Das Dennis-Wolkowicz-Verfahren für unrestringierte Optimierungsaufgaben

Diplomarbeit

vorgelegt von

Andreas Klug

aus

Dortmund

angefertigt am

Institut für Numerische und Angewandte Mathematik

der Georg-August-Universität Göttingen

2002

Inhaltsverzeichnis

Kapitel 1

Einleitung

Wir betrachten in dieser Arbeit ein numerisches Verfahren zur Behandlung der *unrestringierten Optimierungsaufgabe*

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

mit einer Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, welche im Laufe der weiteren Betrachtungen zumindest einmal stetig differenzierbar sein wird. Bevor wir jedoch auf dieses Verfahren näher eingehen können, benötigen wir noch einige elementare Grundlagen.

Ziel aller Verfahren zur numerischen Lösung von (P) ist es, einen *stationären Punkt* von f zu finden oder wenigstens eine hinreichend gute Näherung für einen solchen zu liefern. Es handelt sich dabei um einen Punkt $x^* \in \mathbb{R}^n$, in welchem die notwendige Optimalitätsbedingung erster Ordnung erfüllt ist. Dies bedeutet, daß der Zielfunktionsgradient ∇f in x^* verschwindet, also $\nabla f(x^*) = 0$ gilt. Ist zusätzlich die Zielfunktion zweimal stetig differenzierbar und die hinreichende Optimalitätsbedingung zweiter Ordnung erfüllt, beziehungsweise die Hessesche $\nabla^2 f$ in x^* positiv definit, so ist x^* ein (striktes) lokales Minimum von f und somit auch eine lokale Lösung von (P). Genauer verstehen wir unter einer *lokalen Lösung* $x^* \in \mathbb{R}^n$ von (P) einen Punkt, um den es eine Umgebung $U(x^*)$ derart gibt, daß $f(x^*) \leq f(x)$ für alle $x \in U(x^*)$ gilt. Ferner nennen wir $x^* \in \mathbb{R}^n$ eine *globale Lösung* von (P), falls $f(x^*) \leq f(x)$ für alle $x \in \mathbb{R}^n$ gilt. Unter gewissen Zusatzvoraussetzungen für die Zielfunktion, etwa Konvexität, ist jeder stationäre Punkt von f ein globales Minimum und daher jede lokale Lösung von (P) auch globale Lösung.

Da die in der Praxis auftretenden Zielfunktionen oftmals von sehr vielen Variablen abhängen oder die Bestimmung von stationären Punkten mit Methoden der Analysis oder Algebra aufgrund der spezifischen Eigenschaften der Zielfunktion gar nicht möglich ist, man denke in diesem Zusammenhang beispielsweise an Polynome höheren Grades, werden numerische Verfahren zur Lösung von (P) ein-

gesetzt. Eine Vielzahl dieser Verfahren, wie auch das zu untersuchende, gehören zur Klasse der *line-search-Verfahren*. Diese lassen sich wie folgt beschreiben: Ausgehend von einer aktuellen Näherung x_k für einen stationären Punkt bestimmt man eine sogenannte *Abstiegsrichtung* $p_k \in \mathbb{R}^n$ für f in x_k. Dies ist ein Vektor, für den $\nabla f(x_k)^T p_k < 0$ gilt. Hieraus folgt $f(x_k + tp_k) < f(x_k)$ für alle hinreichend kleinen $t > 0$, also eine Zielfunktionsverminderung falls man von x_k einen hinreichend kleinen Schritt in Richtung p_k geht. Man kann eine Abstiegsrichtung auf verschiedene Arten ermitteln, wir werden später darauf eingehen. Als nächstes bestimmt man eine Schrittweite $t_k > 0$, die, wie der Name schon erahnen läßt, festlegt, wie weit man von x_k in die Richtung p_k geht. Auch hierbei gibt es zahlreiche Möglichkeiten, eine Schrittweite zu bestimmen. Es ist aber sinvoll, eine Forderung der Form $f(x_k + t_k p_k) < f(x_k)$ zu stellen, denn dann ist mit der Wahl $x_{k+1} := x_k + t_k p_k$ für eine neue Näherung die Folge $\{f(x_k)\}$ streng monoton fallend. Schrittweiten, die diese Forderung erfüllen, werden daher auch als monoton bezeichnet. Die wichtigsten hiervon werden wir noch vorstellen. Man führt die eben beschriebenen Schritte solange durch, bis ein stationärer Punkt erreicht oder $\|\nabla f(x_k)\|$ hinreichend nahe bei 0 ist. Dabei bezeichnet $\|\cdot\|$ hier und im weiteren Verlauf dieser Arbeit stets die euklidische Vektornorm, beziehungsweise die zugeordnete Matrixnorm. Formal lassen sich line-search-Verfahren dem folgenden *Modellalgorithmus* unterordnen:

- Gegeben sei ein Startvektor $x_0 \in \mathbb{R}^n$.
- Für $k = 0, 1, \ldots$:
  - Falls $\nabla f(x_k) = 0$, dann STOP. x_k ist stationärer Punkt.
  - Andernfalls:
    1. Bestimme eine (Abstiegs-)Richtung $p_k \in \mathbb{R}^n$ mit $\nabla f(x_k)^T p_k < 0$.
    2. Bestimme eine Schrittweite $t_k > 0$ mit $f(x_k + t_k p_k) < f(x_k)$.
    3. Bestimme die neue Näherung $x_{k+1} := x_k + t_k p_k$.

Line-search-Verfahren setzen sich also aus einer Richtungs- und einer Schrittweitenstrategie zusammen. Auf eine Gruppe von Möglichkeiten für die Wahl der ersteren wollen wir hier schon eingehen und die sogenannten *Quasi-Newton-Verfahren* erläutern. Diese entstehen aus dem auf die Gleichung $\nabla f(x) = 0$ angewandten Newton-Verfahren, indem anstatt der bekannten Richtungswahl

$$p_k := -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

die Richtung

$$p_k := -B_k^{-1} \nabla f(x_k)$$

mit einer nichtsingulären Matrix $B_k \in \mathbb{R}^{n \times n}$ gewählt wird. Ist die Matrix B_k, die man auch als eine Approximation an die Hessesche in der aktuellen Näherung

betrachten kann, symmetrisch und positiv definit, so ist $p_k := -B_k^{-1}\nabla f(x_k)$ eine Abstiegsrichtung für f in x_k. Die Bestimmung dieser Matrizen erfolgt über eine Update-Formel für die Folgenglieder der Folge $\{B_k\}$. In die Berechnung von B_{k+1} gehen üblicherweise die aktuelle Matrix B_k sowie $y_k := \nabla f(x_{k+1}) - \nabla f(x_k)$ und $s_k := x_{k+1} - x_k$ ein. Auf Basis dieser einführenden Grundlagen sind wir nun in der Lage, näher auf konkrete Verfahren zur numerischen Behandlung des Problems (P) einzugehen.

Das *BFGS-Verfahren* wurde 1970 unabhängig voneinander von C. G. Broyden, R. Fletcher, D. Goldfarb und D. F. Shanno entdeckt. Dabei handelt es sich um ein Quasi-Newton-Vefahren mit der Update-Formel

$$B_{k+1} := B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}.$$

Wegen seiner guten Konvergenzeigenschaften ist dieses Verfahren in der Praxis äußerst beliebt und daher auch eines der meist untersuchten Verfahren der unrestringierten Optimierung. Wir werden einige dieser Ergebnisse später noch vorstellen, wenden uns nun jedoch dem eigentlichen Objekt dieser Arbeit zu.

1993 verglichen J. E. Dennis Jr. und H. Wolkowicz in ihrer Arbeit [9] die numerischen Eigenschaften mehrerer bereits bekannter und einiger neuer Quasi-Newton-Verfahren. Hierbei stellte sich ein von den Autoren erstmals vorgestelltes Verfahren als überlegen in den Testkriterien Anzahl der Iterationen und Anzahl der Zielfunktionsauswertungen heraus. Dieses Verfahren Nr. 10 aus [9] nennen wir der Einfachheit halber von nun an das *Dennis-Wolkowicz-Verfahren* oder kurz *DW-Verfahren*. Es benutzt ein zweistufiges Matrix-Update der Form

$$\begin{aligned} B_{k+\frac{1}{2}} &:= B_k + \frac{y_k^T B_k^{-1} y_k - y_k^T s_k}{(y_k^T B_k^{-1} y_k)(y_k^T s_k)} y_k y_k^T, \\ B_{k+1} &:= B_{k+\frac{1}{2}} - \frac{(B_{k+\frac{1}{2}} s_k)(B_{k+\frac{1}{2}} s_k)^T}{s_k^T B_{k+\frac{1}{2}} s_k} + \frac{y_k y_k^T}{y_k^T s_k}. \end{aligned}$$

Dies stellt die unmittelbare Hintereinanderausführung eines inversen schwachen Greenstadt-Updates (in der direkten Darstellung) und eines BFGS-Updates dar. Desweiteren führen Dennis und Wolkowicz im ersten Schritt des Verfahrens ein sogenanntes *initial inverse sizing* aus. Es bedeutet, daß unmittelbar vor dem ersten Matrix-Update

$$B_0^{-1} := \frac{y_0^T s_0}{y_0^T B_0^{-1} y_0} B_0^{-1}$$

gesetzt wird. Dies hat zwar keine Auswirkungen auf die theoretischen Eigenschaften des Verfahrens, bewirkt allerdings Verbesserungen bei den numerischen Tests. Das Dennis-Wolkowicz-Verfahren ist noch recht wenig untersucht, die bisher

untersuchten Konvergenzeigenschaften sind jedoch denen des BFGS-Verfahrens sehr ähnlich.

Ziel dieser Arbeit ist es nun, die theoretischen und numerischen Eigenschaften des Dennis-Wolkowicz-Verfahrens vorzustellen und diese mit denen anderer Verfahren zu vergleichen. Als Referenz-Verfahren dient dabei das BFGS-Verfahren. Für den Vergleich der theoretischen Eigenschaften werden wir zunächst die schon bekannten lokalen und globalen Konvergenzresultate für das BFGS-Verfahren vorstellen. Einige dieser Eigenschaften konnten bereits auf das DW-Verfahren übertragen und bewiesen werden. Wir werden diese ebenfalls vorstellen. Ferner werden wir die Gültigkeit einiger weiterer Sätze über das BFGS-Verfahren für das DW-Verfahren beweisen, beziehungsweise bereits bewiesene Aussagen verallgemeinern. Anschließend werden wir beide Verfahren implementieren und ihre numerischen Eigenschaften miteinander vergleichen. Die weitere Arbeit gliedert sich wie folgt:

In Kapitel 2 werden wir einige allgemeine theoretische Grundlagen von linesearch-Verfahren für die späteren Untersuchungen bereitstellen. Dazu gehören unter anderem Eigenschaften von bestimmten Quasi-Newton-Updates, Schrittweitenstrategien, sowie allgemeine Konvergenzaussagen für den Modellalgorithmus und Quasi-Newton-Verfahren.

In Kapitel 3 werden die globalen und lokalen Konvergenzeigenschaften des BFGS-Verfahrens wiederholt und ein modifiziertes BFGS-Verfahren sowie dessen globale Konvergenz bei nichtkonvexer Zielfunktion vorgestellt.

In Kapitel 4 stellen wir analog zum vorangegangenen Kapitel die globalen und lokalen Konvergenzeigenschaften des Dennis-Wolkowicz-Verfahrens vor und liefern in diesem Zusammenhang einige für dieses Verfahren neue, für das BFGS-Verfahren schon bekannte Ergebnisse. Als Stichworte hierzu seinen bereits jetzt globale R-lineare Konvergenz bei semi-effizienter Schrittweitenstrategie, Gültigkeit eines Bounded-Deterioration-Satzes und daraus folgende lokale Q-lineare Konvergenz des ungedämpften Verfahrens sowie globale Konvergenz bei nichtkonvexer Zielfunktion mittels "cautious" Update genannt.

In Kapitel 5 werden wir Grundlagen für eine effiziente Implementierung beider Verfahren vorstellen und anschließend einige numerische Experimente mit beiden Verfahren beschreiben und auswerten. Dabei werden wir noch ein weiteres neues Verfahren von J. E. Dennis Jr. und H. Wolkowicz berücksichtigen. Da zu diesem Verfahren allerdings bislang keine Konvergenzbeweise vorliegen, beschränken wir uns auf die numerische Untersuchung.

In Kapitel 6 werden wir auf Basis der vorangegangenen Kapitel ein Fazit über die mögliche Überlegenheit des Dennis-Wolkowicz-Verfahrens ziehen.

Kapitel 2

Theoretische Grundlagen

In diesem Kapitel beschäftigen wir uns mit allgemeinen theoretischen Grundlagen von line-search-Verfahren, welche wir dann später bei bei der genaueren Untersuchung der Eigenschaften von BFGS- und DW-Verfahren benötigen werden. Wir betrachten im folgenden wieder die unrestringierte Optimierungsaufgabe

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

mit einer stetig differenzierbaren Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, von welcher wir später noch weitere Eigenschaften fordern werden.

Die wohl wichtigste Klasse von line-search-Verfahren sind die schon beschriebenen Quasi-Newton-Verfahren, bei denen die Folge der Abstiegsrichtungen durch eine Folge von Matrizen und deren Update-Formel bestimmt wird. Läßt sich die Update-Formel in einer gewissen Art und Weise, die wir gleich beschreiben werden, darstellen, so spricht man von einem *Broydenklasse-Verfahren*. Da, soviel sei hier schon vorweg genommen, alle in dieser Arbeit zu untersuchenden Verfahren zu dieser Verfahrensklasse gehören, werden wir einige ihrer Eigenschaften jetzt näher erläutern.

2.1 Broydenklasse-Verfahren

Broydenklasse-Verfahren stellen eine in Theorie und Praxis sehr beliebte Gruppe von Verfahren zur numerischen Behandlung unrestringierter Optimierungsaufgaben dar. Diese Quasi-Newton-Verfahren sind durch Matrix-Update-Formeln der Form

$$B_{k+1} := B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} + (1 - \phi_k)(s_k^T B_k s_k) v_k v_k^T$$

gegeben. Hierbei sind

$$s_k := x_{k+1} - x_k, \qquad y_k := \nabla f(x_{k+1}) - \nabla f(x_k), \qquad v_k := \frac{y_k}{y_k^T s_k} - \frac{B_k s_k}{s_k^T B_k s_k}$$

und $\phi_k \in \mathbb{R}$ der Broydenklasseparameter. Die bekanntesten Vertreter dieser Verfahrensklasse sind das schon erwähnte BFGS-Verfahren mit $\phi_k = 1$ und das auf W. C. Davidon, R. Fletcher und M. J. D. Powell zurückgehende DFP-Verfahren mit $\phi_k = 0$. Wir werden nun einige allgemeine Eigenschaften dieser Update-Klasse angeben, um diese später auf die zu untersuchenden Spezialfälle anwenden zu können.

Unter der Voraussetzung $y_k^T s_k > 0$ sagt der nun folgende Satz aus, daß die Update-Matrizen der Broydenklasse der Quasi-Newton-Gleichung genügen und für $\phi_k \leq 1$ mit B_k auch B_{k+1} symmetrisch und positiv definit ist. Ferner wird die Determinante von B_{k+1} aus der von B_k berechnet.

Satz 2.1 *Seien $y_k, s_k \in \mathbb{R}^n$ mit $y_k^T s_k > 0$ und eine symmetrische und positiv definite Matrix $B_k \in \mathbb{R}^{n \times n}$ gegeben. Sei $\phi_k \in \mathbb{R}$ und*

$$B_{k+1} := B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} + (1 - \phi_k)(s_k^T B_k s_k) v_k v_k^T$$

mit

$$v_k := \frac{y_k}{y_k^T s_k} - \frac{B_k s_k}{s_k^T B_k s_k}.$$

Dann gilt:

1. *Die Quasi-Newton-Gleichung $B_{k+1} s_k = y_k$ ist erfüllt.*
2. *Für $\phi_k \leq 1$ ist die Matrix B_{k+1} symmetrisch und positiv definit.*
3. *Es gilt*

$$\det(B_{k+1}) = \left[\phi_k \frac{y_k^T s_k}{s_k^T B_k s_k} + (1 - \phi_k) \frac{y_k^T B_k^{-1} y_k}{y_k^T s_k} \right] \det(B_k).$$

Beweise der obigen Aussagen lassen sich bei J. Werner [23], S. 196f. und J. D. Pearson [18] finden.

Wie sich im vorangegangenen Satz gezeigt hat, ist die Voraussetzung $y_k^T s_k > 0$ für Broydenklasse-Verfahren von großer Bedeutung, da sie unter $\phi_k \leq 1$ eine hinreichende Bedingung für die positive Definitheit von B_{k+1} und damit für eine Abstiegsrichtung ist. Wie wir später noch sehen werden, kann man $y_k^T s_k > 0$ durch Verwendung gewisser Schrittweiten oder einer gleichmäßig konvexen Zielfunktion sicherstellen.

Bemerkung: Eine weitere wichtige Eigenschaft von Broydenklasse-Verfahren ist die Invarianz unter der affin linearen Variablentransformation $z := Ax + a$ mit

einer nichtsingulären Matrix $A \in \mathbb{R}^{n\times n}$ und einem $a \in \mathbb{R}^n$. Ferner kann man zeigen, daß alle Broydenklasse-Verfaren dieselben Folgen $\{x_k\}$ und $\{B_k\}$ erzeugen, falls dieselben Startwerte x_0 und B_0 und die exakte Schrittweite, auf die wir im nächsten Abschnitt eingehen, verwendet werden.

Für Beweise dieser Aussagen siehe z. B. R. Fletcher [10], S. 57 und S. 66.

2.2 Schrittweitenstrategien

Bekannterweise setzen sich line-search-Verfahren aus einer Richtungs- und einer Schrittweitenstrategie zusammen. Möglichkeiten für die Wahl der letzteren werden wir jetzt vorstellen. Wir betrachten wieder die unrestringierte Optimierungsaufgabe (P) und setzen voraus:

(V_1) Bei gegebenem Startvektor $x_0 \in \mathbb{R}^n$ des Verfahrens ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n \ : \ f(x) \leq f(x_0)\}$ kompakt.

(V_2) Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar.

(V_3) Der Zielfunktionsgradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig. Es existiert also eine Konstante $\gamma > 0$ mit

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

Ausgehend von einer aktuellen Näherung $x_k \in L_0$, die keine stationäre Lösung von (P) ist und einer Abstiegsrichtung p_k für f in x_k berechnen sich die gängigsten Schrittweiten $t_k > 0$ wie folgt.

2.2.1 Die exakte Schrittweite

Bei der Suche nach einer Schrittweite würde es sich natürlich anbieten, t_k als Lösung der restringierten Optimierungsaufgabe

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

zu bestimmen. Dann ist

$$f(x_k + t_k p_k) = \min_{t \geq 0} f(x_k + tp_k),$$

beziehungsweise t_k diejenige Schrittweite, die bei gegebener Abstiegsrichtung p_k die maximale Zielfunktionsverminderung liefert. Genauer kann man zeigen, daß in jedem Iterationsschritt

$$f(x_k) - f(x_k + t_k p_k) \geq \frac{1}{2\gamma} \left(\frac{\nabla f(x_k)^T p_k}{\|p_k\|} \right)^2$$

gilt. Ferner gilt mit dieser Schrittweite $\nabla f(x_{k+1})^T p_k = 0$ für alle k und daher $y_k^T s_k > 0$. Dies ist, wie Satz 2.1 zeigt, im Zusammenhang mit Broydenklasse-Verfahren von besonderer Bedeutung. Denn somit ist es möglich, die Durchführbarkeit dieser Verfahren und den Erhalt der positiven Definitheit der Update-Matrizen zu sichern, ohne etwa eine gleichmäßige Konvexitätsforderung an die Zielfunktion zu stellen. Dies wird uns später noch bei der Vorstellung des CBFGS-Verfahrens beschäftigen. Die Existenz dieser Schrittweite, die auch *exakte Schrittweite* genannt wird, ist unter den gegebenen Voraussetzungen gesichert. Allerdings ist es nur in wenigen Spezialfällen möglich, diese genau zu berechnen. Man muss sich daher mit Näherungen begnügen oder *inexakte Schrittweiten* verwenden. In den Anwendungen wird zumeist letzteres getan, weshalb wir die beiden wichtigsten nun angeben.

2.2.2 Die Wolfe-Schrittweite

Der wohl prominenteste Vertreter der inexakten Schrittweiten ist die *Wolfe-Schrittweite*, zum Teil auch *Powell-Schrittweite* genannt. Hierbei wird $t_k > 0$ so gewählt, daß die beiden *Wolfe-Bedingungen*

$$f(x_k + t_k p_k) \leq f(x_k) + \alpha t_k \nabla f(x_k)^T p_k \quad \text{und} \quad \nabla f(x_k + t_k p_k)^T p_k \geq \beta \nabla f(x_k)^T p_k$$

mit Konstanten $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ erfüllt sind. Auch hier sichern die oben getroffenen Voraussetzungen (V_1) bis (V_3) die Existenz einer solchen Schrittweite. Eine Abschätzung der Zielfunktionsverminderung ist ebenfalls möglich. Unter den Voraussetzungen gilt

$$f(x_k) - f(x_k + t_k p_k) \geq \frac{\alpha(1-\beta)}{\gamma} \left(\frac{\nabla f(x_k)^T p_k}{\|p_k\|} \right)^2.$$

Desweiteren folgt bei Verwendung einer Abstiegsrichtung aus der zweiten Wolfe-Bedingung, daß $y_k^T s_k > 0$ gilt. Die enorme Wichtigkeit dieser Eigenschaft haben wir schon bei der exakten Schrittweite betont. Da zusätzlich die Berechnung einer Wolfe-Schrittweite in endlich vielen Schritten möglich ist, wird diese in den praktischen Anwendungen der exakten Schrittweite zumeist vorgezogen. Im Rahmen der numerischen Experimente werden wir daher die Wolfe-Schrittweite benuzten und auch einen Algorithmus für die Berechnung angeben und implementieren.

2.2.3 Die Armijo-Schrittweite

Eine Schrittweite, deren Vorteile vor allem in der leichten Implementierbarkeit liegen, ist die *Armijo-Schrittweite*. Diese berechnet sich wie folgt:

- Geben seien $\alpha \in (0, \frac{1}{2})$ und $0 < l \leq u < 1$. Setze $\rho_0 := 1$.

- Für $j = 0, 1, \ldots$:

  Falls $f(x_k + \rho_j p_k) \leq f(x_k) + \alpha\rho_j \nabla f(x_k)^T p_k$, dann: Setze $t_k := \rho_j$ und STOP.

