$$

   Mit

   $$\theta := \alpha \min(1, 2l(1-\alpha)/\gamma)$$

   ist die Aussage bewiesen.

2. Gegeben sei das nichtlineare Ausgleichsproblem (siehe P. Spellucci (1993, S. 199))

$$\text{Minimiere} \quad f(x) := \|F(x)\|_2, \quad x \in \mathbb{R}^4,$$

wobei $F\colon \mathbb{R}^4 \longrightarrow \mathbb{R}^{11}$ definiert ist durch

$$F_i(x) := \frac{x_1 + x_2 t_i}{1 + x_3 t_i + x_4 t_i^2} - y_i, \qquad i = 1, \ldots, 11,$$

mit den Daten

i	t_i	y_i
1	4.0000	0.0489250
2	2.0000	0.0973500
3	1.0000	0.1735000
4	0.5000	0.3200000
5	0.2500	0.3376000
6	0.1670	0.3754491
7	0.1250	0.3648000
8	0.1000	0.3420000
9	0.0823	0.3924666
10	0.0714	0.3291317
11	0.0625	0.3936000

Man berechne eine Lösung mit Hilfe des gedämpften Gauß-Newton-Verfahrens.

Lösung: Wir schreiben ein File `Spell.m` mit dem Inhalt:

```
function [F,J]=Spell(x);
t=[4.0;2.0;1.0;0.5;0.25;0.167;0.125;0.1;0.0823;0.0714;0.0625];
y=[0.048925;0.09735;0.1735;0.32;0.3376;0.3754491;0.3648;0.342;
    0.3924666;0.3291317;0.3936];
nom=x(1)+x(2)*t; denom=1+x(3)*t+x(4)*(t.^2);
F=nom./denom -y;
if nargout>1
    J=[1./denom, t./denom, -(nom.*t)./(denom.^2),-(nom.*(t.^2))./(denom.^2)];
end;
```

Der Aufruf

```
>> [x,iter]=GauNew('Spell',[0;0;0;0],100,1e-6);
```

liefert

$$x = \begin{pmatrix} 0.3563 \\ 0.2678 \\ 0.2681 \\ 2.0117 \end{pmatrix}, \qquad \text{iter} = 11.$$

Dagegen ergibt der Aufruf

```
>> [x,iter]=GauNew('Spell',[0;0;0;0],100,1e-8);
```

das Resultat

$$\begin{pmatrix} 0.3565 \\ 0.2641 \\ 0.2649 \\ 1.9949 \end{pmatrix}, \qquad \text{iter} = 17.$$

Wir sehen also, dass wir obigen Resultaten nicht recht trauen können. Benutzt man übrigens die Funktion `lsqnonlin` aus der optimization toolbox, so erhält man nach `x=lsqnonlin('Spell',[0;0;0;0]);` als Näherungslösung

$$x = \begin{pmatrix} 0.3569 \\ 0.2275 \\ 0.1955 \\ 1.8667 \end{pmatrix}.$$

Dies ist also ein Resultat, das mit obigen Ergebnissen nicht recht vereinbar ist. Setzen wir allerdings

```
>> options=optimset('TolX',1e-10,'TolFun',1e-10,'LargeScale','off');
>> x=lsqnonlin('Spell',[0;0;0;0],[],[],options);
```

so erhalten wir

$$x = \begin{pmatrix} 0.3566 \\ 0.2637 \\ 0.2646 \\ 1.9930 \end{pmatrix},$$

was schon eher mit unseren Ergebnissen übereinstimmt.

3. Sei $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$. Es existiere ein $\sigma > 0$ mit

$$\sigma\|p\|_2 + \|b\|_2 \le \|b + Ap\|_2 \qquad \text{für alle } p \in \mathbb{R}^n.$$

Man zeige, dass dann Rang $(A) = n$ und $b = 0$.

Lösung: Aus $Ap = 0$ folgt aus der Voraussetzung unmittelbar, dass $p = 0$. Die n Spalten von A sind also linear unabhängig und daher Rang $(A) = n$. Für ein beliebiges $p \neq 0$ ist $\sigma\|p\|_2 + \|b\|_2 \le \|b + Ap\|_2$ und daher

$$\sigma^2\|p\|_2^2 + 2\sigma\|p\|_2\,\|b\|_2 \le \|Ap\|_2^2 + 2b^T Ap.$$

Ersetzt man hier p durch tp mit $t \neq 0$ und teilt die entstehende Ungleichung durch $|t|$, so erhält man

$$|t|\,\|Ap\|_2^2 + \text{sign}\,(t)2b^T Ap \ge \sigma^2|t|\,\|p\|_2^2 + 2\sigma\|p\|_2\,\|b\|_2.$$

Mit $t \to 0+$ bzw. $t \to 0-$ erhält man

$$b^T Ap \ge \sigma\|p\|_2\,\|b\|_2 \qquad \text{bzw.} \qquad -b^T Ap \ge \sigma\|p\|_2\,\|b\|_2.$$

Wegen $p \neq 0$ folgt hieraus $b = 0$. Die Aufgabe ist gelöst.

4. Sei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ mit $m \geq n$ eine stetig differenzierbare Abbildung. Die Abbildungen $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ und $h\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ seien definiert durch

$$f(x) := \|F(x)\|_2, \qquad h(x) := \frac{1}{2}\|F(x)\|_2^2.$$

Bei gegebenem $x_c \in \mathbb{R}^n$ sei $p^* \in \mathbb{R}^n$ eine Lösung des in x_c linearisierten Ausgleichsproblems

$$(\mathrm{LP}_c) \qquad \text{Minimiere} \quad f_c(p) := \|F(x) + F'(x)p\|_2, \quad p \in \mathbb{R}^n.$$

Man zeige:

(a) Es ist $\nabla h(x_c)^T p^* = f_c(p^*)^2 - f(x_c)^2$.
Hinweis: Es gilt die Normalgleichung $F'(x_c)^T[F(x_c) + F'(x_c)p^*] = 0$.

(b) Ist $f_c(p^*) < f(x_c)$, ist also p^* eine Abstiegsrichtung (für f und h) in x_c, und sind $\alpha, \rho \in (0,1)$, so impliziert

$$h(x_c + \rho p^*) \leq h(x_c) + \alpha\rho\nabla h(x_c)^T p^*,$$

dass

$$f(x_c + \rho p^*) \leq f(x_c) + \alpha\rho[f_c(p^*) - f(x_c)].$$

Lösung: Es ist einerseits

$$\begin{aligned} f_c(p^*)^2 - f(x_c)^2 &= \|F(x_c) + F'(x_c)p^*\|_2^2 - \|F(x_c)\|_2^2 \\ &= 2F(x_c)^T F'(x_c)p^* + \|F'(x_c)p^*\|_2^2 \\ &= -\|F'(x_c)p^*\|_2^2, \end{aligned}$$

andererseits ist

$$\nabla h(x_c)^T p^* = [F'(x_c)^T F(x_c)]^T p^* = -\|F'(x_c)p^*\|_2^2,$$

wobei wir beide Male die Normalgleichung benutzt haben. Damit ist der erste Teil schon bewiesen.

Nun seien $\alpha, \rho \in (0,1)$ und

$$h(x_c + \rho p^*) \leq h(x_c) + \alpha\rho\nabla h(x_c)^T p^*.$$

Wegen $h(x) = \frac{1}{2}f(x)^2$ und dem gerade eben bewiesenen ersten Teil der Aufgabe bedeutet dies, dass

$$\frac{1}{2}[f(x_c + \rho p^*) - f(x_c)][f(x_c + \rho p^*) + f(x_c)] \leq \alpha\rho[f_c(p^*)^2 - f(x_c)^2].$$

Hieraus wiederum folgt

$$\begin{aligned} f(x_c + \rho p^*) - f(x_c) &\leq 2\alpha\rho\frac{f_c(p^*)^2 - f(x_c)^2}{f(x_c + \rho p^*) + f(x_c)} \\ &\leq \alpha\rho\frac{f_c(p^*)^2 - f(x_c)^2}{f(x_c)} \\ &= \alpha\rho\left(\frac{f_c(p^*)^2}{f(x_c)} - f(x_c)\right) \\ &\leq \alpha\rho[f_c(p^*) - f(x_c)]. \end{aligned}$$

Damit ist auch der zweite Teil der Aufgabe bewiesen.

5. Man beweise den Satz von Carathéodory: Sei $S \subset \mathbb{R}^n$ und

$$\mathrm{co}(S) := \left\{ \sum_{i=1}^{m} \lambda_i x_i : x_i \in S,\ \lambda_i \geq 0\ (i = 1, \ldots, m),\ \sum_{i=1}^{m} \lambda_i = 1,\ m \in \mathbb{N} \right\}$$

die Menge aller Konvexkombinationen von Punkten aus S. Dann lässt sich jedes $x \in \mathrm{co}(S)$ als Konvexkombination von höchstens $n+1$ Punkten aus S darstellen. Zu jedem $x \in \mathrm{co}(S)$ existiert also ein $m \in \mathbb{N}$ mit $m \leq n+1$ sowie $\mu_i \geq 0$, $x_i \in S$, $i = 1, \ldots, m$, mit $\sum_{i=1}^{m} \mu_i = 1$ und $x = \sum_{i=1}^{m} \mu_i x_i$.

Hinweis: Sei $x \in \mathrm{co}(S)$ Konvexkombination von m Punkten $x_1, \ldots, x_m$ aus S. Man zeige: Ist $m > n+1$, so ist x auch Konvexkombination von $m-1$ Punkten aus S (nach endlich vielen Schritten hat man dann die Behauptung bewiesen). Hierzu benutze man, dass $n+1$ (und mehr) Vektoren des $\mathbb{R}^n$, etwa $\{x_1 - x_m, \ldots, x_{m-1} - x_m\}$, linear abhängig sind und folglich der Nullvektor des $\mathbb{R}^n$ sich als nichttriviale Linearkombination von $x_1, \ldots, x_m$ darstellen lässt.

Lösung: Wir folgen genau dem Hinweis. Als Element von $\mathrm{co}(S)$ lässt sich x darstellen als

$$x = \sum_{i=1}^{m} \lambda_i x_i,$$

wobei

$$\lambda_i \geq 0,\ x_i \in S \quad (i = 1, \ldots, m), \qquad \sum_{i=1}^{m} \lambda_i = 1.$$

Wir zeigen: Ist $m > n+1$, so lässt sich x als Konvexkombination von $m-1$ Punkten aus S darstellen. Hieraus folgt dann die Behauptung.

O. B. d. A. ist $\lambda_i > 0$, $i = 1, \ldots, m$. Die $m - 1 > n$ Vektoren $\{x_1 - x_m, \ldots, x_{m-1} - x_m\}$ sind linear abhängig. Daher existieren reelle Zahlen $r_1, \ldots, r_{m-1}$, die nicht alle gleich Null sind, mit $\sum_{i=1}^{m-1} r_i (x_i - x_m) = 0$. Definiert man $r_m := -\sum_{i=1}^{m-1} r_i$, so ist daher

$$\sum_{i=1}^{m} r_i = 0, \qquad \sum_{i=1}^{m} r_i x_i = 0.$$

Anschließend definiere man

$$\alpha := \min_{\substack{i=1,\ldots,m \\ r_i > 0}} \frac{\lambda_i}{r_i} = \frac{\lambda_j}{r_j}.$$

Hierbei beachte man, dass die r_i nicht alle verschwinden und ihre Summe gleich Null ist, so dass es negative und positive unter ihnen gibt. Nun sei

$$\mu_i := \lambda_i - \alpha r_i, \qquad i = 1, \ldots, m.$$

Dann ist

$$mu_i \geq 0 \quad (i = 1, \ldots, m), \qquad \sum_{i=1}^{m} \mu_i = 1, \qquad \mu_j = 0.$$

Folglich ist

$$x = \sum_{i=1}^{m} \lambda_i x_i = \sum_{i=1}^{m} \mu_i x_i + \alpha \underbrace{\sum_{i=1}^{m} r_i x_i}_{=0} = \sum_{\substack{i=1 \\ i \neq j}}^{m} \mu_i x_i,$$

womit die gewünschte Darstellung von x als Konvexkombination von $m-1$ Elementen aus S gegeben ist.

5.3 Aufgaben zu Kapitel 4

5.3.1 Aufgaben zu Abschnitt 4.1

1. Sei $f \in \mathbb{R}$, $g \in \mathbb{R}^n \setminus \{0\}$ und $\Delta > 0$. Man gebe eine Lösung von

   $$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := f + g^T p, \quad \|p\|_\infty \le \Delta$$

   an und begründe dies.

   Lösung: Man definiere $p^* = (p_j^*)$ durch

   $$p_j^* := -\text{sign}\,(g_j)\Delta_c, \qquad j = 1, \ldots, n.$$

   Hierbei ist es gleichgültig, wie sign (0) zu verstehen ist. Dann ist

   $$\phi(p^*) = f + g^T p^* = f + \sum_{j=1}^n g_j p_j^* = f_c - \Delta \sum_{j=1}^n |g_j| = f_c - \Delta \|g\|_1.$$

   Für ein beliebiges $p \in \mathbb{R}^n$ mit $\|p\|_\infty \le \Delta$ ist andererseits

   $$\phi(p) = f + g^T p \ge f - \sum_{j=1} |g_j|\,|p_j| \ge f - \Delta \sum_{j=1} |g_j| = f_c - \Delta_c \|g\|_1.$$

   Daher ist das angegebene p^* eine Lösung von (P).

2. Man betrachte die unrestringierte Optimierungsaufgabe

   $$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T B p, \quad p \in \mathbb{R}^n,$$

   wobei $f \in \mathbb{R}$, $g \in \mathbb{R}^n$ und $B \in \mathbb{R}^{n \times n}$ eine symmetrische Matrix ist. Man zeige:

   (a) (P) besitzt genau dann eine Lösung, wenn B positiv semidefinit und $g \in \text{Bild}\,(B)$ ist.

   (b) (P) besitzt genau dann eine eindeutige Lösung, wenn B positiv definit ist.

   Lösung: Angenommen, (P) besitzt eine lokale Lösung p^*. Diese muss den notwendigen Bedingungen zweiter Ordnung genögen, d. h. es ist $\nabla\phi(p^*) = 0$ bzw. $g + Bp^* = 0$, folglich $g \in \text{Bild}\,(B)$, und $\nabla^2\phi(p^*) = B$ ist positiv semidefinit. Die Umkehrung ist trivial, denn ist B positiv semidefinit, so ist ϕ konvex. Wegen $g \in \text{Bild}\,(B)$ existiert ferner ein p^* mit $\nabla\phi(p^*) = 0$, dieses p^* ist Lösung von (P).

   Besitzt (P) eine eindeutige Lösung, so ist wegen des ersten Teils B zumindestens positiv semidefinit und $g \in \text{Bild}\,(B)$. Jedes p^* mit $g + Bp^*$ ist eine Lösung, ist diese eindeutig, so ist der Kern von B notwendigerweise trivial und damit B positiv definit. Die Umkehrung ist natürlich trivial.

3. Sei[10] $f(x) := 10(x_2 - x_1^2)^2 + (1 - x_1)^2$. Für $x_c := (0, -1)^T$ mache man einen Contour-Plot des quadratischen Modells

   $$f_c(p) := f(x_c) + \nabla f(x_c)^T p + \frac{1}{2} p^T \nabla^2 f(x_c) p.$$

[10] Diese Aufgabe findet man bei J. Nocedal, S. J. Wright (1999, S. 97).

Man zeichne in den Contour-Plot ferner noch Kreise um $(0,0)$ mit dem Radius 0.5, 1 und 2. Man wiederhole das ganze mit $x_c = (0, 0.5)^T$.

Lösung: Unsere Lösung ist nicht elegant, es geht sicherlich besser. Zunächst schreiben wir ein function file `Rose10.m` so, dass der Aufruf `[f,g,B]=Rose10(x);` den Funktionswert, Gadienten und Hessesche der Zielfunktion in x liefert. Den Plot in Abbildung 5.7 links haben wir durch die folgenden Befehle (in einem script file angelegt) gewonnen:

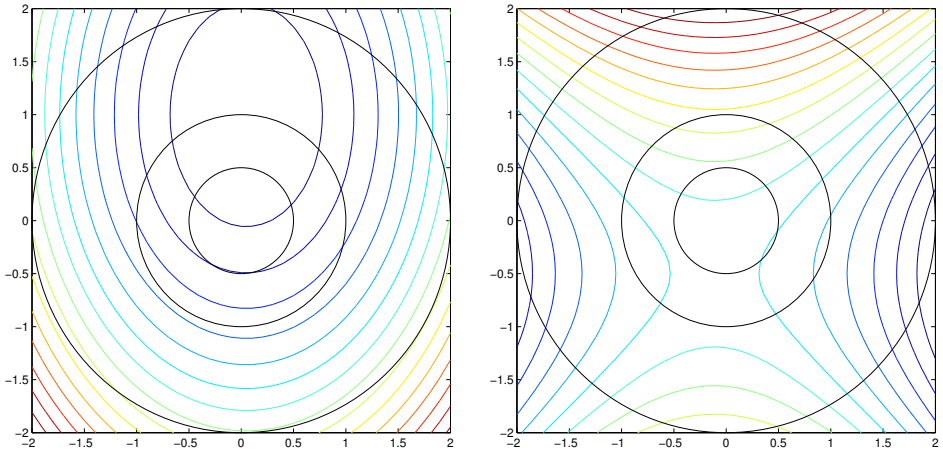

Abbildung 5.7: Höhenlinienplot eines quadratischen Modells, Kreise

```
p_1=-2:0.1:2;p_2=p_1;
[P_1,P_2]=meshgrid(p_1,p_2);
x=[0;-1];
[f,g,B]=Rose10(x);
F_c=f+g(1)*P_1+g(2)*P_2+0.5*(B(1,1)*P_1.^2+2*B(1,2)*P_1.*P_2+B(2,2)*P_2.^2);
contour(P_1,P_2,F_c,15);
hold on;axis square;
rectangle('Position',[-0.5,-0.5,1,1],'Curvature',[1,1]);
rectangle('Position',[-1,-1,2,2],'Curvature',[1,1]);
rectangle('Position',[-2,-2,4,4],'Curvature',[1,1]);
hold off;
```

Für den Plot rechts haben wir nur die Eingabe für `x` entsprechend verändert.