  Andernfalls: Wähle $\rho_{j+1} \in [l\rho_j, u\rho_j]$.

Setzt man beispielsweise $l = u =: \rho \in (0, 1)$, so erhält man $t_k = \rho^j$, wobei j die kleinste nichtnegative ganze Zahl mit

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

ist. Dieser wegen seiner Einfachheit sehr beliebte Spezialfall wird auch *backtracking Armijo-Schrittweite* genannt. Wieder sichern (V_1) bis (V_3) die Existenz dieser Schrittweite und wieder erlauben sie eine Abschätzung der erreichbaren Zielfunktionsverminderung. Genauer existiert unter (V_1) bis (V_3) eine Konstante $\theta > 0$, die zwar von α, γ, l und u abhängt, jedoch unabhängig von x_k und p_k ist, mit

$$f(x_k) - f(x_k + t_k p_k) \geq \theta \min\left[-\nabla f(x_k)^T p_k, \left(\frac{\nabla f(x_k)^T p_k}{\|p_k\|}\right)^2\right].$$

Allerdings kann durch Verwendung der Armijo-Schrittweite $y_k^T s_k > 0$ nicht sichergestellt werden, was den wesentlichen Nachteil dieser Strategie darstellt.

Bemerkung: Die angegebenen Abschätzungen für die Zielfunktionsverminderung bei Verwendung einer bestimmten Schrittweitenstrategie spielen eine zentrale Rolle bei der Konvergenzanalyse des Modellalgorithmus. Denn ist eine solche Abschätzung ersteinmal gezeigt, so kann die weitere Untersuchung unabhängig von der speziellen Wahl der Schrittweite durchgeführt werden. Schrittweitenstrategien, für die unter den Voraussetzungen (V_1) bis (V_3) die Existenz einer von x_k und p_k unabhängigen Konstanten $\theta > 0$ mit

$$f(x_k) - f(x_k + t_k p_k) \geq \theta\left(\frac{\nabla f(x_k)^T p_k}{\|p_k\|}\right)^2$$

gezeigt werden kann, werden gemäß [20] von W. Warth und J. Werner *effizient* genannt. Kann unter (V_1) bis (V_3) gezeigt werden, daß es eine von x_k und p_k unabhängige Konstante $\theta > 0$ mit

$$f(x_k) - f(x_k + t_k p_k) \geq \theta \min\left[-\nabla f(x_k)^T p_k, \left(\frac{\nabla f(x_k)^T p_k}{\|p_k\|}\right)^2\right]$$

existiert, so wird die Schrittweitenstrategie *semi-effizient* genannt, siehe P. Kosmol [13]. Die exakte und die Wolfe-Schrittweite sind folglich effizient. Die Armijo-Schrittweite und jede effiziente Schrittweite sind semi-effizient.

Beweise der Aussagen über die drei vorgestellten Schrittweitenstrategien sind zum Beispiel bei J. Werner [23], S. 163ff. zu finden.

2.3 Allgemeine Konvergenzaussagen

Bei der Untersuchung der theoretischen Konvergenzeigenschaften von BFGS- und Dennis-Wolkowicz-Verfahren werden wir einige allgemein bekannte Konvergenzresultate über line-search-Verfahren, beziehungsweise den Modellalgorithmus sowie Quasi-Newton-Verfahren im speziellen benötigen. Daher stellen wir diese kurz vor, definieren jedoch zuvor noch die auftretenden Konvergenzraten.

Definition 2.2 Sei $\{x_k\} \subset \mathbb{R}^n$ eine Folge von Vektoren, die gegen ein $x^* \in \mathbb{R}^n$ konvergiert. Dann heißt die Konvergenzrate von $\{x_k\}$:

1. *Q-linear*, wenn ein $r \in (0,1)$ mit $\|x_{k+1} - x^*\| \leq r\|x_k - x^*\|$ für alle hinreichend großen k existiert.

2. *R-linear*, wenn eine Folge nichtnegativer Skalare $\{a_k\} \subset \mathbb{R}$ derart existiert, daß $\|x_k - x^*\| \leq a_k$ für alle k gilt und $\{a_k\}$ Q-linear gegen Null konvergiert.

3. *Q-superlinear*, wenn
$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0$$
gilt.

Als erstes geben wir eine hinreichende Bedingung für globale R-lineare Konvergenz von line-search-Verfahren bei gleichmäßig konvexer Zielfunktion an. Wir benötigen die folgenden Voraussetzungen:

(K_1) Bei gegebenem Startvektor $x_0 \in \mathbb{R}^n$ des Verfahrens ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n \ : \ f(x) \leq f(x_0)\}$ konvex.

(K_2) Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar und gleichmäßig konvex auf L_0. Dies bedeutet die Existenz einer Konstanten $c > 0$ mit
$$\frac{c}{2}\|y - x\|^2 + \nabla f(x)^T(y - x) \leq f(y) - f(x) \qquad \text{für alle } x, y \in L_0.$$

(K_3) Der Zielfunktionsgradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig. Es existiert also eine Konstante $\gamma > 0$ mit
$$\|\nabla f(x) - \nabla f(y)\| \leq \gamma\|x - y\| \qquad \text{für alle } x, y \in L_0.$$

Bemerkung: Unter diesen Voraussetzungen ist die Niveaumenge kompakt und die Voraussetzungen (V_1) bis (V_3) damit ebenfalls erfüllt. Ferner besitzt (P) eine globale Lösung $x^* \in L_0$, die der einzige stationäre Punkt in L_0 ist.

Ein Beweis dieser Bemerkung sowie des sich anschließenden Konvergenzsatzes für den Modellalgorithmus ist bei J. Werner [23], S. 171ff. zu finden.

Satz 2.3 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Voraussetzungen (K_1) bis (K_3) seien erfüllt. Man betrachte den Modellalgorithmus mit Abstiegsrichtungen p_k und (zumindest) semi-effizienten Schrittweiten t_k. Es seien $g_k := \nabla f(x_k)$ und*

$$\delta_k := \begin{cases} \min\left[-\frac{g_k^T p_k}{\|g_k\|^2}, \left(\frac{g_k^T p_k}{\|g_k\|\|p_k\|}\right)^2\right] & \text{falls } t_k \text{ semi-effiziente Schrittweite,} \\ \left(\frac{g_k^T p_k}{\|g_k\|\|p_k\|}\right)^2 & \text{falls } t_k \text{ effiziente Schrittweite.} \end{cases}$$

Dann gilt: Existiert für alle $k = 0, 1, \ldots$ eine Konstante $\delta > 0$ mit

$$\delta(k+1) \leq \sum_{j=0}^{k} \delta_j,$$

so konvergiert die vom Modellalgorithmus erzeugte Folge $\{x_k\}$ R-linear gegen die eindeutige Lösung x^* von (P) in L_0. Genauer existieren Konstanten $C > 0$ und $q \in (0,1)$ derart, daß $\|x_k - x^*\| \leq Cq^k$ für alle $k = 0, 1, \ldots$ gilt.

Nun betrachten wir speziell Quasi-Newton-Verfahren zur Behandlung unrestringierter Optimierungsaufgaben. Ein äußerst nützliches Hilfsmittel zur Untersuchung der Konvergenzrate von ungedämpften Quasi-Newton-Verfahren haben J. E. Dennis Jr. und J. J. Moré 1974 in [6] veröffentlicht. Ein mittlerweile nach den Autoren benannter Satz ermöglicht es, Q-superlineare Konvergenz zu beweisen, falls die Bedingung

$$\lim_{k\to\infty} \frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} = 0$$

erfüllt ist. Diese Bedingung wird auch *Dennis-Moré-Bedingung* genannt. Wir benötigen die folgende Voraussetzung:

(V) Die Zielfunktion f ist auf einer konvexen Umgebung der (isolierten) lokalen Lösung $x^* \in \mathbb{R}^n$ zweimal stetig differenzierbar, die Hessesche $\nabla^2 f$ dort in x^* lipschitzstetig und $\nabla^2 f(x^*)$ positiv definit.

Satz 2.4 *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^*. Ferner sei $\nabla f(x_k) \neq 0$ für alle k. Dann gilt: Die Folge $\{x_k\}$ konvergiert genau dann Q-superlinear gegen x^*, wenn

$$\lim_{k\to\infty} \frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} = 0$$

gilt.

Desweiteren zeigen DENNIS und MORÉ, daß die Aussage dieses Satzes auch für gedämpfte Verfahren erhalten bleibt, falls die Folge der Schrittweiten $\{t_k\}$ gegen 1 konvergiert.

Ist die Dennis-Moré-Bedingung erfüllt, so zeigt der nächste Satz, daß bei Dämpfung des Verfahrens durch Wolfe- oder Armijo-Schrittweiten der Übergang in das ungedämpfte Verfahren folgt. Dabei ist es natürlich wichtig, bei der Bestimmung der Wolfe-Schrittweiten immer zuerst $t_k = 1$ zu testen.

Satz 2.5 *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. die Folge $\{B_k^{-1}\}$ 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^*. Ferner sei $\nabla f(x_k) \neq 0$ für alle k und t_k die Wolfe- oder Armijo-Schrittweite. Ist dann

$$\lim_{k\to\infty} \frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} = 0,$$

so gilt: Es ist $t_k = 1$ für alle hinreichend großen k und die Folge $\{x_k\}$ konvergiert superlinear gegen x^*.

Für einen Beweis verweisen wir auf J. WERNER [22].

Damit beenden wir die Zusammenfassung der wichtigsten theoretischen Grundlagen und untersuchen jetzt die beiden zu vergleichenden Verfahren.

Kapitel 3

Das BFGS-Verfahren

In diesem Kapitel werden wir uns mit den theoretischen Eigenschaften des BFGS-Verfahrens beschäftigen und die wichtigsten schon bewiesenen globalen und lokalen Konvergenzaussagen zusammenstellen. Ferner werden wir das auf D.-H. Li und M. Fukushima zurückgehende BFGS-Verfahren mit "cautious" Update vorstellen und seine globale Konvergenz bei nichtkonvexer Zielfunktion beweisen. In der Darstellung dieses Kapitels folgen wir im wesentlichen der Arbeit [22] von J. Werner und betrachten wieder die unrestringierte Optimierungsaufgabe

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

mit einer stetig differenzierbaren Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$.

3.1 Globale Konvergenzaussagen

Beweise für die globale Konvergenz des gedämpften BFGS-Verfahrens sind in der Literatur zahlreich vertreten. Besonders erwähnen wollen wir hier unter anderem M. J. D. Powell, der 1976 in [19] als erster globale Konvergenz des BFGS-Verfahrens mit Wolfe-Schrittweite bei konvexer Zielfunktion zeigen konnte. Eine Verallgemeinerung dieser Aussage auf effiziente Schrittweitenstrategien erfolgte 1978 durch J. Werner in [24]. 1989 wurde von R. H. Byrd und J. Nocedal in [3] eine neue Beweistechnik vorgestellt. Diese ermöglicht es, durch Verwendung der auf der Menge der symmetrischen und positiv definiten Matrizen definierten Funktion ψ mit $\psi(A) := \operatorname{tr}(A) - \ln\det(A)$ die globale R-lineare Konvergenz des BFGS-Verfahrens mit semi-effizienter Schrittweitenstrategie bei gleichmäßig konvexer Zielfunktion zu beweisen. Desweiteren werden wir auf den Spezialfall quadratischer Zielfunktionen und die Frage nach der Konvergenz bei nichtkonvexer Zielfunktion eingehen.

Der nun folgende Satz gibt die globale R-lineare Konvergenz des durch eine zu-

mindest semi-effiziente Schrittweite gedämpften BFGS-Verfahrens an. Wir setzen hierfür voraus:

(K_1) Bei gegebenem Startvektor $x_0 \in \mathbb{R}^n$ des Verfahrens ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n \ : \ f(x) \leq f(x_0)\}$ konvex.

(K_2) Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar und gleichmäßig konvex auf L_0. Dies bedeutet die Existenz einer Konstanten $c > 0$ mit

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

(K_3) Der Zielfunktionsgradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig. Es existiert also eine Konstante $\gamma > 0$ mit

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

Satz 3.1 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Voraussetzungen (K_1) bis (K_3) seien erfüllt. Mit einem $x_0 \in \mathbb{R}^n$ und einer symmetrischen und positiv definiten Matrix $B_0 \in \mathbb{R}^{n\times n}$ sei die Folge $\{x_k\}$ durch das gedämpfte BFGS-Verfahren erzeugt. Ferner gelte (für die gewählte Schrittweitenstrategie) in jedem Iterationsschritt $k = 0, 1, \ldots$*

$$f(x_k) - f(x_k + t_k p_k) \geq \theta \min\left[-\nabla f(x_k)^T p_k, \left(\frac{\nabla f(x_k)^T p_k}{\|p_k\|}\right)^2\right]$$

mit einer Konstanten $\theta > 0$. Dann gilt: Entweder bricht das Verfahren nach endlich vielen Schritten mit der Lösung x^* von (P) in L_0 ab, oder die erzeugte Folge $\{x_k\}$ konvergiert R-linear gegen x^*.

Einen Beweis dieser Aussage findet man bei R. H. Byrd und J. Nocedal in [3] oder bei J. Werner in [22].

Nun wollen wir noch das klassische globale Konvergenzresultat von Powell angeben. Die Voraussetzungen sind schwächer als die von Satz 3.1, dafür ist jedoch keine Angabe der Konvergenzrate möglich, und die Dämpfung des Verfahrens beschränkt sich auf die Wolfe-Schrittweite.

Satz 3.2 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Zielfunktion f sei konvex und auf einer offenen Obermenge der Niveaumenge L_0 zweimal stetig differenzierbar. Ferner existiere eine Konstante $M > 0$ derart, daß $\|\nabla^2 f(x)\| \leq M$ für alle $x \in L_0$ gilt. Mit $x_0 \in \mathbb{R}^n$ und $B_0 \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit erzeuge das durch die Wolfe-Schrittweite gedämpfte BFGS-Verfahren eine Folge $\{x_k\}$. Dann gilt: Ist die Zielfunktion f auf der Niveaumenge L_0 nach unten beschränkt, so gilt*

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0.$$

Für einen Beweis siehe M. J. D. Powell [19] oder J. Werner [22].

Bemerkung: Wegen der vorausgesetzten Konvexität der Zielfunktion folgt aus Satz 3.2 nicht nur Konvergenz einer Teilfolge von $\{x_k\}$ gegen eine lokale Lösung von (P), sondern auch daß jeder Häufungspunkt von $\{x_k\}$ eine globale Lösung des Problems ist.

Bemerkung: Die beiden vorangegangenen Sätze zeigen globale Konvergenz des gedämpften BFGS-Verfahrens für konvexe und spezieller gleichmäßig konvexe Zielfunktionen. Es war jedoch sehr lange unklar, ob das BFGS-Verfahren auch bei nichtkonvexer Zielfunktion global konvergiert. Weder konnte Konvergenz bewiesen noch ein Gegenbeispiel angegeben werden. Erst Ende 2002 gelang es Y.-H. Dai, in der Arbeit [4] eine Zielfunktion zu konstruieren, für die das gedämpfte BFGS-Verfahren nicht konvergiert. Die Hauptaussage aus [4] ist der folgende Satz.

Satz 3.3 *Man betrachte das gedämpfte BFGS-Verfahren mit Wolfe-Schrittweite. Für die Parameter der Schrittweitenstrategie gelte* $0 < \alpha \leq \frac{69}{7480}$ *und* $\beta \in (\alpha, 1)$. *Dann existieren für alle* $n \geq 2$ *eine beliebig oft stetig differenzierbare Zielfunktion* $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, *ein Startvektor* $x_0 \in \mathbb{R}^n$ *und eine Konstante* $\eta > 0$ *derart, daß für die vom Verfahren erzeugte Folge* $\{x_k\}$

$$\|\nabla f(x_k)\| \geq \eta$$

in allen Iterationsschritten k *gilt. Insbesondere konvergiert das Verfahren nicht.*

Bemerkung: Die Aussage des vorangegangenen Satztes kann nach einer Bemerkung von Y.-H. Dai aus [4] auf alle Broydenklasse-Verfahren mit einem Broydenklasse-Parameter $\phi_k \leq 1$ verallgemeinert werden.

Falls man also das gedämpfte BFGS-Verfahren auf eine nichtkonvexe Zielfunktion anwendet, ist die globale Konvergenz nicht sichergestellt. Allerdings konnten D.-H. Li und M. Fukushima in den Arbeiten [14] und [15] die Update-Matrix des BFGS-Verfahrens so modifizieren, daß die geänderten Verfahren auch im nichtkonvexen Fall konvergieren. Hierauf werden wir in Abschnitt 3.3 ausführlich eingehen.

Ist die Zielfunktion f quadratisch und gleichmäßig konvex, so ist es möglich, ein über Satz 3.1 hinausgehendes Konvergenzresultat zu beweisen. In diesem Fall konvergiert das ungedämpfte BFGS-Verfahren global und Q-superlinear.

Satz 3.4 *Gegeben sei die unrestringierte Optimierungsaufgabe*

$$\textit{Minimiere} \quad f(x) := c^T x + \frac{1}{2} x^T Q x, \quad x \in \mathbb{R}^n,$$

wobei $c \in \mathbb{R}^n$ und $Q \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit ist. Dann gilt: Entweder bricht das auf diese Aufgabe angewandte ungedämpfte BFGS-Verfahren nach endlich vielen Schritten mit der eindeutigen Lösung $x^* = -Q^{-1}c$ der Aufgabe ab, oder es liefert eine Folge $\{x_k\}$ die global und Q-superlinear gegen x^* konvergiert.

Ein die ψ-Funktion nutzender Beweis ist bei J. Werner in [22] zu finden.

Damit ist dieser Abschnitt beendet, und wir kommen zu den lokalen Konvergenzaussagen für das BFGS-Verfahren.

3.2 Lokale Konvergenzaussagen

Ziel dieses Abschnitts ist es, die lokale Q-lineare Konvergenz des ungedämpften BFGS-Verfahrens sowie die Dennis-Moré-Bedingung für dieses Verfahren nachzuweisen. Hieraus läßt sich dann unter anderem lokale Q-superlineare Konvergenz folgern. Wesentliches Hilfsmittel bei unseren Untersuchungen ist eine Variante des auf C. G. Broyden, J. E. Dennis Jr. und J. J. Moré zurückgehenden Bounded-Deterioration-Satzes aus [1].

Wir setzen über (P) voraus:

(*V*) Die Zielfunktion f ist auf einer konvexen Umgebung 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.

Als erstes geben wir nun besagten Bounded-Deterioration-Satz für das BFGS-Update an. In der Darstellung folgen wir J. Werner [22] und benutzen die schon genannte Funktion ψ mit $\psi(A) := \operatorname{tr}(A) - \ln\det(A)$ anstelle der früher üblichen gewichteten Frobeniusnorm.

Satz 3.5 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Voraussetzung (V) sei erfüllt. Dann existieren positive Konstanten ϵ und α mit der folgenden Eigenschaft: Sind $x_k, x_{k+1} \in B[x^*;\epsilon] := \{x \in \mathbb{R}^n : \|x - x^*\| \le \epsilon\}$ mit $x_k \neq x_{k+1}$ und $B_k \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit gegeben, so ist*

$$B_{k+1} := B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}$$

mit

$$s_k := x_{k+1} - x_k \quad \textit{und} \quad y_k := \nabla f(x_{k+1}) - \nabla f(x_k)$$

symmetrisch und positiv definit, und es gilt

$$\psi(B^{*-1/2} B_{k+1} B^{*-1/2}) \le \psi(B^{*-1/2} B_k B^{*-1/2}) + \alpha\sigma(x_k, x_{k+1})$$

mit

$$\sigma : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}, \quad \sigma(u,v) := \max(\|u - x^*\|, \|v - x^*\|).$$

Ein Beweis dieser Aussage ist bei J. Werner [22] zu finden.

Mit Hilfe von Satz 3.5 ist es nun möglich, die lokale Q-lineare Konvergenz des ungedämpften BFGS-Verfahrens zu zeigen. Wieder benutzen wir die ψ-Funktion anstelle einer gewichteten Frobeniusnorm.

Satz 3.6 *Gegeben sei die Aufgabe (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 $x_0 \in \mathbb{R}^n$ mit $\|x_0 - x^*\| \leq \epsilon(r)$ und $B_0 \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit mit $\psi(B^{*-1/2} B_0 B^{*-1/2}) - n \leq \delta(r)$, so ist das ungedämpfte BFGS-Verfahren durchführbar und liefert (falls kein vorzeitiger Abbruch stattfindet) 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.

Bemerkung: Durch Inspektion des Beweises des vorangegangenen Satzes bei J. Werner [22] erkennt man, daß die Aussage auch allgemeiner für gewisse Quasi-Newton-Verfahren gilt. Für die Folge der Update-Matrizen $\{B_k\}$ muss lediglich gezeigt werden, daß alle Folgenglieder symmetrisch und positiv definit sind und daß eine Bounded-Deterioration-Eigenschaft gemäß Satz 3.5 vorliegt, also für alle $k = 0, 1, \ldots$

$$\psi(B^{*-1/2} B_{k+1} B^{*-1/2}) \leq \psi(B^{*-1/2} B_k B^{*-1/2}) + \alpha\sigma(x_k, x_{k+1})$$

mit einer Konstanten $\alpha > 0$ gilt.

Nun können wir uns dem zweiten Ziel, dem Nachweis superlinearer Konvergenz widmen. Nach Satz 2.4 genügt es hierzu, die Dennis-Moré-Bedingung

$$\lim_{k\to\infty} \frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} = 0$$

nachzuweisen. Damit sind dann auch noch weitere Folgerungen mittels Satz 2.5 möglich. Zunächst jedoch:

Satz 3.7 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Voraussetzung (V) sei erfüllt. Das gedämpfte oder ungedämpfte BFGS-Verfahren erzeuge eine (o.B.d.A.) nicht vorzeitig abbrechende Folge $\{x_k\}$ mit $\sum_{k=0}^{\infty} \|x_k - x^*\| < \infty$ und eine Folge symmetrischer, positiv definiter Matrizen $\{B_k\} \subset \mathbb{R}^{n\times n}$. Dann sind die Folgen $\{\|B_k\|\}$ und $\{\|B_k^{-1}\|\}$ beschränkt, und es gilt die Dennis-Moré-Bedingung*

$$\lim_{k\to\infty} \frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} = 0.$$

Beweise dieser Aussage sind wiederum bei R. H. Byrd und J. Nocedal [3], sowie J. Werner [22] zu finden. In beiden Arbeiten wird wieder die ψ-Funktion benutzt. Ein früherer Beweis ohne dieses Hilfmittel von J. E. Dennis Jr. und J. J. Moré befindet sich in [7].