4. Sei p^* eine Lösung von

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2} p^T B p, \quad \|p\|_\infty \leq \Delta.$$

Hierbei sind $g \in \mathbb{R}^n$, die symmetrische Matrix $B \in \mathbb{R}^{n \times n}$ und $\Delta > 0$ gegeben. Man zeige:

(a) Ist $-\Delta < p_j^* < \Delta$, so ist $(g + Bp^*)_j = 0$.

(b) Ist $p_j^* = \Delta$, so ist $(g + Bp^*)_j \leq 0$.

(c) Ist $p_j^* = -\Delta$, so ist $(g + Bp^*)_j \geq 0$.

Lösung: Sei $|p_j^*| < \Delta$. Man definiere $p(t) := p^* + te_j$, wobei e_j der j-te Einheitsvektor ist. Dann ist $\|p(t)\|_\infty \leq \Delta$ für alle hinreichend kleinen $|t|$. Wegen $\phi(p^*) = \phi(p(0)) \leq \phi(p(t))$ für alle hinreichend kleinen $|t|$, ist

$$\begin{aligned} 0 &= \frac{d}{dt}\phi(p(t))\Big|_{t=0} \\ &= \nabla\phi(p(0))^T p'(0) \\ &= (g + Bp^*)^T e_j \\ &= (g + Bp^*)_j. \end{aligned}$$

Damit ist der erste Teil der Aufgabe gelöst. Ist $p_j^* = \Delta$, so setze man entsprechend $p(t) := p^* - te_j$. Dann ist $\|p(t)\|_\infty \leq \Delta$ und $\phi(p(t)) \geq \phi(p(0))$ für alle hinreichend kleine $t \geq 0$. Folglich ist

$$\begin{aligned} 0 &\leq \frac{d}{dt}\phi(p(t))\Big|_{t=0} \\ &= \nabla\phi(p(0))^T p'(0) \\ &= (g + Bp^*)^T(-e_j) \\ &= -(g + Bp^*)_j, \end{aligned}$$

woraus die zweite Aussage folgt. Die dritte beweist man natürlich entsprechend.

5. Sei

$$g := \begin{pmatrix} -2 \\ -20 \end{pmatrix}, \qquad B := \begin{pmatrix} 42 & 0 \\ 0 & 20 \end{pmatrix}.$$

Für $\Delta := \frac{1}{2}$, 1, 2 berechne man eine Lösung von

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2} p^T B p, \quad \|p\|_\infty \leq \Delta.$$

Lösung: Sei zunächst $\Delta = 0.5$. Wir zeigen, dass $p^* = (\frac{1}{21}, \frac{1}{2})^T$ eine Lösung ist. Denn für ein beliebiges $p \in \mathbb{R}^2$ mit $\|p\|_\infty \leq \frac{1}{2}$ ist

$$\begin{aligned} \phi(p) - \phi(p^*) &= \nabla\phi(p^*)^T(p - p^*) + \underbrace{\frac{1}{2}(p - p^*)^T B(p - p^*)}_{\geq 0} \\ &\geq \nabla\phi(p^*)^T(p - p^*) \\ &= \begin{pmatrix} 0 \\ -10 \end{pmatrix}^T \begin{pmatrix} p_1 - \frac{1}{21} \\ p_2 - \frac{1}{2} \end{pmatrix} \\ &= 10(\underbrace{\tfrac{1}{2} - p_2}_{\geq 0}) \\ &\geq 0, \end{aligned}$$

womit die Behauptung bewiesen ist. Für $\Delta = 1$ ist entsprechend $p^* = (\frac{1}{21}, 1)^T$, während man für $\Delta = 2$ dieselbe Lösung $p^* = (\frac{1}{21}, 1)^T$ erhält. Steht die Optimization Toolbox zur Verfügung, so erhält man nach

```
options=optimset('LargeScale','off');
g=[-2;-20];B=[42,0;0,20];Delta=0.5;
l=-Delta*ones(2,1);u=-l;
[p,v]=quadprog(B,g,[],[],[],[],l,u,[],options);
```

und `format long`

$$p = \begin{pmatrix} 0.04761904761905 \\ 0.50000000000000 \end{pmatrix}, \qquad v = -7.54761904761905.$$

Die leeren Klammern `[]` bedeuten hierbei jeweils, dass keine Ungleichungen (Matrix und rechte Seite unbesetzt), keine Gleichungen (Matrix und rechte Seite unbesetzt) in der Optimierungsaufgabe vorkommen und dass kein Startwert vorgegeben wird.

5.3.2 Aufgaben zu Abschnitt 4.2

1. Man berechne die Lösung der Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2} p^T B p, \quad \|p\|_2 \le \Delta,$$

wobei

$$g := \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad B := \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \Delta := \frac{1}{2}.$$

Lösung: Offensichtlich ist B positiv definit. Wir definieren $p\colon [0,\infty) \longrightarrow \mathbb{R}^2$ durch $p(\lambda) := -(B+\lambda I)^{-1} g$. Dann ist

$$p(\lambda) = \frac{1}{(2+\lambda)(1+\lambda) - 1} \begin{pmatrix} -\lambda \\ 1+\lambda \end{pmatrix}.$$

Daher ist

$$\psi(\lambda) := \|p(\lambda)\|_2 = \frac{1}{\lambda^2 + 3\lambda + 1} \sqrt{\lambda^2 + (1+\lambda)^2}.$$

In Abbildung 5.8 skizzieren wir $\psi(\cdot)$ auf dem Intervall $[0,3]$. Man erkennt, dass $\psi(0) = 1 > \Delta$, so dass die Lösung λ^* von $\psi(\lambda) = \Delta$ zu bestimmen ist. Nach der Skizze liegt diese etwa bei 0.75. Genauer ist λ^* als Lösung von

$$4[\lambda^2 + (1+\lambda)^2] = (\lambda^2 + 3\lambda + 1)^2$$

zu bestimmen. Wir erhalten also $\lambda^* \approx 0.747535241461$ und hiermit

$$p^* = p(\lambda^*) \approx \begin{pmatrix} -0.196646592915 \\ 0.459706555855 \end{pmatrix}.$$

2. Sei $B \in \mathbb{R}^{n\times n}$ symmetrisch mit kleinstem Eigenwert λ_1 und $g \in \mathbb{R}^n \setminus \{0\}$. Man zeige, dass die durch

$$\psi(\lambda) := \|(B+\lambda I)^{-1} g\|_2$$

definierte Funktion $\psi\colon (-\lambda_1, \infty) \longrightarrow \mathbb{R}$ auf $(-\lambda_1, \infty)$ monoton fallend und konvex ist.

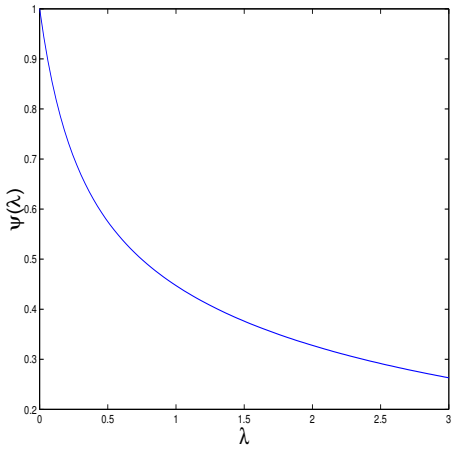

Abbildung 5.8: Die Funktion $\psi(\lambda) := \|p(\lambda)\|_2$

Lösung: Seien $\lambda_1 \leq \cdots \leq \lambda_n$ die Eigenwerte von B und $\{u_1, \ldots, u_n\}$ ein zugehöriges Orthonormalsystem von Eigenvektoren. Mit $\Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ und $U := \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix}$ ist dann

$$\begin{aligned} \psi(\lambda) &= \|(B+\lambda I)^{-1} g\|_2 \\ &= \|U(\Lambda+\lambda I)^{-1} U^T g\|_2 \\ &= \|(\Lambda+\lambda I)^{-1} U^T g\|_2 \\ &= \left(\sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^2} \right)^{1/2} \end{aligned}$$

für $\lambda \in (-\lambda_1, \infty)$. Als Ableitung berechnet man

$$\psi'(\lambda) = -\left(\sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^2} \right)^{-1/2} \left(\sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^3} \right) = -\frac{1}{\psi(\lambda)} \sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^3}.$$

Wegen $g \neq 0$ ist $u_i^T g \neq 0$ für wenigstens ein i und folglich $\psi'(\lambda) < 0$ für alle $\lambda \in (-\lambda_1, \infty)$, also $\psi(\cdot)$ auf $(-\lambda_1, \infty)$ monoton fallend. Erneutes Differenzieren liefert

$$\begin{aligned} \psi''(\lambda) &= \frac{3}{\psi(\lambda)} \sum_{i=1}^n \frac{(u_t^T g)^2}{(\lambda_i+\lambda)^4} + \frac{\psi'(\lambda)}{\psi(\lambda)^2} \sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^3} \\ &= \frac{3}{\psi(\lambda)} \sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^4} - \frac{1}{\psi(\lambda)^3} \left(\sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^3} \right)^2 \\ &= \frac{2}{\psi(\lambda)} \sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^4} \\ &\qquad + \frac{1}{\psi(\lambda)^3} \left[\left(\sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^2} \right) \left(\sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^4} \right) - \left(\sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^3} \right)^2 \right] \\ &\geq \frac{2}{\psi(\lambda)} \sum_{i=1}^n \frac{(u_i^T g)^2}{(\lambda_i+\lambda)^4} \\ &> 0, \end{aligned}$$

insbesondere ist $\psi(\cdot)$ auf $(-\lambda_1, \infty)$ (strikt) konvex.