Bemerkungen:

- Die Voraussetzung $\sum_{k=0}^{\infty} \|x_k - x^*\| < \infty$ für den obigen Satz ist erfüllt, falls die Folge $\{x_k\}$ R-linear oder Q-linear gegen x^* konvergiert. Nach den bisherigen Ergebnissen ist dies global für das durch eine semi-effiziente Schrittweite gedämpfte und lokal für das ungedämpfte BFGS-Verfahren der Fall.

- Da die Dennis-Moré-Bedingung erfüllt ist, folgt aus dem gleichnamigen Satz 2.4 sofort die lokale Q-superlineare Konvergenz des ungedämpften BFGS-Verfahrens.

- Mit Satz 2.5 folgt aus dem vorangegangenen Satz, daß bei Dämpfung des Verfahrens durch die Wolfe- oder die Armijo-Schrittweite $t_k = 1$ für alle hinreichend großen k gilt. Das gedämpfte Verfahren geht also in das ungedämpfte über, und es folgt Q-superlineare Konvergenz. Bei der Wahl der Schrittweite ist es dabei natürlich wichtig, die Zulässigkeit von $t_k = 1$ als erstes zu testen.

Damit endet dieser Abschnitt, und wir wenden uns dem 2001 veröffentlichten BFGS-Verfahren mit "cautious" Update zu.

3.3 Das "cautious" Update

Nun beschäftigen wir uns mit der in Abschnitt 3.1 schon gestellten und lange Zeit unbeantworteten Frage nach der globalen Konvergenz des gedämpften BFGS-Verfahrens bei nichtkonvexer Zielfunktion. Nach dem Satz 3.3 von Dai existiert eine Zielfunktion, für die das Verfahren nicht konvergiert. Daher darf ohne eine Konvexitätsvoraussetzung nicht mehr von globaler Konvergenz ausgegangen werden. Es ist jedoch möglich, die Update-Formel des BFGS-Verfahrens so zu verändern, daß auch im nichtkonvexen Fall globale Konvergenz gezeigt werden kann. Durch das Gegenbeispiel von Dai erhalten solchen Modifikationen eine wesentlich größere Bedeutung. Daher stellen wir nun Ergebnisse von D.-H. Li und M. Fukushima vor, die für zwei modifizierte BFGS-Verfahren globale Konvergenz bei nichtkonvexer Zielfunktion beweisen konnten. Unser hauptsächliches Interesse wird dabei dem 2001 in [15] vorgestellten BFGS-Verfahren mit *"cautious" Update*, kurz *CBFGS-Verfahren*, gelten. Im einzelnen werden wir ein mit Satz 3.2 vergleichbares globales Konvergenzresultat für das CBFGS-Verfahren

mit einer semi-effizienten Schrittweitenstrategie und einer nichtkonvexen Zielfunktion beweisen und unter erheblichen zusätzlichen Voraussetzungen sogar superlineare Konvergenz zeigen. Zunächst wollen wir jedoch das "cautious" Update motivieren.

Wir haben bereits gesehen, daß die Bedingung $y_k^T s_k > 0$ für das BFGS-Verfahren von großer Bedeutung ist. Sie sichert die Durchführbarkeit des Verfahrens und den Erhalt der positiven Definitheit der Update-Matrizen. Ist keine gleichmäßig konvexe Zielfunktion vorgegeben und wird etwa die Armijo-Schrittweite verwendet, so kann $y_k^T s_k > 0$ für alle k nicht sichergestellt werden. Daher geht man dazu über, B_k nur dann upzudaten, wenn $y_k^T s_k > 0$ ist und ansonsten die alte, positiv definite Matrix noch einmal zu verwenden. Es ist dann

$$B_{k+1} := \begin{cases} B_k - \dfrac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \dfrac{y_k y_k^T}{y_k^T s_k}, & \text{falls } y_k^T s_k > 0, \\ B_k & \text{sonst.} \end{cases}$$

Zur besseren Implementierbarkeit wird die Bedingung $y_k^T s_k > 0$ im obigen Update häufig durch $y_k^T s_k \geq \eta$ mit einer (kleinen) Konstanten $\eta > 0$ ersetzt. Dies ist der Ausganspunkt für das "cautious" Update, bei welchem besagte Bedingung noch etwas verändert wird. Mit vorgegebenen Konstanten $\alpha, \epsilon > 0$ lautet das *CBFGS-Update*

$$B_{k+1} := \begin{cases} B_k - \dfrac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \dfrac{y_k y_k^T}{y_k^T s_k}, & \text{falls } \dfrac{y_k^T s_k}{\|s_k\|^2} > \epsilon \|g_k\|^\alpha, \\ B_k & \text{sonst.} \end{cases}$$

Nun wollen wir die Konvergenzeigenschaften des hieraus entstehenden Quasi-Newton-Verfahrens bei nichtkonvexer Zielfunktion untersuchen. Wir benötigen wieder die folgenden Voraussetzungen:

(V_1) Bei gegebenem Startvektor $x_0 \in \mathbb{R}^n$ des Verfahrens ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n \ : \ f(x) \leq f(x_0)\}$ kompakt.

(V_2) Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar.

(V_3) Der Zielfunktionsgradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig. Es existiert also eine Konstante $\gamma > 0$ mit

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

Desweiteren verwenden wir:

Lemma 3.8 *Sind $I \subset \mathbb{N}_0$ eine endliche Indexmenge sowie $a > 0$ und $\alpha_i \geq 0$ für $i \in I$ Konstanten mit*

$$\sum_{i \in I} \alpha_i \leq a \, \#(I),$$

so gibt es eine Indexmenge $J \subset I$, welche mindestens $\frac{1}{2}\#(I)$ Elemente enthält und für die $\alpha_i \leq 2a$ für alle $i \in J$ gilt.

Dieses Lemma entspricht den Aussagen bei C. Geiger und C. Kanzow [11], S. 170 oder J. Werner [23], S. 205. Die Beweisführung kann analog erfolgen.

Nun können wir die Hauptaussage von D.-H. Li und M. Fukushima aus [15] beweisen. Die Autoren beschränken sich bei ihren Untersuchungen auf Wolfe- und Armijo-Schrittweiten, durch Anwendung der ψ-Funktion analog zu J. Werner in [22] können wir die ursprüngliche Aussage allgemeiner auch für semi-effiziente Schrittweiten beweisen.

Satz 3.9 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Voraussetzungen (V_1) bis (V_3) seien erfüllt. Mit einem $x_0 \in \mathbb{R}^n$ und einer symmetrischen und positiv definiten Matrix $B_0 \in \mathbb{R}^{n \times n}$ erzeuge das gedämpfte CBFGS-Verfahren eine Folge $\{x_k\}$. Ferner gelte (für die gewählte Schrittweitenstrategie) in jedem Iterationsschritt $k = 0, 1, \ldots$*

$$f(x_k) - f(x_k + t_k p_k) \geq \theta \min\left[-\nabla f(x_k)^T p_k, \left(\frac{\nabla f(x_k)^T p_k}{\|p_k\|}\right)^2\right]$$

mit einer Konstanten $\theta > 0$. Dann gilt: Entweder bricht das Verfahren nach endlich vielen Schritten mit einer kritischen Lösung ab, oder für die erzeugte Folge gilt

$$\liminf_{k \to \infty} \|\nabla f(x_k)\| = 0.$$

Beweis: Wir führen einen Widerspruchsbeweis und nehmen daher an, daß es keine gegen Null konvergierende Teilfolge von $\{\|\nabla f(x_k)\|\}$ gibt. Folglich ist die gesamte Folge gegen Null beschränkt, und es existiert eine Konstante $\eta > 0$ derart, daß $\|g_k\| = \|\nabla f(x_k)\| \geq \eta$ für alle $k = 0, 1, \ldots$ gilt. Aus der vorausgesetzten Semi-Effizienz der Schrittweitenstrategie folgern wir hiermit, daß

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \underbrace{\theta \min\left[-\frac{g_k^T p_k}{\|g_k\|^2}, \left(\frac{g_k^T p_k}{\|g_k\|\|p_k\|}\right)^2\right]}_{=:\delta_k > 0} \|g_k\|^2 \\ &\geq \theta \eta^2 \delta_k \end{aligned}$$

für alle $k = 0, 1, \ldots$ gilt. Unser Ziel ist es nun, die Existenz eines $\delta > 0$ mit $\delta_k \geq \delta$ für unendlich viele Indices k zu zeigen. Für diese Indices gilt dann

$$f(x_k) - f(x_{k+1}) \geq \theta \eta^2 \delta > 0,$$

was einen Widerspruch zur Beschränktheit der Zielfunktion auf der Niveaumenge ergeben wird. Um dies zu erreichen, definieren wir zunächst die Menge

$$J := \left\{ k \in \mathbb{N}_0 \ : \ \frac{y_k^T s_k}{\|s_k\|^2} \geq \epsilon \|g_k\|^\alpha \right\}$$

derjenigen Iterationsindices, bei denen im CBFGS-Verfahren ein BFGS-Update ausgeführt wird und unterscheiden zwei Fälle.

1. Fall: Die Indexmenge J ist endlich. Daher ist die Folge $\{B_k\}$ ab einem gewissen Iterationsschritt konstant, beziehungsweise $B_k = B$ mit einer symmetrischen und positiv definiten Matrix $B \in \mathbb{R}^{n \times n}$ für fast alle k. Dann ist wegen $B_k p_k = -g_k$ und $s_k = t_k p_k$

$$\begin{aligned}
\delta_k &= \min\left[\frac{p_k^T B_k p_k}{\|B_k p_k\|^2}, \left(\frac{p_k^T B_k p_k}{\|p_k\| \|B_k p_k\|} \right)^2 \right] = \min\left[\frac{s_k^T B_k s_k}{\|B_k s_k\|^2}, \left(\frac{s_k^T B_k s_k}{\|s_k\| \|B_k s_k\|} \right)^2 \right] \\
&= \min\left[\frac{\|B_k^{1/2} s_k\|^2}{\|B_k^{1/2} B_k^{1/2} s_k\|^2}, \frac{\|B_k^{1/2} s_k\|^4}{\|B_k^{-1/2} B_k^{1/2} s_k\|^2 \|B_k^{1/2} B_k^{1/2} s_k\|^2} \right] \\
&\geq \min\left[\frac{1}{\|B_k^{1/2}\|^2}, \frac{1}{\|B_k^{-1/2}\|^2 \|B_k^{1/2}\|^2} \right] = \min\left[\frac{1}{\lambda_{\max}(B_k)}, \frac{\lambda_{\min}(B_k)}{\lambda_{\max}(B_k)} \right].
\end{aligned}$$

Hierbei bezeichnen $\lambda_{\min}(B_k)$ den kleinsten und $\lambda_{\max}(B_k)$ den größten Eigenwert von B_k. Da J endlich ist, gilt

$$\delta_k \geq \min\left[\frac{1}{\lambda_{\max}(B)}, \frac{\lambda_{\min}(B)}{\lambda_{\max}(B)} \right] =: \delta > 0$$

für fast alle k. Die δ_k sind also für fast alle k gegen Null beschränkt. Folglich wird bei einem nicht vorzeitig abbrechenden CBFGS-Verfahren in fast allen Iterationsschritten eine positive, gegen Null beschränkte Zielfunktionsverminderung erreicht. In den endlich vielen restlichen Schritten findet wegen der vorausgesetzten Semi-Effizienz der Schrittweiten keine Vergrößerung statt. Dies ist der gesuchte Widerspruch zur Beschränktheit der Zielfunktion auf L_0. Wir kommen zum

2. Fall: Die Indexmenge J ist nicht endlich. Für die Indexmengen

$$J_k := J \cap \{0, 1, \ldots, k\}$$

gilt dann $\lim_{k \to \infty} \#(J_k) = \infty$. Nun betrachten wir die schon häufig erwähnte ψ-Funktion von R. H. Byrd und J. Nocedal. Sie wird auf der Menge der symmetrischen und positiv definiten Matrizen durch

$$\psi(A) := \operatorname{tr}(A) - \ln \det(A)$$

definiert. Da die Spur einer Matrix A gleich der Summe und die Determinate gleich dem Produkt ihrer Eigenwerte $\lambda_1(A), \ldots, \lambda_n(A)$ ist, gilt allgemein

$$\psi(A) = \sum_{i=1}^{n} \underbrace{\lambda_i(A) - \ln \lambda_i(A)}_{\geq 1} \geq n.$$

Hierbei haben wir die für alle $t > 0$ gültige Ungleichung

$$1 + \ln t \leq t$$

benutzt. Als nächstes werden wir $\psi(B_{k+1})$ ausrechnen und abschätzen. Für $k \notin J$ gilt $B_{k+1} = B_k$ und somit auch $\psi(B_{k+1}) = \psi(B_k)$. Ist dagegen $k \in J$, so wird ein BFGS-Update ausgeführt, und es gilt

$$B_{k+1} = B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}.$$

Da dies ein Broydenklasse-Update mit dem Parameter $\phi_k = 1$ ist, gilt nach Satz 2.1 für $k \in J$

$$\det(B_{k+1}) = \frac{y_k^T s_k}{s_k^T B_k s_k} \det(B_k).$$

Wegen $\operatorname{tr}(uv^T) = u^T v$ für alle $u, v \in \mathbb{R}^n$ erhalten wir für $k \in J$

$$\begin{aligned}
\psi(B_{k+1}) &= \psi(B_k) - \frac{\|B_k s_k\|^2}{s_k^T B_k s_k} + \frac{\|y_k\|^2}{y_k^T s_k} - \ln \frac{y_k^T s_k}{s_k^T B_k s_k} \\
&= \psi(B_k) + \ln\left(\frac{s_k^T B_k s_k}{\|s_k\| \|B_k s_k\|}\right)^2 + \left[1 - \frac{\|B_k s_k\|^2}{s_k^T B_k s_k} + \ln \frac{\|B_k s_k\|^2}{s_k^T B_k s_k}\right] \\
&\quad + \left[\frac{\|y_k\|^2}{y_k^T s_k} - 1 - \ln \frac{y_k^T s_k}{\|s_k\|^2}\right].
\end{aligned}$$

Insgesamt erhalten wir hiermit für alle $k = 0, 1, \ldots$

$$\begin{aligned}
\psi(B_{k+1}) &= \psi(B_0) + \sum_{j \in J_k} \left\{ \ln\left(\frac{s_j^T B_j s_j}{\|s_j\| \|B_j s_j\|}\right)^2 + \left[1 - \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} + \ln \frac{\|B_j s_j\|^2}{s_j^T B_j s_j}\right] \right\} \\
&\quad + \sum_{j \in J_k} \left[\frac{\|y_j\|^2}{y_j^T s_j} - 1 - \ln \frac{y_j^T s_j}{\|s_j\|^2}\right].
\end{aligned}$$

Wir wollen nun das Argument der letzten der beiden Summen gegen eine Konstante abschätzen. Für $j \in J_k$ gilt nach unserer Annahme

$$\frac{y_j^T s_j}{\|s_j\|^2} \geq \epsilon \|g_j\|^\alpha \geq \epsilon \eta^\alpha$$

und daher mit (V_3)

$$\frac{\|y_j\|^2}{y_j^T s_j} \leq \frac{\gamma^2 \|s_j\|^2}{y_j^T s_j} \leq \frac{\gamma^2}{\epsilon \eta^\alpha}.$$

Folglich ist

$$\sum_{j \in J_k} \left[\frac{\|y_j\|^2}{y_j^T s_j} - 1 - \ln \frac{y_j^T s_j}{\|s_j\|^2} \right] \leq \#(J_k) \left[\frac{\gamma^2}{\epsilon \eta^\alpha} - 1 - \ln \epsilon \eta^\alpha \right]$$

und somit

$$\begin{aligned} 0 < \psi(B_{k+1}) \quad \leq \quad & \sum_{j \in J_k} \left\{ \ln \left(\frac{s_j^T B_j s_j}{\|s_j\| \|B_j s_j\|} \right)^2 + \left[1 - \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} + \ln \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} \right] \right\} \\ & + C \#(J_k) \end{aligned}$$

mit einer von k unabhängigen Konstanten $C > 0$. Wir formen kurz um zu

$$\sum_{j \in J_k} \left\{ \ln \left(\frac{\|s_j\| \|B_j s_j\|}{s_j^T B_j s_j} \right)^2 + \left[\frac{\|B_j s_j\|^2}{s_j^T B_j s_j} - 1 - \ln \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} \right] \right\} \leq C \#(J_k)$$

und wenden Lemma 3.8 an. Die Voraussetzungen hierfür sind erfüllt, denn beide Summanden sind nichtnegativ. Dies ist beim ersten Summanden durch Anwenden der Cauchy-Schwarz'schen Ungleichung und beim zweiten durch Benutzung von $t - 1 - \ln t \geq 0$ für positive t einzusehen. Es existiert also eine Indexmenge ${J_k}^* \subset J_k$ mit $\#({J_k}^*) \geq \frac{1}{2} \#(J_k)$ derart, daß für alle $j \in J_k^*$

$$\ln \left(\frac{\|s_j\| \|B_j s_j\|}{s_j^T B_j s_j} \right)^2 + \left[\frac{\|B_j s_j\|^2}{s_j^T B_j s_j} - 1 - \ln \frac{\|B_j s_j\|^2}{s_j^T B_j s_j} \right] \leq 2C$$

gilt. Wegen Nichtnegativität sind beide Summanden durch $2C$ beschränkt. Daher ist für alle $j \in {J_k}^*$

$$\left(\frac{\|s_j\| \|B_j s_j\|}{s_j^T B_j s_j} \right)^2 \leq e^{2C}.$$

Nun zeigen wir, daß auch $\|B_j s_j\|^2 / s_j^T B_j s_j$ beschränkt ist. Dazu definieren wir die Abbildung $h : (0, \infty) \longrightarrow \mathbb{R}$ durch $h(t) := t - 1 - \ln t$. Durch Inspektion der ersten beiden Ableitungen von h sieht man, daß h strikt konvex im gesamten Definitionsbereich, streng monoton fallend auf dem Intervall $(0, 1]$ und streng monoton wachsend auf auf $[1, \infty)$ ist. Außerdem ist $h(1) = 0$ und

$$\lim_{h \to 0+} h(t) = \lim_{h \to \infty} = \infty.$$

Daher existieren eindeutig bestimmte Konstanten $C_1 \in (0, 1)$ und $C_2 \in (1, \infty)$ mit $h(C_1) = h(C_2) = 2C$. Folglich gilt für alle $t > 0$ mit $h(t) \leq 2C$ dann $t \in [C_1, C_2]$ und damit

$$\frac{\|B_j s_j\|^2}{s_j^T B_j s_j} \leq C_2$$

für alle $j \in J_k{}^*$. Also ist für alle $j \in J_k{}^*$

$$\delta_j = \min\left[\frac{s_j^T B_j s_j}{\|B_j s_j\|^2}, \left(\frac{s_j^T B_j s_j}{\|s_j\|\|B_j s_j\|}\right)^2\right] \geq \min\left(C_2^{-1}, e^{-2C}\right) =: \delta > 0.$$

Somit wird in all den Iterationschritten mit einem Iterationsindex $k \in J_k{}^*$ eine positive gegen Null beschränkte Zielfunktionsverminderung erreicht. Da nun $J_k{}^* \subset J_k$ ist und $\#(J_k{}^*) \geq \frac{1}{2}\#(J_k) \to \infty$ gilt, wird dies bei einem nicht vorzeitig abbrechenden Verfahren in unendlich vielen Iterationsschritten geschehen. In den restlichen Iterationsschritten findet wegen $\delta_k > 0$ keine Zielfunktionsvergrößerung statt. Folglich kann die Zielfunktion nicht auf der Niveaumenge nach unten beschränkt sein. Formal hat das die folgende Gestalt:

$$\begin{aligned} f(x_0) - f(x_{k+1}) &= \sum_{j=0}^{k} f(x_j) - f(x_{j+1}) \geq \sum_{j=0}^{k} \theta\eta^2\delta_j \geq \theta\eta^2 \sum_{j\in J_k^*} \delta_j \\ &\geq \theta\eta^2\delta\#(J_k^*) \geq \frac{\theta\eta^2\delta}{2}\#(J_k) \longrightarrow \infty. \end{aligned}$$

Dies ist der gesuchte Widerspruch. □

Nach obigem Satz existiert also zumindest eine Teilfolge von $\{x_k\}$, die gegen einen stationären Punkt von f konvergiert. Ob es sich hierbei um ein lokales oder gar globales Minimum, beziehungsweise eine entsprechende Lösung von (P) handelt, kann aufgrund der geringen Voraussetzungen a priori nicht entschieden werden. Setzt man jedoch die Zielfunktion als konvex voraus, so ergibt sich das folgende Korollar.

Korollar 3.10 *Gegeben sei die unrestringierte Aufgabe (P). Die Voraussetzungen (V_1) bis (V_3) seien erfüllt. Mit $x_0 \in \mathbb{R}^n$ und $B_0 \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit startend sei die Folge $\{x_k\}$ durch das CBFGS-Verfahren mit semi-effizienter Schrittweite erzeugt. Dann gilt: Ist die Zielfunktion f konvex, so ist jeder Häufungspunkt von $\{x_k\}$ eine globale Lösung von (P).*

Falls nun f aber nicht konvex ist, und davon wollten wir ja ursprünglich ausgehen, haben wir also nur Konvergenz einer Teilfolge gegen einen stationären Punkt. Mit erheblichen Zusatzvoraussetzungen "beweisen" D.-H. Li und M. Fukushima auch hier Konvergenz der ganzen Folge $\{x_k\}$ gegen ein lokales Minimum von f. Als wesentlich wichtiger zu bewerten ist, daß unter denselben Voraussetzungen der Übergang des CBFGS-Verfahrens in das BFGS-Verfahren nach endlich vielen Iterationsschritten folgt. Wir formalisieren dies im folgenden Satz.

Satz 3.11 *Gegeben sei die unrestringerte Optimierungsaufgabe (P). Die Voraussetzungen (V_1) bis (V_3) seien erfüllt. Mit einem $x_0 \in \mathbb{R}^n$ und $B_0 \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit erzeuge das durch eine semi-effiziente Schrittweitenstrategie gedämpfte CBFGS-Verfahren eine Folge $\{x_k\}$. Ferner sei die Zielfunktion f*

zweimal stetig differenzierbar und es gelte $s_k \to 0$. Dann gilt: Falls ein Häufungspunkt $x^* \in \mathbb{R}^n$ von $\{x_k\}$ mit $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ symmetrisch und positiv definit existiert, dann konvergiert $\{x_k\}$ gegen x^*, und das CBFGS-Verfahren geht nach endlich vielen Iterationsschritten in das BFGS-Verfahren über.

Beweis: Die erste Aussage ist trivial, da aus $s_k \to 0$ schon Konvergenz von $\{x_k\}$ folgt. Zum Nachweis der zweiten Aussage definieren wir

$$A_k := \int_0^1 \nabla^2 f(x_k + \tau s_k) d\tau.$$

Wegen der Konvergenzen $x_k \to x^*$, $s_k \to 0$ folgt $A_k \to \nabla^2 f(x^*)$ und somit aufgrund der positiven Definitheit von $\nabla^2 f(x^*)$ die gleichmäßige positive Definitheit der Matrizen A_k für alle hinreichend großen k. Es existieren also Konstanten $m, M > 0$ mit

$$m\|s_k\|^2 \leq s_k^T A_k s_k \leq M\|s_k\|^2$$

für alle hinreichend großen k. Eine Formulierung dieser Aussage in Form eines Satzes mit Beweis gibt es zum Beispiel in [11], S. 98. Durch den Mittelwertsatz in Integralform erhalten wir

$$A_k s_k = \int_0^1 \nabla^2 f(x_k + \tau(x_{k+1} - x_k))(x_{k+1} - x_k) d\tau = \nabla f(x_{k+1}) - \nabla f(x_k) = y_k.$$

Für hinreichend große k gilt also

$$y_k^T s_k = s_k^T A_k s_k \geq m\|s_k\|^2$$

und daher wegen $\|g_k\| \to 0$

$$\frac{y_k^T s_k}{\|s_k\|^2} \geq m \geq \epsilon \|g_k\|^\alpha.$$

Für alle hinreichend großen k wird also B_{k+1} immer durch ein BFGS-Update bestimmt. □

Die erste Aussage des obigen Satzes wird in [15] zwar als eine Art Konvergenzresultat angekündigt, es sollte jedoch klar sein, daß die Konvergenz von $\{x_k\}$, beziehungsweise $s_k \to 0$ ja schon Voraussetzung des Satzes ist. Von größerer Bedeutung ist also die zweite Aussage, denn mit obigem Satz übertragen sich im Falle der Konvergenz gegen ein isoliertes lokales Minimum die Konvergenzeigenschaften des BFGS-Verfahrens auf das CBFGS-Verfahren. Falls also zusätzlich die Voraussetzung (V) aus Abschnitt 3.2 erfüllt ist, gilt auch hier die Dennis-Moré-Bedingung und es folgen (zumindest bei Verwendung von Wolfe- oder Armijo-Schrittweiten) der Übergang vom gedämpften in ein ungedämpftes Verfahren und Q-superlineare Konvergenz.

In der zugrundeliegenden Arbeit [15] werden auch umfangreiche numerische Experimente mit dem CBFGS-Verfahren durchgeführt. Für die Wahl der Parameter wird $\epsilon = 10^{-6}$ und $\alpha \in \{0.001, 3\}$ vorgegeben. In den zahlreichen Testergebnissen erkennt man, daß in fast allen Fällen die Bedingung für ein BFGS-Update erfüllt ist. Ferner wird bemerkt, daß die Wahl von α die Konvergenzgeschwindigkeit entscheident beeinflussen kann. Es wird empfohlen, für große Werte von $\|g_k\|$ den Parameter $\alpha = 0.001$ und für kleine Werte $\alpha = 3$ zu wählen. Mit dieser Wahl ist das CBFGS-Verfahren dann noch näher am BFGS-Verfahren. Zum Abschluss dieses Abschnitts weisen wir noch kurz auf ein anderes verändertes BFGS-Verfahren und dessen Konvergenz bei nichtkonvexer Zielfunktion hin.

In der etwas früheren Arbeit [14] stellen D.-H. Li und M. Fukushima ein weiteres modifiziertes BFGS-Verfahren, das *MBFGS-Verfahren*, vor. Hierbei wird in der Update-Formel für die BFGS-Matrix nicht wie üblich $y_k := g_{k+1} - g_k$, sondern

$$y_k := (g_{k+1} - g_k) + r_k s_k$$

mit $r_k \in [0, C]$ und einer Konstanten $C > 0$ gesetzt. Auch für dieses Verfahren gilt ohne Konvexitätsvoraussetzung bei beschränkter Niveaumenge und Verwendung von Wolfe-Schrittweiten

$$\liminf_{k \to \infty} \|g_k\| = 0.$$

Mit einigen weiteren Voraussetzungen kann auch superlineare Konvergenz gezeigt werden. Bei Verwendung von Armijo-Schrittweiten ist eine andere Modifikation mit den gleichen Konvergenzeigenschaften möglich. Auf nähere Einzelheiten wollen wir jedoch nicht eingehen. Damit beenden wir die Ausführungen zu den theoretischen Eigenschaften des BFGS-Verfahrens und wenden uns dem Dennis-Wolkowicz-Verfahren zu.

Kapitel 4

Das Dennis-Wolkowicz-Verfahren

In diesem Kapitel werden wir weitgehend analog zu den Ausführungen zum BFGS-Verfahren im vorangegangenen Kapitel die theoretischen Eigenschaften des Dennis-Wolkowicz-Verfahrens untersuchen. Nach einer kurzen Einführung werden wir einige bereits bekannte aber auch einige neue globale und lokale Konvergenzresultate zu diesem Verfahren beweisen. Wir betrachten im folgenden wieder die unrestringierte Optimierungsaufgabe

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

mit einer stetig differenzierbaren Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$.

Das Dennis-Wolkowicz-Verfahren wurde von den beiden Namensgebern J. E. Dennis Jr. und H. Wolkowicz 1993 in [9] vorgestellt und ist ein Quasi-Newton-Verfahren mit zweistufigem Matrix-Update. In der ersten Stufe wird ein *inverses schwaches Greenstadt-Update*, in der zweiten Stufe ein BFGS-Update ausgeführt. Das inverse schwache Greenstadt-Update B_{k+1}^{-1} von B_k^{-1} ist die eindeutige Lösung der restringierten Optimierungsaufgabe

$$\text{Minimiere} \quad \|W^T(H - B_k^{-1})W\|_F \quad \text{auf} \quad \{H \in \mathbb{R}^{n\times n} \ : \ y_k^T H y_k = y_k^T s_k\}.$$

Hierbei ist $W \in \mathbb{R}^{n\times n}$ eine Matrix mit $B_k = WW^T$ und $\|\cdot\|_F$ die Frobeniusnorm. Die Lösung ist gemäß [9] gegeben durch

$$B_{k+1}^{-1} := B_k^{-1} + \frac{y_k^T s_k - y_k^T B_k^{-1} y_k}{(y_k^T B_k^{-1} y_k)^2} B_k^{-1} y_k y_k^T B_k^{-1}.$$

Hieraus berechnet sich eine direkte Update Formel wie folgt:

$$B_{k+1} = B_k + \frac{y_k^T B_k^{-1} y_k - y_k^T s_k}{(y_k^T B_k^{-1} y_k)(y_k^T s_k)} y_k y_k^T.$$

Genau wie bei einigen Update-Formeln der Broydenklasse gilt auch für das inverse schwache Greestadt-Update, daß unter der Bedingung $y_k^T s_k > 0$ mit der

aktuellen Matrix B_k auch B_{k+1} symmetrisch und positiv definit ist. Ein Beweis dieser Aussage ist wiederum bei Dennis und Wolkowicz in [9] zu finden. Mit den Abkürzungen

$$a_k := y_k^T B_k^{-1} y_k, \qquad b_k := y_k^T s_k, \qquad c_k := s_k^T B_k s_k$$

ist das zweistufige *Dennis-Wolkowicz-Update* dann gegeben durch

$$\begin{aligned} B_{k+\frac{1}{2}} &:= B_k + \frac{a_k - b_k}{a_k}\frac{y_k y_k^T}{b_k}, \\ B_{k+1} &:= B_{k+\frac{1}{2}} - \frac{(B_{k+\frac{1}{2}} s_k)(B_{k+\frac{1}{2}} s_k)^T}{s_k^T B_{k+\frac{1}{2}} s_k} + \frac{y_k y_k^T}{b_k}. \end{aligned}$$

Es ist allerdings auch möglich, dieses zweistufige Update als ein einstufiges darzustellen. Dieses gehört sogar zur Broydenklasse und hat den Broydenklasse-Parameter

$$\phi_k = \frac{1}{b_k/c_k + 1 - b_k^2/a_k c_k}.$$

Da Dennis und Wolkowicz diesen für das direkte Update lediglich angeben und nicht ausrechnen, vergewissern wir uns noch einmal selbst. Zunächst gilt

$$\begin{aligned} s_k^T B_{k+\frac{1}{2}} s_k &= s_k^T \left(B_k + \frac{a_k - b_k}{a_k}\frac{y_k y_k^T}{b_k} \right) s_k = c_k + \frac{a_k - b_k}{a_k} b_k \\ &= c_k \left(1 + \frac{a_k - b_k}{a_k}\frac{b_k}{c_k} \right) = c_k \left(1 + \frac{b_k}{c_k} - \frac{b_k^2}{a_k c_k} \right). \end{aligned}$$

Hiermit können wir nun die zweistufige Update-Formel des DW-Verfahrens in die Form eines Broydenklasse-Updates überführen. Aus Platzgründen werden wir hierbei einige Zwischenschritte nur verkürzt darstellen können. Die betroffenen Rechnungen lassen sich jedoch mit einem mathematischen Anwendersystem sehr leicht verifizieren. Es gilt

$$\begin{aligned} B_{k+1} &= B_{k+\frac{1}{2}} - \frac{(B_{k+\frac{1}{2}} s_k)(B_{k+\frac{1}{2}} s_k)^T}{s_k^T B_{k+\frac{1}{2}} s_k} + \frac{y_k y_k^T}{b_k} = B_k + \left(\frac{a_k - b_k}{a_k} + 1 \right) \frac{y_k y_k^T}{b_k} \\ &\quad - \frac{\left(B_k s_k + ((a_k - b_k)/a_k) y_k\right) \left(B_k s_k + ((a_k - b_k)/a_k) y_k\right)^T}{c_k (b_k/c_k + 1 - b_k^2/a_k c_k)} \\ &= B_k + \left(\frac{a_k - b_k}{a_k} + 1 \right) \frac{y_k y_k^T}{b_k} - \frac{(B_k s_k)(B_k s_k)^T}{c_k (b_k/c_k + 1 - b_k^2/a_k c_k)} \\ &\quad - \frac{(a_k - b_k)\left(B_k s_k y_k^T + y_k (B_k s_k)^T \right)}{a_k c_k (b_k/c_k + 1 - b_k^2/a_k c_k)} - \frac{(a_k - b_k)^2 y_k y_k^T}{a_k^2 c_k (b_k/c_k + 1 - b_k^2/a_k c_k)} \end{aligned}$$

$$
\begin{aligned}
&= B_k - \frac{(B_k s_k)(B_k s_k)^T}{c_k} + \frac{y_k y_k^T}{b_k} + \frac{(a_k - b_k)b_k}{c_k(a_k b_k + a_k c_k - b_k^2)}(B_k s_k)(B_k s_k)^T \\
&\quad + \frac{b_k - a_k}{a_k b_k + a_k c_k - b_k^2}\left(B_k s_k y_k^T + y_k (B_k s_k^T)\right) + \frac{(a_k - b_k)c_k}{b_k(a_k b_k + a_k c_k - b_k^2)} y_k y_k^T \\
&= B_k - \frac{(B_k s_k)(B_k s_k)^T}{c_k} + \frac{y_k y_k^T}{b_k} + c_k \left(1 - \frac{1}{b_k/c_k + 1 - b_k^2/a_k c_k}\right) \\
&\quad \cdot \left(\frac{(B_k s_k)(B_k s_k)^T}{c_k^2} - \frac{B_k s_k y_k^T}{b_k c_k} - \frac{y_k (B_k s_k)^T}{b_k c_k} + \frac{y_k y_k^T}{b_k^2}\right) \\
&= B_k - \frac{(B_k s_k)(B_k s_k)^T}{c_k} + \frac{y_k y_k^T}{b_k} \\
&\quad + c_k \left(1 - \frac{1}{b_k/c_k + 1 - b_k^2/a_k c_k}\right)\left(\frac{y_k}{b_k} - \frac{B_k s_k}{c_k}\right)\left(\frac{y_k}{b_k} - \frac{B_k s_k}{c_k}\right)^T \\
&= B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} + s_k^T B_k s_k (1 - \phi_k) v_k v_k^T
\end{aligned}
$$

mit

$$
\phi_k := \frac{1}{b_k/c_k + 1 - b_k^2/a_k c_k} \quad \text{und} \quad v_k := \frac{y_k}{y_k^T s_k} - \frac{B_k s_k}{s_k^T B_k s_k}.
$$

Folglich gehört auch das Dennis-Wolkowicz-Verfahren zur Broydenklasse, und es gelten die Aussagen von Satz 2.1. Wir wollen jetzt ϕ_k nach oben und unten abschätzen und benötigen dafür

$$
b_k^2 = (y_k^T B_k^{-1/2} B_k^{1/2} s_k)^2 \leq \|B_k^{-1/2} y_k\|^2 \|B_k^{1/2} s_k\|^2 = y_k^T B_k^{-1} y_k s_k^T B_k s_k = a_k c_k.
$$

Damit ist dann

$$
\frac{c_k}{b_k} \geq \phi_k \geq \frac{1}{b_k/c_k + 1}.
$$

Mit $b_k, c_k > 0$ ist dann auch $\phi_k > 0$. Dies kann durch Verwendung der Wolfe- oder der exakten Schrittweite oder einer gleichmäßig konvexen Zielfunktion sichergestellt werden. Allerdings ist $\phi_k > 1$ möglich, weshalb das DW-Update nicht zur konvexen Broydenklasse gehört. Dennoch vererbt sich die positive Definitheit der Update-Matrizen unter der Bedingung $b_k > 0$, da dies für jede einzelne der beiden Stufen des DW-Updates der Fall ist.

Wie in der Einleitung schon erwähnt, wird vor dem ersten Dennis-Wolkowicz-Update ein initial inverse sizing der Form

$$
B_0^{-1} := \frac{y_0^T s_0}{y_0^T B_0^{-1} y_0} B_0^{-1},
$$

beziehungsweise in der direkten Darstellung $B_0 := (b_0/a_0)B_0$, ausgeführt. Da dies ausschließlich zur Verbesserung der numerischen Eigenschaften dient und keinen

Einfluß auf die theoretischen hat, berücksichtigen wir es im weiteren Verlauf dieses Kapitels nicht mehr. Nun werden wir die Konvergenzeigenschaften näher betrachten, zunächst die globalen.

4.1 Globale Konvergenzaussagen

Die globale Konvergenz des Dennis-Wolkowicz-Verfahrens bei gleichmäßig konvexer Zielfunktion unter Verwendung der Wolfe-Schrittweite konnte erstmals 1998 von L. Han und G. Liu in [12] bewiesen werden. Das Hauptergebnis dieser Arbeit ist ein mit Satz 3.2 vergleichbares Konvergenzresultat ohne Angabe einer Konvergenzrate. Durch Anwendung der ψ-Funktion konnte jedoch im selben Jahr die globale R-lineare Konvergenz bei gleichmäßig konvexer Zielfunktion und effizienter Schrittweitenstrategie von J. Werner in [21] bewiesen werden. Wir werden dieses Ergebnis noch einmal ausführlich darstellen und sogar semi-effiziente Schrittweiten zulassen. Zunächst jedoch die üblichen Voraussetzungen:

(K_1) Bei gegebenem Startvektor $x_0 \in \mathbb{R}^n$ des Verfahrens ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n \ : \ f(x) \leq f(x_0)\}$ konvex.

(K_2) Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar und gleichmäßig konvex auf L_0. Dies bedeutet die Existenz einer Konstanten $c > 0$ mit

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

(K_3) Der Zielfunktionsgradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig. Es existiert also eine Konstante $\gamma > 0$ mit

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

Nun folgt der globale Konvergenzsatz für das gedämpfte Dennis-Wolkowicz-Verfahren.

Satz 4.1 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Voraussetzungen (K_1) bis (K_3) seien erfüllt. Mit einem $x_0 \in \mathbb{R}^n$ und einer symmetrischen und positiv definiten Matrix $B_0 \in \mathbb{R}^{n\times n}$ sei die Folge $\{x_k\}$ durch das gedämpfte Dennis-Wolkowicz-Verfahren erzeugt. Ferner gelte (für die gewählte Schrittweitenstrategie) in jedem Iterationsschritt $k = 0, 1, \ldots$*

$$f(x_k) - f(x_k + t_k p_k) \geq \theta \min\left[-\nabla f(x_k)^T p_k, \left(\frac{\nabla f(x_k)^T p_k}{\|p_k\|}\right)^2\right]$$

mit einer Konstanten $\theta > 0$. Dann gilt: Entweder bricht das Verfahren nach endlich vielen Schritten mit der Lösung x^* von (P) in L_0 ab, oder die erzeugte Folge $\{x_k\}$ konvergiert R-linear gegen x^*.

Beweis: Zunächst untersuchen wir die Durchführbarkeit des Verfahrens. Ist $g_k \neq 0$ und die Matrix B_k symmetrisch und positiv definit, so ist p_k eine Abstiegsrichtung und somit $s_k \neq 0$. Wegen (K_2) ist dann

$$\begin{aligned} y_k^T s_k &= -\nabla f(x_{k+1})(x_k - x_{k+1}) - \nabla f(x_k)(x_{k+1} - x_k) \\ &\geq c\|x_{k+1} - x_k\|^2 = c\|s_k\|^2 > 0 \end{aligned}$$

und damit auch $B_{k+\frac{1}{2}}$ und B_{k+1} positiv definit. Das Verfahren ist also durchführbar. Wir nehmen nun an, es würde kein vorzeitiger Abbruch stattfinden und nutzen zum Nachweis der R-linearen Konvergenz Satz 2.3. Zu zeigen ist also die Existenz einer Konstanten $\delta > 0$ mit

$$\delta(k+1) \leq \sum_{j=0}^{k} \min\left[-\frac{g_j^T p_j}{\|g_j\|^2}, \left(\frac{g_j^T p_j}{\|g_j\|\|p_j\|}\right)^2\right]$$

für alle k. Unter den getroffenen Voraussetzungen folgt dann die Behauptung. Wie im Beweis von Satz 3.9 folgen wir der Methode von R. H. Byrd und J. Nocedal aus [3] und definieren die Abbildung ψ auf der Menge der symmetrischen und positiv definiten Matrizen durch

$$\psi(A) := \operatorname{tr}(A) - \ln\det(A).$$

Wegen $1 + \ln t \leq t$ für positive t gilt wieder

$$\psi(A) = \sum_{i=1}^{n} \underbrace{\lambda_i(A) - \ln \lambda_i(A)}_{\geq 1} \geq n,$$

wobei mit $\lambda_1(A), \ldots, \lambda_n(A)$ die Eigenwerte von A bezeichnet seien. Als nächstes werden wir $\psi(B_{k+1})$ ausrechnen und abschätzen. Zuvor sei noch daran erinnert, daß für $u, v \in \mathbb{R}^n$ die Gleichung $\operatorname{tr}(uv^T) = v^T u$ gilt. Für die Spur von B_{k+1} gilt dann

$$\begin{aligned} \operatorname{tr}(B_{k+1}) &= \operatorname{tr}(B_{k+\frac{1}{2}}) - \frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k} + \frac{\|y_k\|^2}{b_k} \\ &= \operatorname{tr}(B_k) + \frac{a_k - b_k}{a_k}\frac{\|y_k\|^2}{b_k} - \frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k} + \frac{\|y_k\|^2}{b_k} \\ &= \operatorname{tr}(B_k) + \left[1 + \frac{a_k - b_k}{a_k}\right]\frac{\|y_k\|^2}{b_k} - \frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k}. \end{aligned}$$

Zur Berechnung der Determinate verwenden wir die Determinantenformel für Broydenklasse-Updates aus Satz 2.1 und erhalten

$$\begin{aligned}
\det(B_{k+1}) &= \det(B_k)\left[\phi_k \frac{y_k^T s_k}{s_k^T B_k s_k} + (1-\phi_k)\frac{y_k^T B_k^{-1} y_k}{y_k^T s_k}\right] \\
&= \det(B_k)\left[\phi_k \frac{b_k}{c_k} + (1-\phi_k)\frac{a_k}{b_k}\right] \\
&= \det(B_k)\frac{\phi_k a_k}{c_k}\left[\frac{b_k}{a_k} + \left(\frac{1}{\phi_k} - 1\right)\frac{c_k}{b_k}\right] \\
&= \det(B_k)\frac{\phi_k a_k}{c_k}\left[\frac{b_k}{a_k} + \left(\frac{b_k}{c_k} - \frac{b_k^2}{a_k c_k}\right)\frac{c_k}{b_k}\right] \\
&= \det(B_k)\frac{\phi_k a_k}{c_k}.
\end{aligned}$$

Zusammen ergibt sich

$$\begin{aligned}
\psi(B_{k+1}) &= \psi(B_k) + \left[1 + \frac{a_k - b_k}{a_k}\right]\frac{\|y_k\|^2}{b_k} - \frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k} - \ln\frac{\phi_k a_k}{c_k} \\
&= \psi(B_k) + \ln\left(\frac{s_k^T B_{k+\frac{1}{2}} s_k}{\|s_k\|\|B_{k+\frac{1}{2}} s_k\|}\right)^2 + \left[1 - \frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k} + \ln\frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k}\right] \\
&\quad + \left[1 + \frac{a_k - b_k}{a_k}\right]\frac{\|y_k\|^2}{b_k} - 1 - \ln\frac{\phi_k a_k}{c_k} - \ln\frac{s_k^T B_{k+\frac{1}{2}} s_k}{\|s_k\|^2}.
\end{aligned}$$