3. Sei $B \in \mathbb{R}^{n\times n}$ symmetrisch mit kleinstem Eigenwert λ_1, sei weiter $\lambda \in \mathbb{R}$.

   (a) Für jedes $u \in \mathbb{R}^n$ mit $\|u\|_2 = 1$ ist

   $$\lambda_+ := \lambda - u^T(B + \lambda I)u \leq -\lambda_1.$$

   (b) Sei $\lambda > -\lambda_1$ und $B + \lambda I = LL^T$ eine Cholesky-Zerlegung, also $L \in \mathbb{R}^{n\times n}$ eine untere Dreiecksmatrix mit positiven Diagonalelementen. Um eine möglichst gute untere Schranke für $-\lambda_1$ zu erhalten, bestimmt man einen Vektor $u \in \mathbb{R}^n$ mit $\|u\|_2 = 1$, für den $u^T(B + \lambda I)u = \|L^T u\|_2^2$ "klein" ist. Eine mögliche Heuristik zur Bestimmung von u ist die folgende (siehe A. R. Conn, N. I. M. Gould, Ph. L. Toint (2000, S. 191)): Mit einem Vektor $v \in \mathbb{R}^n$ mit $v_i \in \{-1, 1\}$, $i = 1, \ldots, n$, setze

   $$u := \frac{(B + \lambda I)^{-1}v}{\|(B + \lambda I)^{-1}v\|_2}.$$

   Dann ist

   $$u^T(B + \lambda I)u = \frac{v^T(B + \lambda I)^{-1}v}{\|(B + \lambda I)^{-1}v\|_2^2}.$$

   Daher sollte man v so wählen, dass

   $$\|(B + \lambda I)^{-1}v\|_2 = \|L^{-T}L^{-1}v\|_2$$

   bzw. $\|L^{-1}v\|$ möglichst groß ist. Hierzu bestimme man $w = Lv$ folgendermaßen. Angenommen, $w_1, \ldots, w_{k-1}$ sind schon bestimmt. Man setze

   $$v_k := \begin{cases} -1, & \text{falls} \quad \sum_{i=1}^{k-1} l_{ki}w_i > 0, \\ +1, & \text{falls} \quad \sum_{i=1}^{k-1} l_{ki}w_i \leq 0 \end{cases}$$

   und anschließend

   $$w_k := \frac{1}{l_{kk}}\left(v_k - \sum_{i=1}^{k-1} l_{ki}w_i\right).$$

   Weiter ist

   $$u = \frac{(B + \lambda I)^{-1}v}{\|(B + \lambda I)^{-1}v\|_2} = \frac{L^{-T}w}{\|L^{-T}w\|_2}.$$

   Man schreibe eine Matlab-function `Estimate.m`, welche den Overhead hat:

```
function [lambda_plus,info]=Estimate(B,lambda);
%Input-parameter:
%    B           symmetric matrix
%    lambda      real number, usually lambda>-lambda_1
%Output-parameter:
%    lambda_plus lower bound of -lambda_1, if successfull
%    info        info=0, if successfull
%                info=1, if the Cholesky decomposition of
%                B+lambda*I does not exist (because lambda is
%                not an upper bound of -lambda_1)
%************************************************************
```

Anschließend erprobe man die Funktion an den Daten

$$B := \begin{pmatrix} 4 & -3 & 5 & -1 \\ -3 & 2 & -6 & 4 \\ 5 & -6 & 1 & 5 \\ -1 & 4 & 5 & -8 \end{pmatrix}, \qquad \lambda = 9,\ 8,\ 7.7.$$

Lösung: Der erste Teil der Aufgabe ist völlig trivial. Denn sei $\lambda \in \mathbb{R}$ und $u \in \mathbb{R}^n$ ein Vektor mit $\|u\|_2 = 1$. Dann ist

$$\lambda_+ := \lambda - u^T(B + \lambda I)u = -u^T B u \leq -\lambda_1$$

und das ist schon der Nachweis der Behauptung.

Der Rumpf der Matlab-Funktion `Estimate.m` könnte folgendermaßen aussehen:

```
n=length(B);[R,p]=chol(B+lambda*eye(n));
if p>0
   info=1; return
else
   L=R';v=zeros(n,1);w=zeros(n,1);
   v(1)=1;w(1)=1/L(1,1);
   for k=2:n
       temp=L(k,1:k-1)*w(1:k-1);
       if temp>0
            v(k)=-1;
       else
            v(k)=1;
       end;
       w(k)=(v(k)-temp)/L(k,k);
   end;
   v=L'\w;u=v/norm(v);
   info=0;lambda_plus=lambda-norm(L'*u)^2;
end;
```

Für die angegebenen Daten erhalten wir die folgenden Ergebnisse (man ist jeweils erfolgreich):

λ	λ_+
9	7.4780
8	7.6003
7.7	7.6128

4. Sei $B \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, $g \in \mathbb{R}^n$ und $\Delta > 0$. Man schreibe eine Matlab-Funktion `TrustStep` zur Berechnung der Lösung des Trust-Region-Hilfsproblems

(P) $$\text{Minimiere} \quad \phi(p) := g^T p + \frac{1}{2} p^T B p, \quad \|p\|_2 \leq \Delta.$$

Anschließend teste man die Funktion für den Spezialfall, dass

$$B := \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{n\times n}, \qquad g := \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} \in \mathbb{R}^n$$

mit $n := 10$ und $\Delta := 1$.

Hinweis: Wegen der Voraussetzung, dass B positiv definit ist, kann der schwere Fall nicht eintreten, so dass die Aufgabenstellung einfach ist.

Lösung: Wir schreiben die folgende Funktion:

```
function [p,lambda,iter]=TrustStep(B,g,Delta,max_iter,tol);
%******************************************************
%Solve the trust region subproblem
%Minimize g^Tp+0.5*p^TBp,   ||p||_2<=Delta
%******************************************************
%Input parameter:
%  B        symmetric and positive definite n-by-n matrix
%  g        n-vector
%  Delta    positive trust region radius
%  max_iter maximal number of iterations
%  tol      tolerance: STOP if | ||p_c||_2-Delta|<=tol
%Output-parameter:
%  p        approximate solution
%  lambda   associated multiplier
%  iter     number of iterations
%******************************************************
p_c=-B\g;
if norm(p_c)<Delta
  p=p_c;lambda=0;iter=0;return
end;
lambda_c=0;
n=length(g);
for k=1:max_iter
    L_c=chol(B+lambda_c*eye(n))';
    p_c=-L_c'\(L_c\g);normpc=norm(p_c);
    if abs(normpc-Delta)<tol
        p=p_c;lambda=lambda_c;iter=k;return
    end;
    w_c=L_c\p_c;
    lambda_c=lambda_c+((normpc-Delta)/Delta)*(normpc/norm(w_c))^2;
end;
iter=max_iter;
```

Zunächst setzen wir

```
>> B=2*eye(10)+diag(-ones(9,1),1)+diag(-ones(9,1),-1);
>> g=ones(10,1);
```

Mit dem Aufruf

```
[p,lambda,iter]=TrustStep(B,g,1.0,100,1e-12);
```

erhalten wir (mit `format long`)

$$p = \begin{pmatrix} -0.26330804211067 \\ -0.31818848044440 \\ -0.32962617353183 \\ -0.33200579892365 \\ -0.33248112890276 \\ -0.33248112890276 \\ -0.33200579892365 \\ -0.32962617353183 \\ -0.31818848044440 \\ -0.26330804211067 \end{pmatrix}, \qquad \lambda = 3.00625985396363, \qquad \text{iter} = 4.$$

5. Sei $f \in \mathbb{R}$, $g \in \mathbb{R}^n$, $B \in \mathbb{R}^{n\times n}$ symmetrisch, $D \in \mathbb{R}^n$ nichtsingulär und $\Delta > 0$. Man gebe notwendige und hinreichende Bedingungen dafür an, dass ein $p^* \in \mathbb{R}^n$ mit $\|Dp^*\|_2 \le \Delta$ eine globale Lösung von

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T B p, \quad \|Dp\|_2 \le \Delta$$

ist.

Lösung: Ist p^* eine Lösung von (P), so ist $q^* := Dp^*$ eine Lösung von

$$\text{Minimiere} \quad \psi(q) := f + (D^{-T} g)^T q + \frac{1}{2} q^T D^{-T} B D^{-1} q, \quad \|q\|_2 \le \Delta.$$

Wegen Satz 2.1 ist dies genau dann der Fall, wenn ein $\lambda^* \ge 0$ existiert derart, dass

$$(D^{-T} B D^{-1} + \lambda^* I) q^* = -D^{-T} g, \qquad \lambda^* (\Delta - \|q^*\|_2) = 0$$

und $D^{-T} B D^{-1} + \lambda^* I$ positiv semidefinit ist. Daher ist p^* genau dann eine Lösung von (P), wenn ein $\lambda^* \ge 0$ existiert derart, dass

$$(B + \lambda^* D^T D) p^* = -g, \qquad \lambda^* (\Delta - \|Dp^*\|_2) = 0$$

und $B + \lambda^* D^T D$ positiv semidefinit ist.

6. Gegeben sei das Trust-Region Hilfsproblem

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := f + g^T p + \frac{1}{2} p^T B p, \quad \|p\|_2 \le \Delta,$$

wobei $f \in \mathbb{R}$, $g \in \mathbb{R}^{n\times n} \setminus \{0\}$, die symmetrische Matrix $B \in \mathbb{R}^{n\times n}$ und $\Delta > 0$ gegeben sind. Sei der sogenannte Cauchy-Punkt definiert durch

$$p^C := -\tau \frac{\Delta}{\|g\|_2} g,$$

wobei

$$\tau := \begin{cases} 1, & \text{falls} \quad g^T B g \le 0, \\ \min(\|g\|_2^3/(\Delta g^T B g), 1), & \text{sonst.} \end{cases}$$

Man zeige, dass

$$f - \phi(p^C) \ge \frac{1}{2}\|g\|_2 \min\left(\Delta, \frac{\|g\|_2}{\|B\|_2}\right).$$

Der Cauchy-Punkt p^C genügt also derselben Abschätzung wie eine globale Lösung p^* von (P), siehe Lemma 2.5. Weiter zeige man, dass p^C Lösung der Aufgabe

$$\text{Minimiere} \quad \phi(p), \quad \|p\|_2 \le \Delta, \; p \in \text{span}\,\{g\}$$

ist.

Lösung: Wir nehmen zunächst an, es sei $g^T B g \le 0$. Dann ist

$$\begin{aligned} f - \phi(p^C) &= -g^T p^C - \frac{1}{2}(p^C)^T B p^C \\ &= \Delta\|g\|_2 - \frac{1}{2}\frac{\Delta^2}{\|g\|_2^2}\underbrace{g^T B g}_{\le 0} \\ &\ge \Delta\|g\|_2 \\ &\ge \|g\|_2 \min\left(\Delta, \frac{\|g\|_2}{\|B\|_2}\right) \\ &\ge \frac{1}{2}\|g\|_2 \min\left(\Delta, \frac{\|g\|_2}{\|B\|_2}\right). \end{aligned}$$

In diesem Fall ist die Behauptung also richtig. Nun nehmen wir $g^T B g > 0$ an. Auch für diesen Fall machen wir eine Fallunterscheidung und setzen zunächst $\|g\|_2^3/(\Delta g^T B g) \le 1$ voraus. Dann ist

$$\begin{aligned} f - \phi(p^C) &= -g^T p^C - \frac{1}{2}(p^C)^T B p^C \\ &= \frac{1}{2}\frac{\|g\|_2^4}{g^T B g} \\ &\ge \frac{1}{2}\frac{\|g\|_2^4}{\|B\|_2\,\|g\|_2^2} \\ &= \frac{1}{2}\frac{\|g\|_2^2}{\|B\|_2} \\ &\ge \frac{1}{2}\|g\|_2 \min\left(\Delta, \frac{\|g\|_2}{\|B\|_2}\right). \end{aligned}$$

Ist dagegen $\|g\|_2^3/(\Delta g^T B g) > 1$, so ist

$$\begin{aligned} f - \phi(p^C) &= -g^T p^C - \frac{1}{2}(p^C)^T B p^C \\ &= \Delta\|g\|_2 - \frac{1}{2}\frac{\Delta^2}{\|g\|_2^2} g^T B g \\ &\ge \Delta\|g\|_2 - \frac{1}{2}\Delta\|g\|_2 \end{aligned}$$

$$\begin{aligned} &= \frac{1}{2}\Delta\|g\|_2 \\ &\geq \frac{1}{2}\|g\|_2 \min\left(\Delta, \frac{\|g\|_2}{\|B\|_2}\right). \end{aligned}$$

Nun wollen wir noch zeigen, dass p^C Lösung der Aufgabe

$$\text{Minimiere} \quad \phi(p), \quad \|p\|_2 \leq \Delta,\ p \in \operatorname{span}\{g\}$$

ist, was man auch als eine Motivation für die Definition des Cauchy-Punktes ansehen kann. Hierzu stelle wir $p \in \operatorname{span}\{g\}$ in der Form

$$p = -\tau \frac{\Delta}{\|g\|} g$$

mit $\tau \in \mathbb{R}$ dar. Hierdurch wird obiges eindimensionales Hilfsproblem auf

$$\text{Minimiere} \quad \psi(\tau) := f - \tau\Delta\|g\|_2 + \frac{\tau^2}{2}\frac{\Delta^2}{\|g\|_2^2} g^T B g, \quad |\tau| \leq 1$$

transformiert. Ist $g^T B g \leq 0$, so ist ψ eine konkave Funktion, die ihr Minimum an einem der Intervallenden, also für $\tau = 1$ annimmt. Sei daher jetzt $g^T B g > 0$ und daher ψ konvex. Ist $\tau = \|g\|_2^3/(\Delta g^T B g) \leq 1$, so ist τ eine Lösung, andernfalls ist es $\tau = 1$. Insgesamt ist die Behauptung bewiesen.

7. Seien $f \in \mathbb{R}$, $g \in \mathbb{R}^n \setminus \{0\}$, die symmetrische und positiv definite Matrix $B \in \mathbb{R}^{n\times n}$ und $\Delta > 0$ gegeben. Ferner sei $\phi\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ durch

$$\phi(p) := f + g^T p + \frac{1}{2} p^T B p$$

definiert. Sei

$$p^B := -B^{-1} g$$

(unrestringiertes Minimum von ϕ) und

$$p^U := -\frac{\|g\|_2^2}{g^T B g} g$$

(unstringiertes Minimum von ϕ in $\operatorname{span}\{g\}$). Wir nehmen an, es sei $p^B \neq p^U$. Man definiere $\tilde{p}\colon [0,2] \longrightarrow \mathbb{R}^n$ durch

$$\tilde{p}(\tau) := \begin{cases} \tau p^U, & 0 \leq \tau \leq 1, \\ p^U + (\tau - 1)(p^B - p^U), & 1 \leq \tau \leq 2. \end{cases}$$

Man zeige:

(a) $\|\tilde{p}(\cdot)\|_2$ ist strikt monoton wachsend,

(b) $\phi(\tilde{p}(\cdot))$ ist strikt monoton fallend.

(c) Sei $\|p^B\|_2 > \Delta$ (andernfalls ist p^B die Lösung des Trust-Region-Hilfsproblems). Man zeige, dass es genau ein $\tau^* \in (0,2)$ mit $\|\tilde{p}(\tau^*)\|_2 = \Delta$ gibt[11].

[11] Die Bestimmung von $\tilde{p}(\tau^*)$ als Näherungslösung des Trust-Region-Hilfsproblems wird in der englischsprachigen Literatur als "dog leg method" bezeichnet.

(d) Sei $\|p^B\|_2 > \Delta$. Man zeige, dass $\phi(\tilde{p}(\tau^*)) \leq \phi(p^C)$, wobei p^C der in Aufgabe 6 definierte Cauchy-Punkt ist. Insbesondere gilt auch für $\tilde{p}(\tau^*)$ die Abschätzung aus Lemma 2.5, also

$$f - \phi(\tilde{p}(\tau^*)) \geq \frac{1}{2}\|g\|_2 \min\left(\Delta, \frac{\|g\|_2}{\|B\|_2}\right).$$

Lösung:Für $\tau \in [0,1]$ ist

$$\|\tilde{p}(\tau)\| = \tau\|p^U\|_2, \qquad \phi(\tilde{p}(\tau)) = f - \tau\frac{\|g\|_2^4}{g^TBg} + \frac{\tau^2}{2}\frac{\|g\|_2^4}{g^TBg}.$$

Hieraus erkennt man, dass $\|\tilde{p}(\cdot)\|_2$ auf $[0,1]$ strikt monoton wachsend und $\phi(\tilde{p}(\cdot))$ auf $[0,1]$ strikt monoton fallend ist. Für die ersten beiden Teile der Aufgabe braucht man also nur das Intervall $[1,2]$ zu betrachten. Für (a) definieren wir

$$\begin{aligned} h(\alpha) &:= \frac{1}{2}\|\tilde{p}(1+\alpha)\|_2^2 \\ &= \frac{1}{2}\|p^U + \alpha(p^B - p^U)\|_2^2 \\ &= \frac{1}{2}\|p^U\|_2^2 + \alpha(p^U)^T(p^B - p^U) + \frac{1}{2}\alpha^2\|p^B - p^U\|_2^2. \end{aligned}$$