Nun fassen wir die letzten beiden Terme zusammen und vereinfachen sie. Wie wir zu Beginn dieses Abschnitts schon gesehen haben, gilt

$$s_k^T B_{k+\frac{1}{2}} s_k = c_k\left(1 + \frac{b_k}{c_k} - \frac{b_k^2}{a_k c_k}\right) = \frac{c_k}{\phi_k}$$

und damit

$$\begin{aligned}
\psi(B_{k+1}) &= \psi(B_k) + \ln\left(\frac{s_k^T B_{k+\frac{1}{2}} s_k}{\|s_k\|\|B_{k+\frac{1}{2}} s_k\|}\right)^2 + \left[1 - \frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k} + \ln\frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k}\right] \\
&\quad + \left[1 + \frac{a_k - b_k}{a_k}\right]\frac{\|y_k\|^2}{b_k} - 1 - \ln\frac{a_k}{\|s_k\|^2} \\
&= \psi(B_k) + \ln\left(\frac{s_k^T B_{k+\frac{1}{2}} s_k}{\|s_k\|\|B_{k+\frac{1}{2}} s_k\|}\right)^2 + \left[1 - \frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k} + \ln\frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k}\right] \\
&\quad + 2\frac{\|y_k\|^2}{b_k} - \frac{\|y_k\|^2}{a_k} - 1 + \ln\left(\frac{\|s_k\|^2}{\|y_k\|^2}\frac{\|y_k\|^2}{a_k}\right)
\end{aligned}$$

$$
\begin{aligned}
= \quad & \psi(B_k) + \ln\left(\frac{s_k^T B_{k+\frac{1}{2}} s_k}{\|s_k\|\|B_{k+\frac{1}{2}} s_k\|}\right)^2 + \left[1 - \frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k} + \ln \frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k}\right] \\
& + \left[1 - \frac{\|y_k\|^2}{a_k} + \ln \frac{\|y_k\|^2}{a_k}\right] + 2\left(\frac{\|y_k\|^2}{b_k} - 1\right) - \ln \frac{\|y_k\|^2}{\|s_k\|^2}.
\end{aligned}
$$

Aufgrund von (K_1) und (K_2) können wir die beiden letzten Terme gegen eine Konstante abschätzen: Wegen (K_2) ist

$$
c\|s_k\|^2 \leq y_k^T s_k \leq \|s_k\|\|y_k\|
$$

und somit

$$
\begin{aligned}
2\left(\frac{\|y_k\|^2}{b_k} - 1\right) - \ln \frac{\|y_k\|^2}{\|s_k\|^2} \quad &\leq \quad 2\left(\frac{\|y_k\|^2}{c\|s_k\|^2} - 1\right) - \ln c^2 \\
&\leq \quad 2\left(\frac{\gamma^2}{c} - 1\right) - \ln c^2.
\end{aligned}
$$

Durch Anwenden dieser Abschätzung auf $\psi(B_{k+1})$ und Aufsummieren erhalten wir

$$
\begin{aligned}
0 \quad &\leq \quad \psi(B_{k+1}) \\
&\leq \quad \psi(B_0) + \sum_{j=0}^{k} \left\{ \ln\left(\frac{s_j^T B_{j+\frac{1}{2}} s_j}{\|s_j\|\|B_{j+\frac{1}{2}} s_j\|}\right)^2 + \left[1 - \frac{\|B_{j+\frac{1}{2}} s_j\|^2}{s_j^T B_{j+\frac{1}{2}} s_j} + \ln \frac{\|B_{j+\frac{1}{2}} s_j\|^2}{s_j^T B_{j+\frac{1}{2}} s_j}\right] \right. \\
&\qquad \left. + \left[1 - \frac{\|y_j\|^2}{a_j} + \ln \frac{\|y_j\|^2}{a_j}\right] \right\} + \hat{C}(k+1)
\end{aligned}
$$

für alle k mit einer Konstanten $\hat{C} > 0$. Wir definieren nun wieder die Abbildung $h : (0, \infty) \longrightarrow \mathbb{R}$ durch $h(t) := t - 1 - \ln t$ und folgern hiermit aus der obigen Abschätzung

$$
\sum_{j=0}^{k} \left\{ \ln\left(\frac{\|s_j\|\|B_{j+\frac{1}{2}} s_j\|}{s_j^T B_{j+\frac{1}{2}} s_j}\right)^2 + h\left(\frac{\|B_{j+\frac{1}{2}} s_j\|^2}{s_j^T B_{j+\frac{1}{2}} s_j}\right) + h\left(\frac{\|y_j\|^2}{a_j}\right) \right\} \leq C(k+1)
$$

mit $C := \hat{C} + \psi(B_0)$. Hierauf wenden wir nun das Lemma 3.8 an. Die hierfür erforderliche Nichtnegativität der Summanden ist erfüllt. Der erste Term ist wegen der der Cauchy-Schwarz'schen Ungleichung nichtnegativ. Für die beiden anderen Terme kann wieder die Ungleichung $t - 1 - \ln t \geq 0$ für positive t verwendet werden. Es gibt also eine Indexmenge $J_k \subset \{0, \ldots, k\}$ mit mindestens $\frac{1}{2}(k+1)$ Elementen derart, daß für alle $j \in J_k$

$$
h\left(\frac{\|B_{j+\frac{1}{2}} s_j\|^2}{s_j^T B_{j+\frac{1}{2}} s_j}\right) \leq 2C \quad \text{und} \quad h\left(\frac{\|y_j\|^2}{a_j}\right) \leq 2C
$$

gilt. Wie schon im Beweis zu Satz 3.9 existieren aufgrund der Eigenschaften der Funktion h eindeutig bestimmte Konstanten $C_0 \in (0,1)$ und $C_1 \in (1,\infty)$ mit $h(C_0) = h(C_1) = 2C$. Insbesondere gilt für alle $t > 0$ mit $h(t) \leq 2C$ dann $t \in [C_0, C_1]$. Also ist

$$C_0 \leq \frac{\|B_{j+\frac{1}{2}} s_j\|^2}{s_j^T B_{j+\frac{1}{2}} s_j} \leq C_1 \quad \text{und} \quad C_0 \leq \frac{\|y_j\|^2}{a_j} \leq C_1$$

für alle $j \in J_k$. Unser Ziel ist es nun, die Existenz zweier von k unabhängigen Konstanten $\hat{\delta}, \bar{\delta} > 0$ mit

$$-\frac{g_j^T p_j}{\|g_j\|^2} \geq \hat{\delta} \quad \text{und} \quad \left(\frac{g_j^T p_j}{\|g_j\|\|p_j\|}\right)^2 \geq \bar{\delta}$$

für alle $j \in J_k$ zu zeigen. Aus der zweiten der beiden obigen Abschätzungen, den Voraussetzungen (K_2) und (K_3) sowie $b_j^2 \leq a_j c_j$ erhalten wir

$$C_0 \leq \frac{\|y_j\|^2}{a_j} \leq \frac{\gamma^2 \|s_j\|^2}{a_j} \leq \frac{\gamma^2 \|s_j\|^2 c_j}{b_j^2} \leq \left(\frac{\gamma}{c}\right)^2 \frac{c_j}{\|s_j\|^2}$$

für alle $j \in J_k$. Damit gilt für diese Indices

$$\frac{\|s_j\|}{\sqrt{c_j}} \leq \frac{\gamma}{c\sqrt{C_0}}.$$

Aus der noch nicht benutzten ersten Abschätzung erhalten wir für alle $j \in J_k$

$$\begin{aligned} \sqrt{C_1} &\geq \frac{\|B_{j+\frac{1}{2}} s_j\|}{\sqrt{s_j^T B_{j+\frac{1}{2}} s_j}} = \sqrt{\frac{\phi_j}{c_j}} \left\| B_j s_j + \frac{a_j - b_j}{a_j} y_j \right\| \\ &\geq \sqrt{\frac{\phi_j}{c_j}} \left[\|B_j s_j\| - \left| \frac{a_j - b_j}{a_j} \right| \|y_j\| \right] \end{aligned}$$

und somit

$$\frac{\|B_j s_j\|}{\sqrt{c_j}} \leq \sqrt{\frac{C_1}{\phi_j}} + \frac{|a_j - b_j|}{a_j \sqrt{c_j}} \|y_j\|.$$

Wir wollen diesen Term im folgenden weiter abschätzen, unterscheiden jedoch bei festem $j \in J_k$ nach der Größe von ϕ_j.

Fall 1: Es gilt $\phi_j \geq 1$. Hieraus folgt

$$\frac{1}{\phi_j} - 1 = \frac{b_j}{c_j} - \frac{b_j^2}{a_j c_j} = \underbrace{\frac{b_j}{c_j}}_{\geq 0} \left(1 - \frac{b_j}{a_j}\right) \leq 0$$

und damit $a_j \leq b_j$. Hiermit und mit $b_j^2 \leq a_j c_j$ können wir nun weiter abschätzen. Es folgt

$$\begin{aligned}\frac{\|B_j s_j\|}{\sqrt{c_j}} &\leq \sqrt{C_1} + \frac{|a_j - b_j|}{a_j\sqrt{c_j}}\|y_j\| \leq \sqrt{C_1} + \frac{b_j}{a_j\sqrt{c_j}}\|y_j\| \\ &\leq \sqrt{C_1} + \frac{c_j}{b_j\sqrt{c_j}}\|y_j\| \leq \sqrt{C_1} + \frac{\sqrt{c_j}}{c\|s_j\|^2}\|y_j\| \\ &\leq \sqrt{C_1} + \left(\frac{\gamma}{c}\right)\frac{\sqrt{c_j}}{\|s_j\|}.\end{aligned}$$

Durch Multiplikation dieser Ungleichung mit $\|s_j\|/\sqrt{c_j}$ erhalten wir unter Verwendung der Abschätzung für diesen Faktor

$$\frac{\|g_j\|\|p_j\|}{(-g_j^T p_j)} = \frac{\|B_j s_j\|\|s_j\|}{c_j} \leq \sqrt{C_1}\frac{\|s_j\|}{\sqrt{c_j}} + \frac{\gamma}{c} \leq \frac{\gamma}{c}\left(\sqrt{\frac{C_1}{C_0}} + 1\right)$$

und damit

$$\left(\frac{g_j^T p_j}{\|g_j\|\|p_j\|}\right)^2 \geq \left[\frac{\gamma}{c}\left(\sqrt{\frac{C_1}{C_0}} + 1\right)\right]^{-2} =: \bar{\delta}_1.$$

Wenn wir in den obigen Ausführungen etwas anders vorgehen, erhalten wir

$$\begin{aligned}\frac{\|B_j s_j\|}{\sqrt{c_j}} &\leq \sqrt{C_1} + \frac{b_j}{a_j\sqrt{c_j}}\|y_j\| \leq \sqrt{C_1} + \frac{\|y_j\|^2}{a_j}\frac{\|s_j\|}{\sqrt{c_j}} \\ &\leq \sqrt{C_1} + C_1\frac{\|s_j\|}{\sqrt{c_j}} \leq \sqrt{C_1} + C_1\frac{\gamma}{c\sqrt{C_0}}.\end{aligned}$$

Daher gilt

$$-\frac{g_j^T p_j}{\|g_j\|^2} = \frac{c_j}{\|B_j s_j\|^2} \geq \left[\sqrt{C_1} + \frac{\gamma C_1}{c\sqrt{C_0}}\right]^{-2} =: \hat{\delta}_1.$$

Damit ist der Fall $\phi_j \geq 1$ abgeschlossen. Nun zu

Fall 2: Es gilt $\phi_j < 1$ und daher mit dem Argument aus dem vorangegangenen Fall $b_j < a_j$. Die Fortführung der Abschätzung unter Verwendung der Ungleichung für $\|s_j\|/\sqrt{c_j}$ ergibt hier

$$\begin{aligned}\frac{\|B_j s_j\|}{\sqrt{c_j}} &\leq \sqrt{\frac{C_1}{\phi_j}} + \frac{|a_j - b_j|}{a_j\sqrt{c_j}}\|y_j\| \leq \sqrt{\frac{C_1}{\phi_j}} + \frac{\|y_j\|}{\sqrt{c_j}} \\ &\leq \sqrt{\frac{C_1}{\phi_j}} + \gamma\frac{\|s_j\|}{\sqrt{c_j}} \leq \sqrt{\frac{C_1}{\phi_j}} + \frac{\gamma^2}{c\sqrt{C_0}}.\end{aligned}$$

Jetzt unterscheiden wir zwei Unterfälle.

Fall 2a: Es ist $\phi_j < 1$ und $c_j \geq b_j$. Hieraus folgt

$$\phi_j = \frac{1}{b_j/c_j + 1 - b_j^2/a_j c_j} \geq \frac{1}{b_j/c_j + 1} \geq \frac{1}{2}$$

und daraus

$$\frac{\|B_j s_j\|}{\sqrt{c_j}} \leq \sqrt{2C_1} + \frac{\gamma^2}{c\sqrt{C_0}}.$$

Mit der Abschätzung für $\|s_j\|/\sqrt{c_j}$ ergibt sich dann

$$\begin{aligned} \frac{\|g_j\|\|p_j\|}{(-g_j^T p_j)} &= \frac{\|B_j s_j\|\|s_j\|}{c_j} \leq \left(\sqrt{\frac{C_1}{\phi_j}} + \frac{\gamma^2}{c\sqrt{C_0}}\right)\frac{\gamma}{c\sqrt{C_0}} \\ &\leq \left(\sqrt{2C_1} + \frac{\gamma^2}{c\sqrt{C_0}}\right)\frac{\gamma}{c\sqrt{C_0}}. \end{aligned}$$

Also ist

$$\left(\frac{g_j^T p_j}{\|g_j\|\|p_j\|}\right)^2 \geq \left[\left(\sqrt{2C_1} + \frac{\gamma^2}{c\sqrt{C_0}}\right)\frac{\gamma}{c\sqrt{C_0}}\right]^{-2} =: \bar{\delta}_2$$

und

$$-\frac{g_j^T p_j}{\|g_j\|^2} = \frac{c_j}{\|B_j s_j\|^2} \geq \left[\sqrt{2C_1} + \frac{\gamma^2}{c\sqrt{C_0}}\right]^{-2} =: \hat{\delta}_2.$$

Fall 2b: Es gilt $\phi_j < 1$ und $c_j < b_j$. Daher ist $1 < b_j/c_j$ und damit

$$\phi_j \geq \frac{1}{b_j/c_j + 1} > \frac{c_j}{2b_j} \geq \frac{c_j}{2\|y_j\|\|s_j\|} \geq \frac{c_j}{2\gamma\|s_j\|^2}.$$

Hieraus folgt

$$\frac{\|B_j s_j\|}{\sqrt{c_j}} \leq \sqrt{2\gamma C_1}\frac{\|s_j\|}{\sqrt{c_j}} + \frac{\gamma^2}{c\sqrt{C_0}} \leq \left(\sqrt{2\gamma C_1} + \gamma\right)\frac{\gamma}{c\sqrt{C_0}}.$$

Analog zum vorherigen Fall ist dann

$$\frac{\|g_j\|\|p_j\|}{(-g_j^T p_j)} = \frac{\|B_j s_j\|\|s_j\|}{c_j} \leq \left(\sqrt{2\gamma C_1} + \gamma\right)\frac{\gamma^2}{c^2 C_0}$$

und deswegen

$$\left(\frac{g_j^T p_j}{\|g_j\|\|p_j\|}\right)^2 \geq \left[\left(\sqrt{2\gamma C_1} + \gamma\right)\frac{\gamma^2}{c^2 C_0}\right]^{-2} =: \bar{\delta}_3.$$

Weiter gilt

$$-\frac{g_j^T p_j}{\|g_j\|^2} = \frac{c_j}{\|B_j s_j\|^2} \geq \left[\left(\sqrt{2\gamma C_1} + \gamma\right)\frac{\gamma}{c\sqrt{C_0}}\right]^{-2} =: \hat{\delta}_3.$$

Insgesamt erhalten wir also für $j \in J_k$

$$-\frac{g_j^T p_j}{\|g_j\|^2} \geq \min(\hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3) =: \hat{\delta} \quad \text{und} \quad \left(\frac{g_j^T p_j}{\|g_j\|\|p_j\|}\right)^2 \geq \min(\bar{\delta}_1, \bar{\delta}_2, \bar{\delta}_3) =: \bar{\delta}.$$

Mit der Abkürzung

$$\delta_j := \min\left[-\frac{g_j^T p_j}{\|g_j\|^2}, \left(\frac{g_j^T p_j}{\|g_j\|\|p_j\|}\right)^2\right]$$

folgt dann

$$\sum_{j=0}^{k} \delta_j \geq \sum_{j \in J_k} \delta_j \geq \#(J_k) \min(\hat{\delta}, \bar{\delta}) \geq (k+1)\frac{\min(\hat{\delta}, \bar{\delta})}{2} = (k+1)\delta$$

mit $\delta := \min(\hat{\delta}, \bar{\delta})/2$. Die Voraussetzungen für Satz 2.3 sind also erfüllt. Es folgt die R-lineare Konvergenz der erzeugten Folge $\{x_k\}$ gegen die unter den Voraussetzungen eindeutige Lösung x^* von (P) in L_0. □

Mit dem obigen Satz ist es also gelungen, das globale Konvergenzverhalten des BFGS-Verfahrens bei gleichmäßig konvexer Zielfunktion aus Satz 3.1 auch für das Dennis-Wolkowicz-Verfahren zu beweisen. Eine Übertragung des Konvergenzsatzes von Powell scheint aufgrund der geringen Voraussetzungen von Satz 3.2 nur schwer möglich zu sein. Ein Indiz hierfür ist die schon genannte Arbeit [12] von L. Han und G. Liu, in der für einen vergleichbaren Satz stärkere Voraussetzungen und eine extrem lange Argumentationskette benötigt werden. Der Beweis eines zu Satz 3.4 äquivalenten Konvergenzresultats bei quadratischer Zielfunktion ist möglich. Wir verschieben diesen jedoch auf einen späteren Abschnitt, da wir hierzu andere, noch zu beweisende Ergebnisse nutzen wollen. Wir fahren also mit der Untersuchung der lokalen Konvergenzeigenschaften fort.

4.2 Lokale Konvergenzaussagen

Als Grundlage für die Betrachtung der lokalen Konvergenzeigenschaften werden wir analog zu Abschnitt 3.2 eine Variante des Bounded-Deterioration-Satzes von C. G. Broyden, J. E. Dennis Jr. und J. J. Moré benutzen. Hieraus können wir dann die lokale Q-lineare Konvergenz des ungedämpften DW-Verfahrens folgern. Anschließend werden wir die Dennis-Moré-Bedingung für dieses Verfahren nachweisen und einige Folgerungen, unter anderem lokale Q-superlineare Konvergenz daraus ziehen. Über (P) setzen wir in diesem Abschnitt voraus:

(*V*) Die Zielfunktion f ist auf einer konvexen Umgebung 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.

Wir geben nun besagten Bounded-Deterioration-Satz für das Dennis-Wolkowicz-Update an. Die Darstellung dieser Eigenschaft erfolgt wie schon beim BFGS-Verfahren über die ψ-Funktion. Der sich anschließende Beweis ist weitgehend eine Übertragung des Beweises für das BFGS-Update von J. Werner aus [22].

Satz 4.2 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Voraussetzung (V) sei erfüllt. Dann existieren positive Konstanten ϵ und α mit der folgenden Eigenschaft: Sind $x_k, x_{k+1} \in B[x^*;\epsilon] := \{x \in \mathbb{R}^n : \|x - x^*\| \leq \epsilon\}$ mit $x_k \neq x_{k+1}$ und $B_k \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit gegeben, so ist $B_{k+1} \in \mathbb{R}^{n\times n}$ mit*

$$\begin{aligned} B_{k+\frac{1}{2}} &:= B_k + \frac{a_k - b_k}{a_k}\frac{y_k y_k^T}{b_k}, \\ B_{k+1} &:= B_{k+\frac{1}{2}} - \frac{(B_{k+\frac{1}{2}}s_k)(B_{k+\frac{1}{2}}s_k)^T}{s_k^T B_{k+\frac{1}{2}} s_k} + \frac{y_k y_k^T}{b_k} \end{aligned}$$

symmetrisch und positiv definit, und es gilt

$$\psi(B^{*-1/2}B_{k+1}B^{*-1/2}) \leq \psi(B^{*-1/2}B_kB^{*-1/2}) + \alpha\sigma(x_k, x_{k+1})$$

mit

$$\sigma : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}, \quad \sigma(u,v) := \max(\|u - x^*\|, \|v - x^*\|).$$

Beweis: Aus der Voraussetzung (V) folgt die Existenz eines $\epsilon_1 > 0$ derart, daß f zweimal stetig differenzierbar auf $B[x^*;\epsilon_1]$ und $\nabla^2 f$ lipschitzstetig auf $B[x^*;\epsilon_1]$ in x^* mit einer Lipschitz-Konstanten $L > 0$ ist. Mit $\lambda^*_{\min}$ sei der kleinste Eigenwert von B^* bezeichnet. Ferner seien $\tilde{s}_k := B^{*1/2}s_k$ und $\tilde{y}_k := B^{*-1/2}y_k$. Bevor wir uns dem eigentlichen Vorhaben, der Abschätzung von $\psi(B^{*-1/2}B_{k+1}B^{*-1/2})$ widmen, beweisen wir noch einige dafür notwendige Hilfsabschätzungen.