Zu zeigen ist, dass $h'(\alpha) \geq 0$ für alle $\alpha \in (0,1]$. Nun ist

$$\begin{aligned} h'(\alpha) &= -(p^U)^T(p^U - p^B) + \underbrace{\alpha\|p^B - p^U\|_2^2}_{>0} \\ &> -(p^U)^T(p^U - p^B) \\ &= \frac{\|g\|_2^2}{g^TBg}g^T\left(-\frac{\|g\|_2^2}{g^TBg}g + B^{-1}g\right) \\ &= \|g\|_2^2\frac{g^TB^{-1}g}{g^TBg}\underbrace{\left[1 - \frac{\|g\|_2^4}{(g^TBg)(g^TB^{-1}g)}\right]}_{\geq 0} \\ &\geq 0. \end{aligned}$$

Hierbei haben wir am Schluss ausgenutzt, dass wegen der Cauchy-Schwarzschen Ungleichung

$$\frac{\|g\|_2^4}{(g^TBg)(g^TB^{-1}g)} = \frac{((B^{1/2}g)^T(B^{-1/2}g))^2}{\|B^{1/2}g\|_2^2\,\|B^{-1/2}g\|_2^2} \leq 1.$$

Für (b) definieren wir $k(\alpha) := \phi(\tilde{p}(\alpha))$ und zeigen, dass $k'(\alpha) < 0$ für alle $\alpha \in (0,1)$. Nun ist

$$\begin{aligned} k'(\alpha) &= \nabla\phi(\tilde{p}(\alpha))^T(p^B - p^U) \\ &= (g + B\tilde{p}(\alpha))^T(p^B - p^U) \\ &= (g + Bp^U)^T(p^B - p^U) + \alpha\underbrace{(p^B - p^U)^TB(p^B - p^U)}_{>0} \\ &< (g + Bp^U)^T(p^B - p^U) + (p^B - p^U)^TB(p^B - p^U) \\ &= (p^B - p^U)^T(g + Bp^B) \\ &= 0 \end{aligned}$$

für alle $\alpha \in [0,1)$. Damit sind (a) und (b) bewiesen.

Zum Beweis von (c) nehmen wir $\|p^B\|_2 > \Delta$ an. Es ist $\|\tilde{p}(0)\|_2 = 0 < \Delta$ und $\|\tilde{p}(2)\|_2 = \|p^B\|_2 > \Delta$. Da $\|\tilde{p}(\cdot)\|_2$ auf $[0,2]$ strikt monoton wachsend ist, gibt es genau ein $\tau^* \in (0,2)$ mit $\|\tilde{p}(\tau^*)\|_2 = \Delta$. Ist $\|p^U\| \geq \Delta$ bzw. $\|g\|_2^3/(\Delta g^T B g) \geq 1$, so ist $\tau^* \in (0,1]$ gegeben durch

$$\tau^* = \frac{\Delta}{\|p^U\|_2} = \frac{\Delta g^T B g}{\|g\|_2^3}.$$

Andernfalls ist $\tau^* \in (1,2)$ als Lösung von

$$\|p^U + (\tau - 1)(p^B - p^U)\|_2^2 = \Delta^2$$

zu bestimmen.

Wieder sei $\|p^B\|_2 > \Delta$. Ist $\|p^U\|_2 \geq \Delta$ bzw. $\|g\|_2^3/(\Delta g^T B g) \geq 1$, so ist

$$\tau^* = \frac{\Delta g^T B g}{\|g\|_2^3} \in (0,1]$$

und daher

$$\tilde{p}(\tau^*) = -\frac{\Delta}{\|g\|_2} g = p^C.$$

In diesem Falle ist also $\phi(\tilde{p}(\tau^*)) = \phi(p^C)$, die Behauptung also richtig. Ist dagegen $\|p^U\|_2 < \Delta$ bzw. $\|g\|_2^3/(\Delta g^T B g) < 1$, so ist $\tau^* \in (1,2)$ und

$$p^C = -\frac{\|g\|_2^3}{g^T B g} g = p^U = \tilde{p}(1).$$

Da $\phi(\tilde{p}(\cdot))$ auf $[0,2]$ monoton fallend ist, ist $\phi(p^C) \geq \phi(\tilde{p}(1)) \geq \phi(\tilde{p}(\tau^*))$. Damit ist die Aufgabe scgließlich vollständig gelöst.

8. Seien $f \in \mathbb{R}$, $g \in \mathbb{R}^n \setminus \{0\}$, die symmetrische und positiv definite Matrix $B \in \mathbb{R}^{n \times n}$ und $\Delta > 0$ gegeben. Man gebe ein Verfahren zur Berechnung der Lösung p^M von

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad \phi(p) := f + g^T p + \dfrac{1}{2} p^T B p \quad \text{auf} \\ M := \{p \in \mathbb{R}^n : p \in \operatorname{span}\{g, B^{-1}g\},\ \|p\|_2 \leq \Delta\} \end{cases}$$

an und zeige, dass

$$f - \phi(p^M) \geq \frac{1}{2}\|g\|_2 \min\left(\Delta, \frac{\|g\|_2}{\|B\|_2}\right).$$

Hinweis: Für den letzten Teil der Aufgabe kann man die Aussage von Aufgabe 6 verwenden.

Lösung: O. B. d. A. können wir annehmen, dass g und $B^{-1}g$ linear unabhängig sind. Denn andernfalls ist der Cauchy-Punkt p^C Lösung von (P), wie wir in Aufgabe 6 gesehen haben. Das Problem (P) ist äquivalent zu

$$\begin{cases} \text{Minimiere} \quad f + \begin{pmatrix} \|g\|_2^2 \\ g^T B g \end{pmatrix}^T \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \dfrac{1}{2} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}^T \begin{pmatrix} g^T B g & \|g\|_2^2 \\ \|g\|_2^2 & g^T B^{-1} g \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \\ \text{unter der Nebenbedingung} \quad \|\alpha g + \beta B^{-1} g\|_2 \leq \Delta. \end{cases}$$

Im Prinzip hat man das Problem (P) also auf ein zweidimensionales Trust-Region-Hilfsproblem zurückgeführt. Eine Lösung (α^*, β^*) ist durch die Existenz eines $\lambda^* \geq 0$ mit

$$\left[\begin{pmatrix} g^TBg & \|g\|_2^2 \\ \|g\|_2^2 & g^TB^{-1}g \end{pmatrix} + \lambda^* \begin{pmatrix} \|g\|_2^2 & g^TB^{-1}g \\ g^TB^{-1}g & \|B^{-1}g\|_2^2 \end{pmatrix}\right] \begin{pmatrix} \alpha^* \\ \beta^* \end{pmatrix} = -\begin{pmatrix} \|g\|_2^2 \\ g^TBg \end{pmatrix}$$

und

$$\lambda^*(\Delta - \|\alpha^* g + \beta^* B^{-1}g\|_2) = 0$$

charakterisiert. Für $\lambda \geq 0$ definiere man

$$\begin{pmatrix} \alpha(\lambda) \\ \beta(\lambda) \end{pmatrix} := -\left[\begin{pmatrix} g^TBg & \|g\|_2^2 \\ \|g\|_2^2 & g^TB^{-1}g \end{pmatrix} + \lambda \begin{pmatrix} \|g\|_2^2 & g^TB^{-1}g \\ g^TB^{-1}g & \|B^{-1}g\|_2^2 \end{pmatrix}\right]^{-1} \begin{pmatrix} \|g\|_2^2 \\ g^TBg \end{pmatrix}.$$

Natürlich könnte man eine explizite Darstellung angeben, was aber nicht sonderlich hilfreich wäre. Ist $\|\alpha(0)g + \beta(0)B^{-1}g\|_2 \leq \Delta$, so ist $p^M := \alpha(0)g + \beta(0)B^{-1}g$ die Lösung von (P). Andernfalls hat man die Lösung $\lambda^* > 0$ von

$$\|\alpha(\lambda)g + \beta(\lambda)B^{-1}g\|_2 = \Delta$$

zu bestimmen und es ist $p^M := \alpha(\lambda^*)g + \beta(\lambda^*)B^{-1}g$ die Lösung von (P). Da der Cauchy-Punkt p^C als Vielfaches von g trivialerweise in $\operatorname{span}\{g, B^{-1}g\}$ liegt, folgt aus Aufgabe 6, dass

$$f - \phi(p^M) \geq f - \phi(p^C) \geq \frac{1}{2}\|g\|_2 \min\left(\Delta, \frac{\|g\|_2}{\|B\|_2}\right).$$

Die Aufgabe ist damit gelöst.

9. In Aufgabe 1 sollte die exakte Lösung p^* von

$$\text{(P)} \qquad \text{Minimiere} \quad \phi(p) := g^Tp + \frac{1}{2}p^TBp, \quad \|p\|_2 \leq \Delta,$$

berechnet werden, wobei

$$g := \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad B := \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \Delta := \frac{1}{2}.$$

Zum Vergleich berechne man den Cauchy-Punkt p^C (siehe Aufgabe 6) und den Dog Leg Punkt $\tilde{p}(\tau^*)$ (siehe Aufgabe 7). Man vergleiche die Werte $\phi(p^*)$, $\phi(p^C)$ und $\phi(\tilde{p}(\tau^*))$. Schließlich plotte man den "Dog Leg Pfad" $\{\tilde{p}(\tau) : \tau \in [0, 2]\}$ und (gestrichelt) den optimalen Pfad $\{p^*(\Delta) : \Delta \in [0, 1]\}$, wobei $p^*(\Delta)$ die Lösung von (P) (mit variablem Δ) ist.