(i) Es seien $x_k, x_{k+1} \in B[x^*;\epsilon_1]$ gegeben. Dann gilt

$$\|\tilde{y}_k - \tilde{s}_k\| \leq \frac{L}{\lambda^*_{\min}}\|\tilde{s}_k\|\,\sigma(x_k, x_{k+1}).$$

Denn es ist

$$\begin{aligned} \|\tilde{y}_k - \tilde{s}_k\| &= \|B^{*-1/2}(y_k - B^*s_k)\| \leq \|B^{*-1/2}\|\|y_k - B^*s_k\| \\ &= \frac{1}{\sqrt{\lambda^*_{\min}}}\|\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x^*)s_k\| \\ &= \frac{1}{\sqrt{\lambda^*_{\min}}}\left\|\int_0^1 [\nabla^2 f(x_k + t(x_{k+1} - x_k)) - \nabla^2 f(x^*)]s_k dt\right\| \\ &\leq \frac{1}{\sqrt{\lambda^*_{\min}}}\|s_k\|\int_0^1 \|\nabla^2 f(x_k + t(x_{k+1} - x_k)) - \nabla^2 f(x^*)\|dt \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{L}{\sqrt{\lambda^*_{\min}}}\|s_k\| \int_0^1 \|x_k + t(x_{k+1} - x_k) - x^*\| dt \\
&= \frac{L}{\sqrt{\lambda^*_{\min}}}\|s_k\| \int_0^1 \|(1-t)(x_k - x^*) + t(x_{k+1} - x^*)\| dt \\
&\leq \frac{L}{\sqrt{\lambda^*_{\min}}}\|s_k\| \left(\frac{\|x_k - x^*\|}{2} + \frac{\|x_{k+1} - x^*\|}{2} \right) \\
&\leq \frac{L}{\sqrt{\lambda^*_{\min}}}\|s_k\| \max(\|x_k - x^*\|, \|x_{k+1} - x^*\|) \\
&= \frac{L}{\sqrt{\lambda^*_{\min}}}\|B^{*-1/2}\tilde{s}_k\| \, \sigma(x_k, x_{k+1}) \\
&\leq \frac{L}{\lambda^*_{\min}}\|\tilde{s}_k\| \, \sigma(x_k, x_{k+1}).
\end{aligned}
$$

(ii) Hieraus folgt

$$
\begin{aligned}
\tilde{y}_k^T \tilde{s}_k &= \|\tilde{s}_k\|^2 - (\tilde{s}_k - \tilde{y}_k)^T \tilde{s}_k \geq (\|\tilde{s}_k\| - \|\tilde{y}_k - \tilde{s}_k\|)\|\tilde{s}_k\| \\
&\geq \left[1 - \frac{L}{\lambda^*_{\min}} \sigma(x_k, x_{k+1})\right] \|\tilde{s}_k\|^2.
\end{aligned}
$$

(iii) Mit

$$
\epsilon := \min\left(\epsilon_1, \frac{\lambda^*_{\min}}{2L}\right),
$$

erhalten wir für alle $x_k, x_{k+1} \in B[x^*; \epsilon] \subset B[x^*; \epsilon_1]$

$$
\sigma(x_k, x_{k+1}) \leq \epsilon \leq \frac{\lambda^*_{\min}}{2L}
$$

und folgern aus (ii)

$$
\tilde{y}_k^T \tilde{s}_k \geq \left[1 - \frac{L}{\lambda^*_{\min}} \sigma(x_k, x_{k+1})\right] \|\tilde{s}_k\|^2 \geq \frac{1}{2}\|\tilde{s}_k\|^2.
$$

Insbesondere folgt hieraus wegen $y_k^T s_k = \tilde{y}_k^T \tilde{s}_k > 0$ die positive Definitheit der Matrizen $B_{k+\frac{1}{2}}$ und B_{k+1}. Die Durchführbarkeit des Verfahrens ist also gesichert.

(iv) Für ungleiche $x_k, x_{k+1} \in B[x^*; \epsilon]$ gilt

$$
-\ln\left(\frac{\|\tilde{y}_k\|}{\|\tilde{s}_k\|}\right)^2 \leq \frac{4L}{\lambda^*_{\min}} \sigma(x_k, x_{k+1}).
$$

Denn nach (ii) ist

$$\begin{aligned} -\ln\left(\frac{\|\tilde{y}_k\|}{\|\tilde{s}_k\|}\right)^2 &= -2\ln\left(\frac{\|\tilde{y}_k\|\|\tilde{s}_k\|}{\|\tilde{s}_k\|^2}\right) \leq -2\ln\left(\frac{\tilde{y}_k^T\tilde{s}_k}{\|\tilde{s}_k\|^2}\right) \\ &\leq -2\ln\left(1-\frac{L}{\lambda^*_{\min}}\sigma(x_k,x_{k+1})\right). \end{aligned}$$

Nun wenden wir die für $t \in [0,1/2]$ gültige Ungleichung $-\ln(1-t) \leq 2t$ an, vergewissern uns jedoch zuvor, daß

$$\frac{L}{\lambda^*_{\min}}\sigma(x_k,x_{k+1}) \leq \frac{L}{\lambda^*_{\min}}\epsilon \leq \frac{1}{2}$$

ist. Damit gilt dann

$$-\ln\left(\frac{\|\tilde{y}_k\|}{\|\tilde{s}_k\|}\right)^2 \leq \frac{4L}{\lambda^*_{\min}}\sigma(x_k,x_{k+1}).$$

(v) Aus (i) und (iii) folgt für $x_k, x_{k+1} \in B[x^*;\epsilon]$ mit $x_k \neq x_{k+1}$

$$\begin{aligned} \frac{\|\tilde{y}_k\|^2}{\tilde{y}_k^T\tilde{s}_k} - 1 &= \frac{\tilde{y}_k^T(\tilde{y}_k-\tilde{s}_k)}{\tilde{y}_k^T\tilde{s}_k} \leq \frac{\|\tilde{y}_k\|\|\tilde{y}_k-\tilde{s}_k\|}{\tilde{y}_k^T\tilde{s}_k} \\ &\leq \frac{\|\tilde{y}_k\|}{\tilde{y}_k^T\tilde{s}_k}\frac{L}{\lambda^*_{\min}}\|\tilde{s}_k\|\sigma(x_k,x_{k+1}) \\ &\leq \frac{2L}{\lambda^*_{\min}}\frac{\|\tilde{y}_k\|}{\|\tilde{s}_k\|}\sigma(x_k,x_{k+1}) \\ &\leq \frac{2L}{\lambda^*_{\min}}\frac{\|\tilde{y}_k-\tilde{s}_k\|+\|\tilde{s}_k\|}{\|\tilde{s}_k\|}\sigma(x_k,x_{k+1}) \\ &\leq \frac{2L}{\lambda^*_{\min}}\left(\frac{L}{\lambda^*_{\min}}\sigma(x_k,x_{k+1})+1\right)\sigma(x_k,x_{k+1}) \\ &\leq \frac{2L}{\lambda^*_{\min}}\left(\frac{L}{\lambda^*_{\min}}\epsilon+1\right)\sigma(x_k,x_{k+1}). \end{aligned}$$

Nun kommen wir zur eigentlichen Behauptung. Es seien $x_k, x_{k+1} \in B[x^*;\epsilon]$ mit $x_k \neq x_{k+1}$ und $B_k \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit gegeben. Wir setzen

$$\tilde{B}_k := B^{*-1/2}B_kB^{*-1/2}, \qquad \tilde{B}_{k+\frac{1}{2}} := B^{*-1/2}B_{k+\frac{1}{2}}B^{*-1/2}.$$

Dann ist

$$\begin{aligned} \tilde{B}_{k+1} &:= B^{*-1/2}B_{k+1}B^{*-1/2} \\ &= \tilde{B}_{k+\frac{1}{2}} - \frac{(\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k)(\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k)^T}{\tilde{s}_k^T\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k} + \frac{\tilde{y}_k^T\tilde{y}_k}{\tilde{y}_k^T\tilde{s}_k}. \end{aligned}$$

Es gilt dann wie im Beweis von Satz 4.1 für die Spur von $\tilde{B}_{k+1}$

$$\begin{aligned}\operatorname{tr}(\tilde{B}_{k+1}) &= \operatorname{tr}(\tilde{B}_k) + \frac{a_k - b_k}{a_k}\frac{\|\tilde{y}_k\|^2}{\tilde{y}_k^T\tilde{s}_k} - \frac{\|\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k} + \frac{\|\tilde{y}_k\|^2}{\tilde{y}_k^T\tilde{s}_k} \\ &= \operatorname{tr}(\tilde{B}_k) + \left[1 + \frac{a_k - b_k}{a_k}\right]\frac{\|\tilde{y}_k\|^2}{\tilde{y}_k^T\tilde{s}_k} - \frac{\|\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k}.\end{aligned}$$

Für die Determinante von $\tilde{B}_{k+1}$ gilt nach dem Determinantenmultiplikationssatz

$$\begin{aligned}\det(\tilde{B}_{k+1}) &= \det(B^{*-1/2})\det(B_{k+1})\det(B^{*-1/2}) \\ &= \frac{\phi_k a_k}{c_k}\det(B^{*-1/2})\det(B_k)\det(B^{*-1/2}) \\ &= \frac{\phi_k a_k}{c_k}\det(\tilde{B}_k).\end{aligned}$$

Zusammen mit (iv) und (v) ergibt sich dann wie im Beweis zu Satz 4.1 für ungleiche $x_k, x_{k+1} \in B[x^*;\epsilon]$

$$\begin{aligned}\psi(\tilde{B}_{k+1}) &= \psi(\tilde{B}_k) + \left[1 + \frac{a_k - b_k}{a_k}\right]\frac{\|\tilde{y}_k\|^2}{\tilde{y}_k^T\tilde{s}_k} - \frac{\|\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k} - \ln\frac{\phi_k a_k}{c_k} \\ &\vdots \\ &= \psi(\tilde{B}_k) + \underbrace{\ln\left(\frac{\tilde{s}_k^T\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k}{\|\tilde{s}_k\|\|\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k\|}\right)^2}_{\leq 0} + \underbrace{\left[1 - \frac{\|\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k} + \ln\frac{\|\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k}\right]}_{\leq 0} \\ &\quad + \underbrace{\left[1 - \frac{\|\tilde{y}_k\|^2}{a_k} + \ln\frac{\|\tilde{y}_k\|^2}{a_k}\right]}_{\leq 0} + 2\left(\frac{\|\tilde{y}_k\|^2}{\tilde{y}_k^T\tilde{s}_k} - 1\right) - \ln\frac{\|\tilde{y}_k\|^2}{\|\tilde{s}_k\|^2} \\ &\leq \psi(\tilde{B}_k) + 2\left(\frac{\|\tilde{y}_k\|^2}{\tilde{y}_k^T\tilde{s}_k} - 1\right) - \ln\frac{\|\tilde{y}_k\|^2}{\|\tilde{s}_k\|^2} \\ &\leq \psi(\tilde{B}_k) + \frac{4L}{\lambda^*_{\min}}\left(\frac{L}{\lambda^*_{\min}}\epsilon + 1\right)\sigma(x_k, x_{k+1}) + \frac{4L}{\lambda^*_{\min}}\sigma(x_k, x_{k+1}) \\ &= \psi(\tilde{B}_k) + \alpha\sigma(x_k, x_{k+1})\end{aligned}$$

mit

$$\alpha := \frac{4L}{\lambda^*_{\min}}\left(\frac{L}{\lambda^*_{\min}}\epsilon + 2\right).$$

Die Behauptung ist somit bewiesen. □

Bei der Darstellung von Satz 4.2, beziehungsweise der Bounded-Deterioration-Eigenschaft des DW-Updates haben wir uns wie schon beim BFGS-Update für

die ψ-Funktion von R. H. Byrd und J. Nocedal entschieden. Es ist jedoch möglich und früher auch üblich gewesen, diese Eigenschaft mit einer gewichteten Frobeniusnorm zu formulieren. Hierzu müsste man in Satz 4.2 die Ungleichung

$$\psi(B^{*-1/2}B_{k+1}B^{*-1/2}) \leq \psi(B^{*-1/2}B_kB^{*-1/2}) + \alpha\sigma(x_k, x_{k+1})$$

durch

$$\left\|B^{*-1/2}(B_{k+1} - B^*)B^{*-1/2}\right\|_F \leq \left\|B^{*-1/2}(B_k - B^*)B^{*-1/2}\right\|_F + \alpha\sigma(x_k, x_{k+1})$$

ersetzen. Für das BFGS-Update findet man deratige Formulierungen beispielsweise bei C. G. Broyden, J. E. Dennis Jr. und J. Moré in [1] oder auch bei J. E. Dennis Jr., J. Martínez und R. A. Tapia in [5].

Nun drängt sich natürlich die Frage auf, warum wir für den vorangegangenen Satz nicht die klassische Frobeniusnorm-Darstellung gewählt haben. Die Antwort ist recht simpel: Da sich die Beweisführung durch den Einsatz der ψ-Funktion erheblich vereinfacht. Um diese Aussage zu untermauern, erläutern wir nun anhand der Ausführungen von Dennis, Martínez und Tapia über das BFGS-Verfahren eine mögliche Beweisstruktur für einen klassischen Bounded-Deterioration-Satz mit gewicheteter Forbeniusnorm für das Dennis-Wolkowicz-Verfahren. Wir verwenden die gleichen Bezeichnungen und Voraussetzungen wie im Beweis von Satz 4.2. Ferner sei die Matrixnorm $\|\cdot\|_*$ durch

$$\|A\|_* := \left\|B^{*-1/2}AB^{*-1/2}\right\|_F$$

für alle $A \in \mathbb{R}^{n\times n}$ definiert. Zu zeigen ist also, daß für alle $k = 0, 1, \ldots$

$$\|B_{k+1} - B^*\|_* \leq \|B_k - B^*\|_* + \alpha\sigma(x_k, x_{k+1})$$

mit einer Konstanten $\alpha > 0$ gilt. Die weitere Beweisführung erfolgt in mehreren Schritten. Zunächst wählt man sich eine geeignete Matrix $B' \in \mathbb{R}^{n\times n}$ und zeigt hiermit die Abschätzungen

$$\|B' - B^*\|_* \leq \|B_k - B^*\|_*$$

und

$$\|B_{k+1} - B'\|_* \leq \alpha\sigma(x_k, x_{k+1})$$

mit einer Konstanten $\alpha > 0$. Beim Beweis dieser beiden Ungleichungen werden neben einigen sehr umständlichen Rechnungen auch Hilfsaussagen, die den Aussagen (i), (ii) und (iii) im vorangegangenen Beweis sehr ähnlich sind, benötigt. Durch Anwenden der Dreicksungleichung erhält man dann mit den beiden obigen Abschätzungen die Behauptung. Dennis, Martínez und Tapia setzen in ihrem Beweis für das BFGS-Verfahren

$$B' := B_k - \frac{(B_ks_k)(B_ks_k)^T}{s_k^TB_ks_k} + \frac{(B^*s_k)(B^*s_k)^T}{s_k^TB^*s_k}.$$

Ob diese Wahl auch für das Dennis-Wolkowicz-Verfahren sinnvoll ist, ist fraglich. Die erste der beiden zu zeigenden Ungleichungen läßt sich dann zwar samt Beweis übernehmen, jedoch verkompliziert sich der Ausdruck $\|B_{k+1} - B'\|_*$ derartig, daß ein Nachweis der zweiten Ungleichung nur schwer möglich scheint. Der erfolgreiche Beweis eines klassischen Bounded-Deterioration-Satzes für das Dennis-Wolkowicz-Update wird also von einer (noch zu treffenden) geschickten Wahl der Matrix B' abhängen. Ob ein B', mit welchem sich beide Abschätzungen zeigen lassen, existiert, ist nicht sicher. Falls eine derartige Wahl von B' möglich ist, wird wie beim BFGS-Verfahren in [5] die wesentliche Arbeit im Nachweis der beiden Abschätzungen bestehen. Schon beim BFGS-Verfahren ist allein der Aufwand für eine der beiden Abschätzungen erheblich größer als für die Abschätzung von $\psi(\tilde{B}_{k+1})$ im Beweis von Satz 4.2. Berücksichtigt man jetzt noch die im Vergleich zum BFGS-Update kompliziertere Struktur der DW-Update-Formel, so wird sich der Beweisaufwand sicherlich noch weiter erhöhen. Insgesamt läßt sich also sagen, daß zum Beweis von Satz 4.2 zwar zwei Hilfsabschätzungen mehr zu zeigen sind als bei einem klassischen Satz, daß dafür aber der Aufwand zur Abschätzung der ψ-Funktion sehr viel geringer ist als der Aufwand zur Abschätzung der gewichteten Frobeniusnorm. Daher erscheint die Formulierung eines klassischen Bounded-Deterioration-Satzes für das Dennis-Wolkowiz-Verfahren nicht sinvoll. Gleiches gilt im übrigen auch für das BFGS-Verfahren, dort fallen die Unterschiede nicht ganz so erheblich, aber immer noch deutlich aus.

Nach diesem kurzen Exkurs kehren wir nun zur Untersuchung der lokalen Konvergenzeigenschaften des Dennis-Wolkowicz-Verfahrens zurück. Aus Satz 4.2 ergibt sich unmittelbar die lokale Q-lineare Konvergenz des ungedämpften Dennis-Wolkowicz-Verfahrens.

Korollar 4.3 *Gegeben sei die Aufgabe (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 $x_0 \in \mathbb{R}^n$ mit $\|x_0 - x^*\| \leq \epsilon(r)$ und $B_0 \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit mit $\psi(B^{*-1/2} B_0 B^{*-1/2}) - n \leq \delta(r)$, so ist das ungedämpfte Dennis-Wolkowicz-Verfahren durchführbar und liefert (falls kein vorzeitiger Abbruch stattfindet) 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.

Auf einen Beweis dieser Aussage können wir verzichten, da, wie schon im Anschluss an Satz 3.6 bemerkt, ein solcher Beweis unabhängig von der speziellen Wahl der Matrix-Update-Formel ist, solange die erzeugten Matrizen positiv definit sind und die Bounded-Deterioration-Eigenschaft aus Satz 3.5 besitzen. Beides ist hier nach Satz 4.2 erfüllt.

Unser nächstes Ziel ist es, eine bessere Aussage über die Konvergenzgeschwindigkeit zu machen. Hierzu weisen wir nach, daß die Dennis-Moré-Bedingung erfüllt ist. Daraus folgt lokale Q-superlineare Konvergenz des ungedämpften DW-Verfahrens sowie der Übergang des gedämpften in das ungedämpfte Verfahren nach endlich vielen Schritten. Bei der Beweisführung folgen wir im wesentlichen J. Werner [21], nutzen allerdings analog zu [22] die eben bewiesene Bounded-Deterioration-Eigenschaft des DW-Updates.

Satz 4.4 *Gegeben sei die Optimierungsaufgabe (P). Die Voraussetzung (V) sei erfüllt. Das gedämpfte oder ungedämpfte Dennis-Wolkowicz-Verfahren erzeuge eine (o.B.d.A.) nicht vorzeitig abbrechende Folge* $\{x_k\}$ *mit* $\sum_{k=0}^{\infty} \|x_k - x^*\| < \infty$ *und eine Folge symmetrischer, positiv definiter Matrizen* $\{B_k\} \subset \mathbb{R}^{n\times n}$*. Dann sind die Folgen* $\{\|B_k\|\}$ *und* $\{\|B_k^{-1}\|\}$ *beschränkt, und es gilt die Dennis-Moré-Bedingung*

$$\lim_{k\to\infty} \frac{\|[B_k - \nabla^2 f(x^*)](x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} = 0.$$

Beweis: Zunächst nutzen wir den Bounded-Deterioration-Satz 4.2 und seinen Beweis. Mit den dortigen Bezeichungen gilt für ungleiche $x_k, x_{k+1} \in B[x^*; \epsilon]$

$$\psi(\tilde{B}_{k+1}) \le \psi(\tilde{B}_k) + \alpha \max(\|x_k - x^*\|, \|x_{k+1} - x^*\|)$$

mit einer Konstanten $\alpha > 0$. Aufsummieren liefert

$$\psi(\tilde{B}_{k+1}) \le \psi(\tilde{B}_0) + \alpha \sum_{j=0}^{k} \max(\|x_j - x^*\|, \|x_{j+1} - x^*\|).$$

Da nach Voraussetzung $\sum_{k=0}^{\infty} \|x_k - x^*\| < \infty$ gilt, ist die Folge $\{\psi(\tilde{B}_k)\}$ beschränkt. Hieraus wollen wir nun die Beschränktheit der Folgen $\{\|B_k\|\}$ und $\{\|B_k^{-1}\|\}$ folgern. Es gilt nämlich

$$\psi(\tilde{B}_k) = \sum_{i=1}^{n} \underbrace{\lambda_i(\tilde{B}_k) - \ln \lambda_i(\tilde{B}_k)}_{\ge 1} \le C$$

mit einer hinreichend großen Konstanten $C > 0$ und daher für alle $i = 1, \ldots, n$

$$\lambda_i(\tilde{B}_k) - \ln \lambda_i(\tilde{B}_k) \le C.$$

Dabei bezeichnen $\lambda_i(\tilde{B}_k)$ mit $i = 1, \ldots, n$ die Eigenwerte von $\tilde{B}_k$. Von diesen sei $\lambda_{\max}(\tilde{B}_k)$ der größte und $\lambda_{\min}(\tilde{B}_k)$ der kleinste. Wegen der positiven Definitheit von $\tilde{B}_k$ gilt

$$C \ge \lambda_{\min}(\tilde{B}_k) - \ln \lambda_{\min}(\tilde{B}_k) > -\ln \lambda_{\min}(\tilde{B}_k) = \ln \frac{1}{\lambda_{\min}(\tilde{B}_k)} = \ln \|\tilde{B}_k^{-1}\|.$$

Durch eine Anwendung der Exponentialfunktion erhalten wir

$$\|B_k^{-1}\| = \|B^{*-1/2}\tilde{B}_k^{-1}B^{*-1/2}\| \le e^C\|B^{*-1}\| = \frac{e^C}{\lambda^*_{\min}},$$

wobei $\lambda^*_{\min}$ wieder den kleinsten Eigenwert von B^* bezeichnet. Hieraus folgt die Beschränktheit von $\{\|B_k^{-1}\|\}$. Zur Herleitung der Beschränktheit der Folge $\{\|B_k\|\}$ verwenden wir die Beziehung $t - 2\ln t \ge 0$ für $t > 0$, erhalten damit

$$C \ge \lambda_{\max}(\tilde{B}_k) - \ln\lambda_{\max}(\tilde{B}_k) \ge \ln\lambda_{\max}(\tilde{B}_k) = \ln\|\tilde{B}_k\|$$

und dadurch

$$\|B_k\| \le \|B^*\|\|\tilde{B}_k\| \le \lambda^*_{\max}e^C$$

mit $\lambda^*_{\max}$ als dem größten Eigenwert von B^*. Zum Nachweis der Dennis-Moré-Bedingung verwenden wir noch einmal den Beweis des Satzes 4.2. Wir haben nachgewiesen, daß

$$\begin{aligned}\psi(\tilde{B}_{k+1}) \;\le\; & \psi(\tilde{B}_k) + \ln\left(\frac{\tilde{s}_k^T\tilde{B}_{k+\frac12}\tilde{s}_k}{\|\tilde{s}_k\|\|\tilde{B}_{k+\frac12}\tilde{s}_k\|}\right)^2 + \left[1 - \frac{\|\tilde{B}_{k+\frac12}\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_{k+\frac12}\tilde{s}_k} + \ln\frac{\|\tilde{B}_{k+\frac12}\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_{k+\frac12}\tilde{s}_k}\right] \\ & + \left[1 - \frac{\|\tilde{y}_k\|^2}{a_k} + \ln\frac{\|\tilde{y}_k\|^2}{a_k}\right] + \alpha\max(\|x_k - x^*\|, \|x_{k+1} - x^*\|)\end{aligned}$$

für alle ungleichen $x_k, x_{k+1} \in B[x^*;\epsilon]$ und $k = 0, 1, \ldots$ gilt. Durch Aufsummieren erhalten wir wegen $\sum_{k=0}^{\infty}\|x_k - x^*\| < \infty$ die Existenz einer Konstanten $\hat{C}$ mit

$$\begin{aligned}0 \;\le\; & \psi(\tilde{B}_{k+1}) \\ \le\; & \psi(\tilde{B}_0) + \sum_{j=0}^{k}\left\{\ln\left(\frac{\tilde{s}_j^T\tilde{B}_{j+\frac12}\tilde{s}_j}{\|\tilde{s}_j\|\|\tilde{B}_{j+\frac12}\tilde{s}_j\|}\right)^2 + \left[1 - \frac{\|\tilde{B}_{j+\frac12}\tilde{s}_j\|^2}{\tilde{s}_j^T\tilde{B}_{j+\frac12}\tilde{s}_j} + \ln\frac{\|\tilde{B}_{j+\frac12}\tilde{s}_j\|^2}{\tilde{s}_j^T\tilde{B}_{j+\frac12}\tilde{s}_j}\right]\right. \\ & \left. + \left[1 - \frac{\|\tilde{y}_j\|^2}{a_j} + \ln\frac{\|\tilde{y}_j\|^2}{a_j}\right]\right\} + \hat{C}\end{aligned}$$

für alle k. Hieraus folgt für alle $k = 0, 1, \ldots$

$$\begin{aligned}\psi(\tilde{B}_0) + \hat{C} \;\ge\; & \sum_{j=0}^{k}\Bigg\{\underbrace{\ln\left(\frac{\|\tilde{s}_j\|\|\tilde{B}_{j+\frac12}\tilde{s}_j\|}{\tilde{s}_j^T\tilde{B}_{j+\frac12}\tilde{s}_j}\right)^2}_{\ge 0} + \underbrace{\left[\frac{\|\tilde{B}_{j+\frac12}\tilde{s}_j\|^2}{\tilde{s}_j^T\tilde{B}_{j+\frac12}\tilde{s}_j} - 1 - \ln\frac{\|\tilde{B}_{j+\frac12}\tilde{s}_j\|^2}{\tilde{s}_j^T\tilde{B}_{j+\frac12}\tilde{s}_j}\right]}_{\ge 0} \\ & + \underbrace{\left[\frac{\|\tilde{y}_j\|^2}{a_j} - 1 - \ln\frac{\|\tilde{y}_j\|^2}{a_j}\right]}_{\ge 0}\Bigg\}.\end{aligned}$$

Daher gilt notwendigerweise für die Terme in der Summe

$$\lim_{k\to\infty}\ln\left(\frac{\|\tilde{s}_k\|\|\tilde{B}_{k+\frac12}\tilde{s}_k\|}{\tilde{s}_k^T\tilde{B}_{k+\frac12}\tilde{s}_k}\right)^2 = 0, \quad \lim_{k\to\infty}\left[\frac{\|\tilde{B}_{k+\frac12}\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_{k+\frac12}\tilde{s}_k} - 1 - \ln\frac{\|\tilde{B}_{k+\frac12}\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_{k+\frac12}\tilde{s}_k}\right] = 0$$

und

$$\lim_{k\to\infty}\left[\frac{\|\tilde{y}_k\|^2}{a_k} - 1 - \ln\frac{\|\tilde{y}_k\|^2}{a_k}\right] = 0.$$

Daraus folgern wir

$$\lim_{k\to\infty}\frac{\|\tilde{s}_k\|\|\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k\|}{\tilde{s}_k^T\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k} = 1, \quad \lim_{k\to\infty}\frac{\|\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k\|^2}{\tilde{s}_k^T\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k} = 1, \quad \lim_{k\to\infty}\frac{\|\tilde{y}_k\|^2}{a_k} = 1.$$

Durch Multiplikation des zweiten Grenzwertes mit dem Kehrwert des ersten und des zweiten Grenzwertes mit dem quadrierten Kehrwert des ersten ergibt sich

$$\lim_{k\to\infty}\frac{\|\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k\|}{\|\tilde{s}_k\|} = 1 \quad \text{und} \quad \lim_{k\to\infty}\frac{\tilde{s}_k^T\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k}{\|\tilde{s}_k\|^2} = 1.$$

Damit gilt dann

$$\frac{\|(\tilde{B}_{k+\frac{1}{2}} - I)\tilde{s}_k\|^2}{\|\tilde{s}_k\|^2} = \underbrace{\frac{\|\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k\|^2}{\|\tilde{s}_k\|^2}}_{\to 1} - 2\underbrace{\frac{\tilde{s}_k^T\tilde{B}_{k+\frac{1}{2}}\tilde{s}_k}{\|\tilde{s}_k\|^2}}_{\to 1} + 1 \longrightarrow 0.$$

Hiermit können wir nun den uns eigentlich interessierenden Term abschätzen:

$$\begin{aligned}
\frac{\|(B_k - \nabla^2 f(x^*))s_k\|}{\|s_k\|} &= \frac{\|B^{*1/2}(\tilde{B}_k - I)B^{*1/2}s_k\|}{\|s_k\|} = \frac{\|B^{*1/2}(\tilde{B}_k - I)\tilde{s}_k\|}{\|B^{*-1/2}\tilde{s}_k\|} \\
&\le \|B^*\|\frac{\|(\tilde{B}_k - I)\tilde{s}_k\|}{\|\tilde{s}_k\|} \\
&\le \|B^*\|\frac{\|(\tilde{B}_{k+\frac{1}{2}} - I)\tilde{s}_k\|}{\|\tilde{s}_k\|} + \frac{\|B^*\|}{\|\tilde{s}_k\|}\left\|\frac{a_k - b_k}{a_k}\frac{\tilde{y}_k\tilde{y}_k^T}{\tilde{y}_k^T\tilde{s}_k}\tilde{s}_k\right\| \\
&= \|B^*\|\frac{\|(\tilde{B}_{k+\frac{1}{2}} - I)\tilde{s}_k\|}{\|\tilde{s}_k\|} + \|B^*\|\frac{|a_k - b_k|}{a_k}\frac{\|\tilde{y}_k\|}{\|\tilde{s}_k\|}.
\end{aligned}$$

Die Konvergenz des ersten Summanden gegen Null haben wir schon bewiesen. Es verbleibt also noch

$$\lim_{k\to\infty}\frac{|a_k - b_k|}{a_k}\frac{\|\tilde{y}_k\|}{\|\tilde{s}_k\|} = 0$$

zu zeigen. Mit den Hilfsabschätzungen (i) und (iii) aus dem Beweis zu Satz 4.2 gilt für $x_k, x_{k+1} \in B[x^*;\epsilon]$ mit $x_k \neq x_{k+1}$

$$\frac{\|\tilde{y}_k\|}{\|\tilde{s}_k\|} \le \frac{\|\tilde{s}_k\| + \|\tilde{y}_k - \tilde{s}_k\|}{\|\tilde{s}_k\|} \le 1 + \frac{L}{\lambda^*_{\min}}\sigma(x_k, x_{k+1})$$

und

$$\frac{\|\tilde{y}_k\|}{\|\tilde{s}_k\|} = \frac{\|\tilde{y}_k\|\|\tilde{s}_k\|}{\|\tilde{s}_k\|^2} \ge \frac{\tilde{y}_k^T\tilde{s}_k}{\|\tilde{s}_k\|^2} \ge \frac{1}{2}.$$

Mit diesen Ungleichungen gilt nun

$$\begin{aligned}\left|\frac{a_k - b_k}{a_k}\right| \frac{\|\tilde{y}_k\|}{\|\tilde{s}_k\|} &= \left|1 - \frac{b_k}{a_k}\right| \frac{\|\tilde{y}_k\|}{\|\tilde{s}_k\|} \leq \left|1 - \frac{\tilde{y}_k^T \tilde{s}_k}{a_k}\right| \left(1 + \frac{L}{\lambda^*_{\min}} \sigma(x_k, x_{k+1})\right) \\ &= \left|1 - \frac{\|\tilde{y}_k\|^2}{a_k} \frac{\tilde{y}_k^T \tilde{s}_k}{\|\tilde{y}_k\|^2}\right| \left(1 + \frac{L}{\lambda^*_{\min}} \sigma(x_k, x_{k+1})\right).\end{aligned}$$

Wiederum mit (i) und der Abschätzung für $\|\tilde{y}_k\|/\|\tilde{s}_k\|$ gilt

$$\begin{aligned}\left|1 - \frac{\tilde{y}_k^T \tilde{s}_k}{\|\tilde{y}_k\|^2}\right| &= \frac{|\tilde{y}_k^T (\tilde{y}_k - \tilde{s}_k)|}{\|\tilde{y}_k\|^2} \leq \frac{\|\tilde{y}_k - \tilde{s}_k\|}{\|\tilde{y}_k\|} = \frac{\|\tilde{s}_k\|}{\|\tilde{y}_k\|} \frac{\|\tilde{y}_k - \tilde{s}_k\|}{\|\tilde{s}_k\|} \\ &\leq \frac{2L}{\lambda^*_{\min}} \sigma(x_k, x_{k+1}).\end{aligned}$$

Der letzte Term konvergiert wegen $\sum_{k=0}^{\infty} \|x_k - x^*\| < \infty$ beziehungsweise $x_k \to x^*$ gegen Null. Hieraus folgt

$$\lim_{k \to \infty} \frac{\tilde{y}_k^T \tilde{s}_k}{\|\tilde{y}_k\|^2} = 1.$$

Wegen $\lim_{k\to\infty} \|\tilde{y}_k\|/a_k = 1$ und $\lim_{k\to\infty} \sigma(x_k, x_{k+1}) = 0$ gilt dann

$$\lim_{k \to \infty} \frac{|a_k - b_k|}{a_k} \frac{\|\tilde{y}_k\|}{\|\tilde{s}_k\|} = 0,$$

woraus die Behauptung folgt. □

Die zu obigem Satz erforderlichen Bemerkungen sind dieselben wie beim BFGS-Verfahren. Aus Gründen der Vollständigkeit wiederholen wir sie.

Bemerkungen:

- Die Voraussetzung $\sum_{k=0}^{\infty} \|x_k - x^*\| < \infty$ für den obigen Satz ist erfüllt, falls die Folge $\{x_k\}$ R-linear oder Q-linear gegen x^* konvergiert. Nach den bisherigen Ergebnissen ist dies global für das durch eine semi-effiziente Schrittweite gedämpfte und lokal für das ungedämpfte DW-Verfahren der Fall.

- Da die Dennis-Moré-Bedingung erfüllt ist, folgt aus dem gleichnamigen Satz 2.4 sofort die lokale Q-superlineare Konvergenz des ungedämpften DW-Verfahrens.

- Mit Satz 2.5 folgt aus dem vorangegangenen Satz, daß bei Dämpfung des Verfahrens durch die Wolfe- oder die Armijo-Schrittweite $t_k = 1$ für alle hinreichend großen k gilt. Das gedämpfte Verfahren geht also in das ungedämpfte über, und es folgt Q-superlineare Konvergenz. Bei der Wahl der Schrittweite ist es dabei natürlich wichtig, die Zulässigkeit von $t_k = 1$ als erstes zu testen.

Insgesamt ist es nun gelungen, die in Abschnitt 3.2 vorgestellten lokalen Konvergenzeigenschaften des BFGS-Verfahrens auf das Dennis-Wolkowicz-Verfahren zu übertragen. Ferner haben wir jetzt die nötigen Hilfsmittel, um die bisher versäumte Untersuchung der globalen Konvergenz bei quadratischer Zielfunktion durchzuführen.

4.3 Konvergenz bei quadratischer Zielfunktion

Im folgenden Satz werden wir die globale Q-superlineare Konvergenz des ungedämpften Dennis-Wolkowicz-Verfahrens bei strikt konvexer quadratischer Zielfunktion beweisen. Die Beweisführung erfolgt analog zu J. Werner [21].

Satz 4.5 *Gegeben sei die unrestringierte Optimierungsaufgabe*

$$\textit{Minimiere} \quad f(x) := c^T x + \frac{1}{2} x^T Q x, \quad x \in \mathbb{R}^n,$$

wobei $c \in \mathbb{R}^n$ und $Q \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit ist. Dann gilt: Entweder bricht das auf diese Aufgabe angewandte ungedämpfte Dennis-Wolkowicz-Verfahren nach endlich vielen Schritten mit der eindeutigen Lösung $x^* = -Q^{-1}c$ der Aufgabe ab, oder es liefert eine Folge $\{x_k\}$ die global und Q-superlinear gegen x^* konvergiert.

Beweis: Wegen der positiven Definitheit von Q ist f gleichmäßig konvex und das Verfahren somit durchführbar. Da Broydenklasse-Verfahren invariant unter linearer Variablentransformation sind, können wir $Q^{-1/2}x$ anstelle von x setzen und erhalten als neue, aber äquivalente Zielfunktion

$$f(x) = c^T Q^{-1/2} x + \frac{1}{2} x^T x, \quad \nabla f(x) = Q^{-1/2} c + x, \quad \nabla^2 f(x) = I.$$

Daher gilt dann $y_k = s_k$ und somit $b_k = \|y_k\|^2 = \|s_k\|^2$. Wie schon im Beweis des globalen Konvergenzsatzes 4.1 gilt

$$\begin{aligned}
\psi(B_{k+1}) &= \psi(B_k) + \underbrace{\ln\left(\frac{s_k^T B_{k+\frac{1}{2}} s_k}{\|s_k\| \|B_{k+\frac{1}{2}} s_k\|}\right)^2}_{\leq 0} + \underbrace{\left[1 - \frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k} + \ln \frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k}\right]}_{\leq 0} \\
&\quad + \underbrace{\left[1 - \frac{\|y_k\|^2}{a_k} + \ln \frac{\|y_k\|^2}{a_k}\right]}_{\leq 0} + 2 \underbrace{\left(\frac{\|y_k\|^2}{b_k} - 1\right)}_{=0} - \underbrace{\ln \frac{\|y_k\|^2}{\|s_k\|^2}}_{=1} \\
&\leq \psi(B_k).
\end{aligned}$$

Also ist die Folge $\{\psi(B_k)\}$ monoton fallend und beschränkt durch

$$n \leq \psi(B_k) \leq \psi(B_0),$$

also konvergent. Wie im Beweis von Satz 4.4 folgt hieraus die Beschränktheit der Folgen $\{\|B_k\|\}$ und $\{\|B_k^{-1}\|\}$. Andererseits ist wegen der Konvergenz von $\{\psi(B_k)\}$ die Folge $\{\psi(B_{k+1})-\psi(B_k)\}$ eine Nullfolge und daher notwendigerweise

$$\lim_{k\to\infty} \ln\left(\frac{s_k^T B_{k+\frac{1}{2}} s_k}{\|s_k\|\|B_{k+\frac{1}{2}} s_k\|}\right)^2 = 0, \quad \lim_{k\to\infty}\left[1 - \frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k} + \ln \frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k}\right] = 0$$

und

$$\lim_{k\to\infty}\left[1 - \frac{\|y_k\|^2}{a_k} + \ln \frac{\|y_k\|^2}{a_k}\right] = 0.$$

Hieraus folgt

$$\lim_{k\to\infty} \frac{s_k^T B_{k+\frac{1}{2}} s_k}{\|s_k\|\|B_{k+\frac{1}{2}} s_k\|} = 1, \quad \lim_{k\to\infty} \frac{\|B_{k+\frac{1}{2}} s_k\|^2}{s_k^T B_{k+\frac{1}{2}} s_k} = 1, \quad \lim_{k\to\infty} \frac{\|y_k\|^2}{a_k} = 1.$$

Wieder erhalten wir durch Multiplikation der Grenzwerte

$$\lim_{k\to\infty} \frac{\|B_{k+\frac{1}{2}} s_k\|}{\|s_k\|} = 1 \quad \text{und} \quad \lim_{k\to\infty} \frac{s_k^T B_{k+\frac{1}{2}} s_k}{\|s_k\|^2} = 1$$

und folgern

$$\frac{\|(B_{k+\frac{1}{2}} - I)s_k\|^2}{\|s_k\|^2} = \underbrace{\frac{\|B_{k+\frac{1}{2}} s_k\|^2}{\|s_k\|^2}}_{\to 1} - 2\underbrace{\frac{s_k^T B_{k+\frac{1}{2}} s_k}{\|s_k\|^2}}_{\to 1} + 1 \longrightarrow 0.$$

Hiermit können wir analog zum Beweis von Satz 4.4 fortfahren. Es gilt

$$\begin{aligned}\frac{\|(B_k - I)s_k\|}{\|s_k\|} &\leq \frac{\|(B_{k+\frac{1}{2}} - I)s_k\|}{\|s_k\|} + \frac{1}{\|s_k\|}\left\|\frac{a_k - b_k}{a_k}\frac{y_k y_k^T}{y_k^T s_k} s_k\right\| \\ &= \underbrace{\frac{\|(B_{k+\frac{1}{2}} - I)s_k\|}{\|s_k\|}}_{\to 0} + \frac{|a_k - b_k|}{a_k}\underbrace{\frac{\|y_k\|}{\|s_k\|}}_{=1}.\end{aligned}$$

Zum Nachweis der Konvergenz des zweiten Summanden gegen Null betrachten wir

$$\left|\frac{a_k - b_k}{a_k}\right| = \left|1 - \frac{b_k}{a_k}\right| = \left|1 - \underbrace{\frac{\|y_k\|^2}{a_k}}_{\to 1}\right| \longrightarrow 0.$$

Damit haben wir

$$\lim_{k\to\infty} \frac{\|(B_k - I)s_k\|}{\|s_k\|} = 0$$

gezeigt. Ferner gilt

$$x_{k+1} = x_k - B_k^{-1}\nabla f(x_k) = x_k - B_k^{-1}(Q^{-1/2}c + x_k) = x_k - B_k^{-1}(x_k - x^*).$$

Daher ist $x_k - x^* = -B_k(x_{k+1} - x_k)$ und somit

$$\begin{aligned} x_{k+1} - x^* &= (x_{k+1} - x_k) + (x_k - x^*) \\ &= (x_{k+1} - x_k) - B_k(x_{k+1} - x_k) \\ &= (I - B_k)s_k. \end{aligned}$$

Damit erhalten wir

$$\begin{aligned} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} &= \frac{\|(B_k - I)s_k\|}{\|B_k(x_{k+1} - x_k)\|} = \frac{\|B_k^{-1}\|\|(B_k - I)s_k\|}{\|B_k^{-1}\|\|B_k s_k\|} \\ &\leq \|B_k^{-1}\|\frac{\|(B_k - I)s_k\|}{\|s_k\|} \longrightarrow 0, \end{aligned}$$

da die Folge $\{B_k^{-1}\}$ beschränkt ist. Die Folge $\{x_k\}$ konvergiert also superlinear gegen x^*. □

Auch bei strikt konvexer quadratischer Zielfunktion ist es also möglich, das globale Konvergenzergebnis des ungedämpften BFGS-Verfahrens auf das DW-Verfahren zu übertragen. Gleiches gilt auch für den Konvergenzsatz zum BFGS-Verfahren mit "cautious" Update. Dies wird Inhalt des nächsten Abschnitts sein.

4.4 Das DW-Verfahren mit "cautious" Update

Natürlich stellt sich auch für das Dennis-Wolkowicz-Verfahren die Frage nach dem Konvergenzverhalten bei einer nichtkonvexen Zielfunktion. Im Gegensatz zum BFGS-Verfahren können wir hierauf keine genaue Antwort geben. Die im nachfolgenden Kapitel beschriebenen numerischen Tests lassen zwar vermuten, daß das gedämpfte Dennis-Wolkowicz-Verfahren auch bei nichtkonvexer Zielfunktion global konvergiert, gesichert ist dies jedoch nicht. Desweiteren können wir das von Y.-H. Dai konsturierte Gegenbeispiel aus Satz 3.3 hier nicht anwenden, da es nur auf Broydenklasse-Verfahren mit $\phi_k \leq 1$ erweitert werden kann. Daher beschränken wir uns in den Ausführungen auf eine Modifikation des Verfahrens, durch welche wir globale Konvergenz bei nichtkonvexer Zielfunktion beweisen können. Wir verwenden erneut die Idee von D.-H. Li und M. Fukushima und übertragen ihr "cautious" Update für das BFGS-Verfahren aus [15] auf das DW-Verfahren. Mit gegebenen Konstanten $\alpha, \epsilon > 0$ hat dies dann die folgende Gestalt:

$$B_{k+1} := \begin{cases} B_{k+\frac{1}{2}} - \dfrac{(B_{k+\frac{1}{2}}s_k)(B_{k+\frac{1}{2}}s_k)^T}{s_k^T B_{k+\frac{1}{2}} s_k} + \dfrac{y_k y_k^T}{y_k^T s_k}, & \text{falls } \dfrac{y_k^T s_k}{\|s_k\|^2} > \epsilon\|g_k\|^\alpha, \\ B_k & \text{sonst.} \end{cases}$$

Hierbei ist $B_{k+\frac{1}{2}}$ mit den üblichen Abkürzungen wieder durch

$$B_{k+\frac{1}{2}} := B_k + \frac{a_k - b_k}{a_k} \frac{y_k y_k^T}{b_k}$$

gegeben. Unser Ziel ist es nun, die globale Konvergenz des hieraus entstehenden Quasi-Newton-Verfahrens bei nichtkonvexer Zielfunktion und Verwendung einer zumindest semi-effizienten Schrittweitenstrategie zu beweisen. Wieder benötigen wir folgende Voraussetzungen:

(V_1) Bei gegebenem Startvektor $x_0 \in \mathbb{R}^n$ des Verfahrens ist die Niveaumenge $L_0 := \{x \in \mathbb{R}^n \ : \ f(x) \leq f(x_0)\}$ kompakt.

(V_2) Die Zielfunktion f ist auf einer offenen Obermenge von L_0 stetig differenzierbar.

(V_3) Der Zielfunktionsgradient $\nabla f(\cdot)$ ist auf L_0 lipschitzstetig. Es existiert also eine Konstante $\gamma > 0$ mit

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

Hiermit können wir nun Satz 3.9 und dessen Beweistechnik auf das Dennis-Wolkowicz-Verfahren übertragen.