Lösung: Es ist $\|g\|_2^3/(\Delta g^TBg) > 1$ und daher

$$p^C = \tilde{p}(\tau^*) = -\frac{\Delta}{\|g\|_2}g = \frac{1}{2\sqrt{2}}\begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Wir berechnen

$$\phi(p^*) = -0.42161847956689, \qquad \phi(\tilde{p}(\tau^*)) = \phi(p^C) = -0.39460678118655.$$

In Abbildung 5.9 geben wir den Dog Leg Pfad und (gestrichelt) den optimalen Pfad an. Diese Abbildung haben wir mit Hilfe der Funktion `TrustStep` aus Aufgabe 4 durch das folgende (sicher verbesserungswürdige) Script file hergestellt:

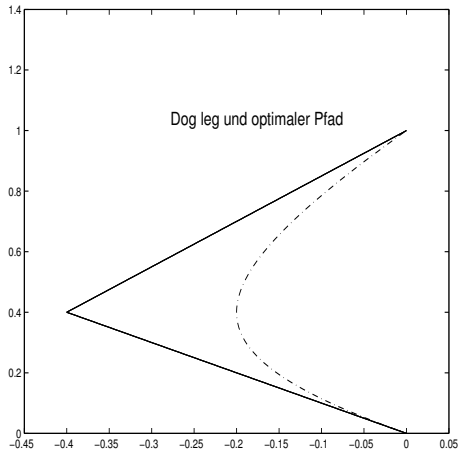

Abbildung 5.9: Dog Leg und optimaler Pfad

```
g=[1;-1];B=[2 -1;-1 1];
p_B=-B\g;
p_U=-(norm(g)^2/(g'*B*g))*g;
tau=linspace(0,1);
p_1=p_U(1)*tau;
p_2=p_U(2)*tau;
tau=linspace(1,2);
q_1=p_U(1)+(tau-1)*(p_B(1)-p_U(1));
q_2=p_U(2)+(tau-1)*(p_B(2)-p_U(2));
r_1=[p_1 q_1];
r_2=[p_2 q_2];
plot(r_1,r_2);
hold on
h=0.01;p=[];
for k=1:100
  Delta=k*h;
  q=TrustStep(B,g,Delta,100,1e-10);
  p=[p q];
end;
plot(p(1,:),p(2,:),'-.');
```

5.3.3 Aufgaben zu Abschnitt 4.3

1. Sei $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ konvex und $M \subset \mathbb{R}^n$ konvex. Ist dann x^* eine Lösung der konvexen Optimierungsaufgabe

   $$\text{(P)} \qquad \text{Minimiere} \quad f(x), \quad x \in M,$$

   welche im *Inneren* von M liegt (insbesondere sei dieses also nichtleer), so ist $f(x^*) \leq f(x)$ für alle $x \in \mathbb{R}^n$, d. h. in x^* liegt ein unrestringiertes Minimum von f.

   Lösung: Sei $x \notin M$ gegeben. Zu zeigen bleibt, dass $f(x^*) \leq f(x)$. Es existiert ein $\lambda_0 \in (0, 1]$ mit $x^* + \lambda_0(x - x^*) \in M$. Da x^* eine Lösung von (P) ist, ist $f(x^*) \leq f(x^* + \lambda_0(x - x^*))$. Die Konvexität von f auf dem $\mathbb{R}^n$ sichert, dass

   $$f(x^*) \leq f(x^* + \lambda_0(x - x^*)) \leq (1 - \lambda_0)f(x^*) + \lambda_0 f(x),$$

woraus $f(x^*) \leq f(x)$ und damit die Behauptung folgt.

2. Gegeben sei die nichtlineare Approximationsaufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x) := \|F(x)\|, \quad x \in \mathbb{R}^n.$$

Hierbei sei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ in $x \in \mathbb{R}^n$ stetig partiell differenzierbar. Mit gegebenem $\Delta > 0$, einer symmetrischen und positiv semidefiniten Matrix $B \in \mathbb{R}^{n\times n}$ und einer Norm $\|\cdot\|$ auf dem $\mathbb{R}^n$ sei p^* eine (globale) Lösung von

$$(\mathrm{P}_{x,\Delta}) \qquad \text{Minimiere} \quad f_x(p) := \|F(x) + F'(x)p\| + \frac{1}{2}p^T Bp, \quad \|p\| \leq \Delta.$$

Dann gilt die folgende Verallgemeinerung von Lemma 3.1 bzw. des ersten Teils von Lemma 3.2:

(a) Es ist x genau dann eine stationäre Lösung von (P), wenn $f(x) = f_x(p^*)$.

(b) Bezeichnet man den Optimalwert von $(\mathrm{P}_{x,\Delta})$ mit $v(x,\Delta)$ und ist $\Delta^* > 0$, so ist

$$f(x) - v(x,\Delta) \geq \frac{\Delta}{\Delta^*}[f(x) - v(x,\Delta^*)] \qquad \text{für alle } \Delta \in (0,\Delta^*].$$

Lösung: Die Abbildung $g\colon \mathbb{R}^m \longrightarrow \mathbb{R}$ sei durch $g(y) := \|y\|$ definiert. Sei x eine stationäre Lösung von (P). Dann ist insbesondere

$$\begin{aligned} 0 &\leq f'(x;p^*) \\ &= g'(F(x);F'(x)p^*) \\ &\leq \|F(x) + F'(x)p^*\| - \|F(x)\| \\ &= f_x(p^*) - \frac{1}{2}(p^*)^T Bp^* - f(x) \\ &\leq f_x(p^*) - f(x) \\ &\leq 0, \end{aligned}$$

woraus die Behauptung folgt, da ja natürlich trivialerweise $f_x(p^*) \leq f_x(0) = f(x)$. Sei umgekehrt $f_x(p^*) = f(x)$. Dann ist auch $p^{**} := 0$ eine Lösung von $(\mathrm{P}_{x,\Delta})$, insbesondere von $(\mathrm{P}_{x,\Delta})$, insbesondere auch der unrestringierten Aufgabe, $f_x(\cdot)$ auf dem $\mathbb{R}^n$ zu minimieren. Für ein beliebiges $p \in \mathbb{R}^n$ ist daher

$$\begin{aligned} 0 &\leq f_x'(0;p) \\ &= \lim_{t\to 0+} \frac{f_x(tp) - f_x(0)}{t} \\ &= \lim_{t\to 0+} \frac{\|F(x) + tF'(x)p\| + \frac{1}{2}t^2 p^T Bp - \|F(x)\|}{t} \\ &= g'(F(x);F'(x)p) \\ &= f'(x;p), \end{aligned}$$

also ist x eine stationäre Lösung von (P). Damit ist der erste Teil der Aufgabe gelöst.

Sei $\Delta \in (0, \Delta^*]$ und p^* eine Lösung von $(\mathrm{P}_{x,\Delta^*})$. Dann ist $p := \lambda p^*$ mit $\lambda := \Delta/\Delta^* \in (0, 1]$ zulässig für $(\mathrm{P}_{x,\Delta})$ und daher

$$\begin{aligned} f(x) - v(x,\Delta) &\geq f_x(0) - f_x(\lambda p^*) \\ &= f_x(0) - f_x((1-\lambda)0 + \lambda p^*) \\ &\geq f_x(0) - [(1-\lambda) f_x(0) + \lambda f_x(p^*)] \\ &= \lambda[f_x(0) - f_x(p^*)] \\ &= \frac{\Delta}{\Delta^*}[f(x) - v(x,\Delta^*)], \end{aligned}$$

womit die Behauptung bewiesen ist.

3. Man schreibe eine Matlab-Funktion, mit welcher mit Hilfe des Madsen-Verfahrens das nichtlineare Tschebyscheffproblem

$$\text{(P)} \qquad \text{Minimiere} \quad \|F(x)\|_\infty, \quad x \in \mathbb{R}^n,$$

gelöst werden kann, wobei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ stetig differenzierbar ist.

Hinweis: Das auftretende Hilfsproblem

$$\text{Minimiere} \quad \|F(x) + F'(x)p\|_\infty, \quad \|p\|_\infty \leq \Delta$$

ist äquivalent zur linearen Optimierungsaufgabe

$$\begin{gathered} \text{Minimiere} \quad \delta \quad \text{unter den Nebenbedingungen} \\ -\delta e \leq F(x) + F'(x)p \leq \delta e, \qquad -\Delta e \leq p \leq \Delta e. \end{gathered}$$

Hierbei ist e der Vektor des $\mathbb{R}^m$ bzw. $\mathbb{R}^n$, dessen Komponenten alle gleich 1 sind. Steht die Optimization toolbox von Matlab zur Verfügung, so kann man die Funktion `linprog` benutzen, andernfalls ist man leider gezwungen, sich selbst einen Löser für linease Programme zu schreiben. Anschließend teste man die Funktion an den folgenden Beispielen:

(a) Man setze $t_i := (i-11)/10$, $i = 1, \ldots, 21$. Die Abbildung $F\colon \mathbb{R}^5 \longrightarrow \mathbb{R}^{21}$ sei durch

$$F_i(x_1, x_2, x_3, x_4, x_5) := \frac{x_1 + x_2 t_i}{1 + x_3 t_i + x_4 t_i^2 + x_5 t_i^3} - \exp(t_i), \qquad i = 1, \ldots, 21,$$

gegeben.

(b) Sei $F\colon \mathbb{R}^2 \longrightarrow \mathbb{R}^3$ durch

$$F(x_1, x_2) := \begin{pmatrix} x_1^2 + x_2^2 + x_1 x_2 \\ \sin x_1 \\ \cos x_2 \end{pmatrix}$$

gegeben.

(c) Die Abbildung $F\colon \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ sei durch

$$F(x_1, x_2) := \begin{pmatrix} x_1 - x_2^3 + 5x_2^2 - 2x_2 - 13 \\ x_1 + x_2^3 + x_2^2 - 14x_2 - 29 \end{pmatrix}$$

gegeben.

Lösung: Wir haben das function file `TrustTsch.m` mit dem folgenden Inhalt geschrieben:

```
function [x,iter]=TrustTsch(Fun,x_0,Delta_0,max_iter,tol);
%*********************************************************
%The Madsen algorithm is used to solve the nonlinear
%Tschebyscheff problem
%              minimize   f(x):=||F(x)||_inf
%*********************************************************
%Input parameter:
%            Fun            function [F(x),F'(x)]=Fun(x)
%            x_0            initial iterate
%            Delta_0        initial trust region radius
%            max_iter       maximal number of iterations
%            tol            tolerance
%Output parameter:
%            x              approximate solution
%            iter           number of iterations
%*********************************************************
rho_1=0.01;rho_2=0.25;rho_3=0.25;sigma_1=0.25;sigma_2=2;
options=optimset('LargeScale','off','Display','off');iter=0;
x_c=x_0;Delta_c=Delta_0;[F_c,J_c]=feval(Fun,x_c);
[m,n]=size(J_c);e_m=ones(m,1);e_n=ones(n,1);
c=[zeros(n,1);1];A=[J_c,-e_m;-J_c,-e_m];b=[-F_c;F_c];
l=[-Delta_c*e_n;-Inf];u=[Delta_c*e_n;Inf];
z=linprog(c,A,b,[],[],l,u,[],options);
p=z(1:n);f=norm(F_c,inf);f_c=z(n+1);
while (f-f_c>tol)&(iter<max_iter)
    iter=iter+1;
    x_plus=x_c+p;
    [F_plus,J_plus]=feval(Fun,x_plus);f_plus=norm(F_plus,inf);
    r_c=(f-f_plus)/(f-f_c);norm_p=norm(p,inf);
    if (r_c<=rho_2)
         Delta_c=sigma_1*norm_p;
    else
       if (norm(F_plus-F_c-J_c*p,inf)<=rho_3*(f-f_plus))
           Delta_c=sigma_2*norm_p;
       else
           Delta_c=norm_p;
       end;
    end;
    if (r_c>=rho_1)
       x_c=x_plus;F_c=F_plus;J_c=J_plus;
    end;
    A=[J_c,-e_m;-J_c,-e_m];b=[-F_c;F_c];
    l=[-Delta_c*e_n;-Inf];u=[Delta_c*e_n;Inf];
    z=linprog(c,A,b,[],[],l,u,[],options);
    p=z(1:n);f=norm(F_c,inf);f_c=z(n+1);
end;
x=x_c;
```

Zur Bearbeitung des ersten Testbeispieles schreiben wir ein function file `Madsen1.m` mit dem Inhalt

```
function [F,J]=Madsen1(x);
t=linspace(-1,1,21)';y=exp(t);
num=x(1)+x(2)*t;den=1+x(3)*t+x(4)*(t.^2)+x(5)*(t.^3);
F=num./den-y;
if nargout>1
  den2=den.^2;
  J=[1./den,t./den,-(num.*t)./den2,-(num.*(t.^2))./den2,-(num.*(t.^3))./den2];
end;
```

Die Befehle

```
>> x_0=zeros(5,1);
>> Delta_0=1;
>> [x,iter]=TrustTsch('Madsen1',x_0,Delta_0,100,1e-12);
```

liefern

$$x = \begin{pmatrix} 0.99987762874885 \\ 0.25358844041148 \\ -0.74660757174630 \\ 0.24520150190227 \\ -0.03749029100843 \end{pmatrix}, \qquad \text{iter} = 10.$$

Für das zweite Testbeispiel schreiben wir ein function file `Madsen2.m` mit dem Inhalt

```
function [F,J]=Madsen2(x);
F=[x(1)^2+x(2)^2+x(1)*x(2);sin(x(1));cos(x(2))];
J=[2*x(1)+x(2),2*x(2)+x(1);cos(x(1)),0;0,-sin(x(2))];
```

Nach den Befehlen

```
>>x_0=[3;1];Delta_0=1.2;
>>[x,iter]=TrustTsch('Madsen2',x_0,Delta_0,100,1e-12);
```

erhalten wir das Ergebnis

$$x = \begin{pmatrix} 0.45329632005639 \\ -0.90659247409256 \end{pmatrix}, \qquad \text{iter} = 31.$$

Selbst mit dem Startwert $x_0 = (10, -10)^T$ erhält man nach 36 Iterationen im wesentlichen dasselbe Ergebnis. Startet man $x_0 = (0, 0)^T$, so erhält man gleich im ersten Schritt einen Abbruch. In der Tat ist $x^* = (0, 0)^T$ eine stationäre Lösung, aber kein lokales Minimum der zugehörigen Approximationsaufgabe.

Für das dritte Beispiel schreiben wir das File `FleWat.m` mit dem Inhalt

```
function [F,J]=FleWat(x);
F=[x(1)-x(2)^3+5*x(2)^2-2*x(2)-13;x(1)+x(2)^3+x(2)^2-14*x(2)-29];
if nargout>1
   J=[1,-3*x(2)^2+10*x(2)-2;1,3*x(2)^2+2*x(2)-14];
end;
```

Bei diesem Beispiel hatten wir beim Gauß-Newton-Verfahren mit Line search Schwierigkeiten. Mit dem Startwert $x_0 = (3, 9)^T$ und $\Delta_0 = 1$ erhalten wir nach

```
[x,iter]=TrustTsch('FleWat',x_0,Delta_0,100,1e-10);
```

das Ergebnis

$$x = \begin{pmatrix} 5.00000000000001 \\ 4.00000000000000 \end{pmatrix}, \qquad \text{iter} = 8.$$

In der Tat ist $x^* = (5,4)^T$ eine Nullstelle von F und daher eine globale Lösung der zugehörigen Tschebyscheffschen Approximationsaufgabe (die auch lokal stark eindeutig ist). Es gibt aber noch eine weitere lokale Lösung, die aber nicht lokal stark eindeutig ist. Nach

```
>>x_0=[10;-2];Delta_0=1;
>>[x,iter]=TrustTsch('FleWat',x_0,Delta_0,100,1e-12);
```

erhalten wir

$$x = \begin{pmatrix} 11.41277858189174 \\ -0.89680528354248 \end{pmatrix}, \qquad \text{iter} = 27.$$

Die exakte lokale Lösung ist

$$x^* = \frac{1}{3}\begin{pmatrix} 53 - 4\sqrt{22} \\ 2 - \sqrt{22} \end{pmatrix} = \begin{pmatrix} 11.41277898690209 \\ -0.89680525327448 \end{pmatrix}.$$

4. Man löse das nichtlineare Ausgleichsproblem

(P), $\qquad$ Minimiere $\quad f(x) := \|F(x)\|_2, \quad x \in \mathbb{R}^n,$

wobei F wie in den drei Beispielen in Aufgabe 3 gegeben ist.

Lösung: Für die drei Aufgaben benutzen wir natürlich die function files `Madse1.m`, `Madsen2.m` und `FleWat.m`. Nach

```
[x,iter]=TrustLeast('Madsen1',zeros(5,1),1,100,1e-12);
```

erhalten wir das Ergebnis

$$x = \begin{pmatrix} 0.99989763243960 \\ 0.25461105075651 \\ -0.74552381361937 \\ 0.24418740089446 \\ -0.03717219743750 \end{pmatrix}, \qquad \text{iter} = 9.$$

Entprechend ist

$$x = \begin{pmatrix} 0.15543784360214 \\ -0.69456373720168 \end{pmatrix}, \qquad \text{iter} = 30$$

das Ergebnis von

```
[x,iter]=TrustLeast('Madsen2',[3;1],1.2,100,1e-12);
```

Nach

```
[x,iter]=TrustLeast('FleWat',[10;-2],1,100,1e-12);
```

erhalten wir

$$x = \begin{pmatrix} 11.41277885557161 \\ -0.89680532100874 \end{pmatrix}, \qquad \text{iter} = 29.$$

Nach

```
[x,iter]=TrustLeast('FleWat',[3;9],1,100,1e-12);
```

ist

$$x = \begin{pmatrix} 5.00000000000000 \\ 4.00000000000000 \end{pmatrix}, \qquad \text{iter} = 8.$$