Satz 4.6 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Voraussetzungen (V_1) bis (V_3) seien erfüllt. Mit einem $x_0 \in \mathbb{R}^n$ und einer symmetrischen und positiv definiten Matrix $B_0 \in \mathbb{R}^{n \times n}$ erzeuge das gedämpfte DW-Verfahren mit "cautious" Update eine Folge $\{x_k\}$. Ferner gelte (für die gewählte Schrittweitenstrategie) in jedem Iterationsschritt $k = 0, 1, \ldots$*

$$f(x_k) - f(x_k + t_k p_k) \geq \theta \min\left[-\nabla f(x_k)^T p_k, \left(\frac{\nabla f(x_k)^T p_k}{\|p_k\|}\right)^2\right]$$

mit einer Konstanten $\theta > 0$. Dann gilt: Entweder bricht das Verfahren nach endlich vielen Schritten mit einer kritischen Lösung ab, oder für die erzeugte Folge gilt

$$\liminf_{k \to \infty} \|\nabla f(x_k)\| = 0.$$

Beweis: Die Durchführbarkeit des Verfahrens ist gesichert, da das "cautious" Update in jedem Fall die positive Definitheit erhält und im Falle eines DW-Updates $y_k^T s_k > 0$ erfüllt ist. Wir folgen dem Widerspruchsbeweis von Satz 3.9 und nehmen an, daß es keine gegen Null konvergierende Teilfolge von $\{\|\nabla f(x_k)\|\}$ gibt. Es existiere also eine Konstante $\eta > 0$ derart, daß $\|g_k\| = \|\nabla f(x_k)\| \geq \eta$ für alle $k = 0, 1, \ldots$ gilt. Für diese Indices definieren wir erneut

$$\delta_k := \min\left[-\frac{g_k^T p_k}{\|g_k\|^2}, \left(\frac{g_k^T p_k}{\|g_k\| \|p_k\|}\right)^2\right].$$

Unser Ziel ist es jetzt wieder, die Existenz einer Konstanten $\delta > 0$ mit $\delta_k \geq \delta$ für unendlich viele Indices k zu zeigen. Aus der vorausgesetzten Semi-Effizienz der Schrittweitenstrategie folgt

$$f(x_k) - f(x_{k+1}) \geq \theta\eta^2\delta > 0$$

für unendlich viele Iterationsschritte. Dies liefert uns den gewünschten Widerspruch zur Beschränktheit der Zielfunktion auf der Niveaumenge, da in den restlichen Iterationen die erreichte Zielfunktionsverminderung nichtnegativ ist. Wir definieren die Menge derjenigen Iterationsindices, bei denen ein Dennis-Wolkowicz-Update durchgeführt wird als

$$J := \left\{ k \in \mathbb{N}_0 \ : \ \frac{y_k^T s_k}{\|s_k\|^2} \geq \epsilon \|g_k\|^\alpha \right\}$$

und unterscheiden zwei Fälle.

Fall 1: Die Menge J enthält nur endlich viele Elemente. Dann gilt völlig analog zum Beweis von Satz 3.9 in fast allen Iterationschritten $B_k = B$ mit einer symmetrischen und positiv definiten Matrix $B \in \mathbb{R}^{n\times n}$ und daher wieder

$$\delta_k \geq \min\left[\frac{1}{\lambda_{\max}(B)}, \frac{\lambda_{\min}(B)}{\lambda_{\max}(B)}\right] =: \delta > 0$$

für fast alle k. Folglich wird in diesen Iterationsschritten eine gegen Null beschränkte Zielfunktionsverminderung erreicht, und der gesuchte Widerspruch ist gefunden.

Fall 2: Die Indexmenge J ist nicht endlich. Wir definieren wieder

$$J_k := J \cap \{0, 1, \ldots, k\}$$

und erinnern uns, daß $\lim_{k\to\infty} \#(J_k) = \infty$ gilt. Nun wenden wir die ψ-Funktion auf die neue Update-Formel an. Wegen $\psi(B_{k+1}) = \psi(B_k)$ für alle $k \notin J$ gilt dann wie im Beweis von Satz 4.1

$$\begin{aligned} 0 \ &< \ \psi(B_{k+1}) \\ &= \ \psi(B_0) + \sum_{j\in J_k} \left\{ \ln\left(\frac{s_j^T B_{j+\frac{1}{2}} s_j}{\|s_j\|\|B_{j+\frac{1}{2}} s_j\|}\right)^2 + \left[1 - \frac{\|B_{j+\frac{1}{2}} s_j\|^2}{s_j^T B_{j+\frac{1}{2}} s_j} + \ln \frac{\|B_{j+\frac{1}{2}} s_j\|^2}{s_j^T B_{j+\frac{1}{2}} s_j}\right] \\ &\quad + \left[1 - \frac{\|y_j\|^2}{a_j} + \ln\frac{\|y_j\|^2}{a_j}\right] + 2\left(\frac{\|y_j\|^2}{b_j} - 1\right) - \ln\frac{\|y_j\|^2}{\|s_j\|^2} \right\}. \end{aligned}$$

Wir schätzen nun die letzten beiden Argumente in der Summe gegen Konstanten ab. Für $j \in J_k$ erhalten wir aus aus dem "cautious" Update und unserer anfänglichen Annahme

$$\frac{y_j^T s_j}{\|s_j\|^2} \geq \epsilon\|g_j\|^\alpha \geq \epsilon\eta^\alpha$$

und daraus mit der Voraussetzung (V_3)

$$\frac{\|y_j\|^2}{b_j} \leq \frac{\gamma^2\|s_j\|^2}{y_j^T s_j} \leq \frac{\gamma^2}{\epsilon\eta^\alpha}.$$

Ferner folgt aus $\|y_j\|\|s_j\| \geq y_j^T s_j$ die Relation

$$\frac{\|y_j\|}{\|s_j\|} \geq \frac{y_j^T s_j}{\|s_j\|^2} \geq \epsilon\eta^\alpha.$$

Folglich gilt

$$\sum_{j\in J_k}\left\{2\left(\frac{\|y_j\|^2}{b_j}-1\right)-\ln\frac{\|y_j\|^2}{\|s_j\|^2}\right\} \leq \#(J_k)\left[2\left(\frac{\gamma^2}{\epsilon\eta^\alpha}-1\right)-2\ln(\epsilon\eta^\alpha)\right].$$

Hieraus folgt

$$\sum_{j\in J_k}\ln\left(\frac{\|s_j\|\|B_{j+\frac{1}{2}}s_j\|}{s_j^T B_{j+\frac{1}{2}}s_j}\right)^2 + h\left(\frac{\|B_{j+\frac{1}{2}}s_j\|^2}{s_j^T B_{j+\frac{1}{2}}s_j}\right) + h\left(\frac{\|y_j\|^2}{a_j}\right) \leq C\#(J_k)$$

mit einer von k unabhängigen Konstanten $C > 0$ und der schon oft verwendeten auf $(0,\infty)$ definierten Funktion $h(t) = t - 1 - \ln t$. Nun können wir wieder Lemma 3.8 anwenden und erhalten die Existenz einer Indexmenge $J_k^* \subset J_k$ mit $\#(J_k^*) \geq \frac{1}{2}\#(J_k)$ und

$$\ln\left(\frac{\|s_j\|\|B_{j+\frac{1}{2}}s_j\|}{s_j^T B_{j+\frac{1}{2}}s_j}\right)^2 \leq 2C, \quad h\left(\frac{\|B_{j+\frac{1}{2}}s_j\|^2}{s_j^T B_{j+\frac{1}{2}}s_j}\right) \leq 2C, \quad h\left(\frac{\|y_j\|^2}{a_j}\right) \leq 2C$$

für alle $j \in J_k^*$. Analog zum Beweis von Satz 4.1 existieren daher Konstanten $C_0 \in (0,1)$ und $C_1 \in (1,\infty)$ mit

$$C_0 \leq \frac{\|B_{j+\frac{1}{2}}s_j\|^2}{s_j^T B_{j+\frac{1}{2}}s_j} \leq C_1 \quad \text{und} \quad C_0 \leq \frac{\|y_j\|^2}{a_j} \leq C_1$$

für alle $j \in J_k^*$. Hieraus wollen wir nun die Existenz zweier Konstanten $\hat{\delta}, \bar{\delta} > 0$ mit

$$-\frac{g_j^T p_j}{\|g_j\|^2} \geq \hat{\delta} \quad \text{und} \quad \left(\frac{g_j^T p_j}{\|g_j\|\|p_j\|}\right)^2 \geq \hat{\delta}$$

für alle $j \in J_k^*$ folgern. Hierzu folgen wir weiterhin der Argumentation des globalen Konvergenzbeweises, müssen jedoch auf die Verwendung der gleichmäßigen Konvexitätsvoraussetzung verzichten. Für alle $j \in J_k^*$ gilt erneut

$$\frac{\|B_j s_j\|}{\sqrt{c_j}} \leq \sqrt{\frac{C_1}{\phi_j}} + \frac{|a_j - b_j|}{a_j\sqrt{c_j}}\|y_j\|.$$

Wegen $b_j^2 \leq a_j c_j$ und $b_j \geq \epsilon\eta^\alpha \|s_j\|^2$ gilt für alle $j \in J_k^*$

$$C_0 \leq \frac{\|y_j\|^2}{a_j} \leq \frac{\gamma^2\|s_j\|^2}{a_j} \leq \frac{\gamma^2 c_j \|s_j\|^2}{b_j^2} \leq \left(\frac{\gamma}{\epsilon\eta^\alpha}\right)^2 \frac{c_j}{\|s_j\|^2}$$

und somit

$$\frac{\|s_j\|}{\sqrt{c_j}} \leq \frac{\gamma}{\epsilon\eta^\alpha\sqrt{C_0}}$$

für alle $j \in J_k^*$. Für die weitere Betrachtung untersuchen wir wieder mehrere Unterfälle bei festem $j \in J_k^*$.

Fall 2a: Es gilt $\phi_j \geq 1$. Dann gilt wieder $0 < a_j \leq b_j$ und damit wie schon zuvor

$$\frac{\|B_j s_j\|}{\sqrt{c_j}} \leq \sqrt{C_1} + \frac{\sqrt{c_j}}{b_j}\|y_j\| \leq \sqrt{C_1} + \frac{\gamma\sqrt{c_j}}{b_j}\|s_j\| \leq \sqrt{C_1} + \frac{\gamma}{\epsilon\eta^\alpha}\frac{\sqrt{c_j}}{\|s_j\|}.$$

Nach Multiplikation mit $\|s_j\|/\sqrt{c_j}$ und Verwendung der obigen Abschätzung hierfür ergibt sich

$$\frac{\|g_j\|\|p_j\|}{(-g_j^T p_j)} = \frac{\|B_j s_j\|\|s_j\|}{c_j} \leq \sqrt{C_1}\frac{\|s_j\|}{\sqrt{c_j}} + \frac{\gamma}{\epsilon\eta^\alpha} \leq \left(\sqrt{\frac{C_1}{C_0}} + 1\right)\frac{\gamma}{\epsilon\eta^\alpha}$$

und damit

$$\left(\frac{g_j^T p_j}{\|g_j\|\|p_j\|}\right)^2 \geq \left[\left(\sqrt{\frac{C_1}{C_0}} + 1\right)\frac{\gamma}{\epsilon\eta^\alpha}\right]^{-2} =: \bar{\delta}_1.$$

Andererseits gilt auch wieder

$$\frac{\|B_j s_j\|}{\sqrt{c_j}} \leq \sqrt{C_1} + C_1\frac{\|s_j\|}{\sqrt{c_j}} \leq \sqrt{C_1} + \frac{\gamma C_1}{\epsilon\eta^\alpha\sqrt{C_0}}$$

und deshalb

$$-\frac{g_j^T p_j}{\|g_j\|^2} = \frac{c_j}{\|B_j s_j\|^2} \geq \left[\sqrt{C_1} + \frac{\gamma C_1}{\epsilon\eta^\alpha\sqrt{C_0}}\right]^{-2} =: \hat{\delta}_1.$$

Fall 2b: Es gilt $\phi_j < 1$ und damit $b_j < a_j$. Daher gilt hier erneut

$$\frac{\|B_j s_j\|}{\sqrt{c_j}} \leq \sqrt{\frac{C_1}{\phi_j}} + \gamma\frac{\|s_j\|}{\sqrt{c_j}} \leq \sqrt{\frac{C_1}{\phi_j}} + \frac{\gamma^2}{\epsilon\eta^\alpha\sqrt{C_0}}.$$

Ab hier müssen wir wieder zwei Unterfälle unterscheiden.

Fall 2b-1: Es gilt $\phi_j < 1$ und $c_j \geq b_j$. Dann ist wieder $\phi_j \geq \frac{1}{2}$ und damit

$$\frac{\|B_j s_j\|}{\sqrt{c_j}} \leq \sqrt{2C_1} + \frac{\gamma^2}{\epsilon\eta^\alpha\sqrt{C_0}}.$$

Hieraus folgern wir

$$\frac{\|g_j\|\|p_j\|}{(-g_j^T p_j)} = \frac{\|B_j s_j\|\|s_j\|}{c_j} \leq \left(\sqrt{2C_1} + \frac{\gamma^2}{\epsilon\eta^\alpha\sqrt{C_0}}\right)\frac{\gamma}{\epsilon\eta^\alpha\sqrt{C_0}}$$

und erhalten somit

$$\left(\frac{g_j^T p_j}{\|g_j\|\|p_j\|}\right)^2 \geq \left[\left(\sqrt{2C_1} + \frac{\gamma^2}{\epsilon\eta^\alpha\sqrt{C_0}}\right)\frac{\gamma}{\epsilon\eta^\alpha\sqrt{C_0}}\right]^{-2} =: \bar{\delta}_2.$$

Außerdem folgt

$$-\frac{g_j^T p_j}{\|g_j\|^2} = \frac{c_j}{\|B_j s_j\|^2} \geq \left[\sqrt{2C_1} + \frac{\gamma^2}{\epsilon\eta^\alpha\sqrt{C_0}}\right]^{-2} =: \hat{\delta}_2.$$

Fall 2b-2: Es gilt $\phi_j < 1$ und $c_j < b_j$ und somit

$$\phi_j \geq \frac{c_j}{2\gamma\|s_j\|^2}.$$

Einsetzen ergibt dann

$$\frac{\|B_j s_j\|}{\sqrt{c_j}} \leq \sqrt{2\gamma C_1}\frac{\|s_j\|}{\sqrt{c_j}} + \frac{\gamma^2}{\epsilon\eta^\alpha\sqrt{C_0}} \leq \left(\sqrt{2\gamma C_1} + \gamma\right)\frac{\gamma}{\epsilon\eta^\alpha\sqrt{C_0}}.$$

Folglich gilt

$$\frac{\|g_j\|\|p_j\|}{(-g_j^T p_j)} = \frac{\|B_j s_j\|\|s_j\|}{c_j} \leq \left(\sqrt{2\gamma C_1} + \gamma\right)\frac{\gamma^2}{\epsilon^2\eta^{2\alpha}C_0}$$

und daher

$$\left(\frac{g_j^T p_j}{\|g_j\|\|p_j\|}\right)^2 \geq \left[\left(\sqrt{2\gamma C_1} + \gamma\right)\frac{\gamma^2}{\epsilon^2\eta^{2\alpha}C_0}\right]^{-2} =: \bar{\delta}_3.$$

Desweiteren folgt

$$-\frac{g_j^T p_j}{\|g_j\|^2} = \frac{c_j}{\|B_j s_j\|^2} \geq \left[\left(\sqrt{2\gamma C_1} + \gamma\right)\frac{\gamma}{\epsilon\eta^\alpha\sqrt{C_0}}\right]^{-2} =: \hat{\delta}_3.$$

Zusammen gilt also für $j \in J_k^*$

$$-\frac{g_j^T p_j}{\|g_j\|^2} \geq \min(\hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3) =: \hat{\delta} \quad \text{und} \quad \left(\frac{g_j^T p_j}{\|g_j\|\|p_j\|}\right)^2 \geq \min(\bar{\delta}_1, \bar{\delta}_2, \bar{\delta}_3) =: \bar{\delta}.$$

Daher gilt dann für alle $j \in J_k^*$

$$\delta_j \geq \min(\hat{\delta}, \bar{\delta}) =: \delta > 0.$$

Dies liefert uns wieder die gewünschte gegen Null beschränkte Zielfunktionsverminderung, und wir können den Beweis analog zu dem von Satz 3.9 abschließen. Es gilt

$$\begin{aligned} f(x_0) - f(x_{k+1}) &= \sum_{j=0}^{k} f(x_j) - f(x_{j+1}) \geq \sum_{j=0}^{k} \theta\eta^2\delta_j \geq \theta\eta^2 \sum_{j\in J_k^*} \delta_j \\ &\geq \theta\eta^2\delta \#(J_k^*) \geq \frac{\theta\eta^2\delta}{2} \#(J_k) \longrightarrow \infty. \end{aligned}$$

Der gesuchte Widerspruch ist gefunden. □

Wir haben also auch für das Dennis-Wolkowicz-Verfahren mit "cautious" Update und semi-effizienter Schrittweitenstrategie die Konvergenz einer Teilfolge von $\{x_k\}$ gegen einen stationären Punkt der nichtkonvexen Zielfunktion bewiesen. Wegen der geringen Voraussetzungen muss es sich bei diesem stationären Punkt allerdings nicht um ein lokales Minimum handeln. Falls wir die Zielfunktion entgegen der ursprünglichen Annahme wieder als konvex voraussetzen, überträgt sich auch Korollar 3.10 auf das Dennis-Wolkowicz-Verfahren mit "cautious" Update. Aus Vollständigkeitsgründen geben wir das entsprechende Resultat jetzt an.

Korollar 4.7 *Gegeben sei die unrestringierte Aufgabe (P). Die Voraussetzungen (V_1) bis (V_3) seien erfüllt. Mit $x_0 \in \mathbb{R}^n$ und $B_0 \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit startend sei die Folge $\{x_k\}$ durch das DW-Verfahren mit "cautious" Update und semi-effizienter Schrittweitenstrategie erzeugt. Dann gilt: Ist die Zielfunktion f konvex, so konvergiert die ganze Folge $\{g_k\}$ gegen Null und jeder Häufungspunkt von $\{x_k\}$ ist eine globale Lösung von (P).*

Für das CBFGS-Verfahren haben D.-H. Li und M. Fukushima den in dieser Arbeit angegebenen Satz 3.11 bewiesen. Da im zugehörigen Beweis keine verfahrensspezifischen Aussagen benötigt werden, ist dieser auch für das Dennis-Wolkowicz-Verfahren mit "cautious" Update gültig.

Satz 4.8 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Die Voraussetzungen (V_1) bis (V_3) seien erfüllt. Mit einem $x_0 \in \mathbb{R}^n$ und $B_0 \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit erzeuge das durch eine semi-effiziente Schrittweitenstrategie gedämpfte Dennis-Wolkowicz-Verfahren mit "cautious" Update eine Folge $\{x_k\}$. Ferner sei die Zielfunktion f zweimal stetig differenzierbar und es gelte $s_k \to 0$. Dann gilt: Falls ein Häufungspunkt $x^* \in \mathbb{R}^n$ von $\{x_k\}$ mit $\nabla f(x^*) = 0$ und $\nabla^2 f(x^*)$ symmetrisch und positiv definit existiert, dann konvergiert $\{x_k\}$ gegen x^*, und das Verfahren geht nach endlich vielen Iterationsschritten in das unveränderte Dennis-Wolkowicz-Verfahren über.*

Falls das Dennis-Wolkowicz-Verfahren mit "cautious" Update also gegen ein isoliertes lokales Minimum der Zielfunktion konvergiert, folgt der Übergang zu einem unveränderten Dennis-Wolkowicz-Verfahren. Unter den entsprechenden Voraussetzungen gelten dann die Konvergenzaussagen der Abschnitte 4.1 und 4.2.

Insgesamt ist es uns in diesem Kapitel also gelungen, fast alle Konvergenzresultate zum BFGS-Verfahrens aus [22], beziehungsweise Kapitel 3 auch für das Dennis-Wolkowicz-Verfahren zu beweisen. Einzige Aussnahme ist hierbei ein dem Konvergenzsatz von POWELL entsprechendes Resultat. Damit beenden wir das Kapitel über die theoretischen Eigenschaften des Dennis-Wolkowicz-Verfahrens und beginnen mit der Untersuchung der numerischen Eigenschaften.

Kapitel 5

Numerische Tests

In diesem Kapitel werden wir die numerischen Eigenschaften des DW-Verfahrens untersuchen und mit denen des BFGS-Verfahrens vergleichen. Zwar haben J. E. Dennis Jr. und H. Wolkowicz in [9] bereits numerische Experimente mit denen von ihnen neu vorgestellten und einigen älteren Verfahren vorgenommen, die Darstellung ihrer Testergebnisse beschränkt sich jedoch auf die Angabe von Durchschnittswerten und einer Rangfolge der Verfahren bezüglich ihrer Testkriterien. Diese sind Anzahl der benötigten Iterationen, Anzahl der benötigten Zielfunktionsauswertungen und Fehleranfälligkeit des Verfahrens. Die Tests erfolgen in drei Gruppen, die sich durch die gewählte Schrittweitenstrategie unterscheiden. Die drei Möglichkeiten hierfür sind Armijo- und Wolfe-Schrittweite sowie ungedämpfte Verfahren. Für uns sind folgende Ergebnisse aus [9] von Interesse:

- In allen drei Schrittweiten-Testgruppen ist das DW-Verfahren mit initial inverse sizing den anderen Verfahren überlegen und belegt jeweils den ersten Platz in den Kriterien Anzahl der Iterationen und Anzahl der Zielfunktionsauswertungen.

- In den beiden Testgruppen mit Dämpfung belegt ein ebenfalls neu vorgestelltes, bisher noch namenloses Broydenklasse-Verfahren mit initial inverse sizing und dem Kürzel "optimal ϕ" jeweils den zweiten Platz bei den Kriterien Anzahl der Iterationen und Anzahl der Zielfunktionsauswertungen.

- Das klassische BFGS-Verfahren ohne sizing belegt bei den Tests in allen Testgruppen und bei allen Kriterien nur hintere Plätze.

Besonders das überaus schlechte Abschneiden des BFGS-Verfahrens läßt die numerischen Untersuchungen von Dennis und Wolkowicz etwas fragwürdig erscheinen. Schließlich handelt es sich um das in der Praxis meistbenutzte Verfahren für niedrigdimensionale unrestringierte Optimierunsaufgaben. Um uns einen objektiven Eindruck über die numerischen Eigenschaften der Verfahren machen

zu können, führen wir eigene Tests mit ihnen durch. Wegen des guten Abschneidens des "optimal ϕ"-Verfahrens in den Tests aus [9] werden wir es bei unseren numerischen Experimenten berücksichtigen. Auf eine genauere Betrachtung der theoretischen Eigenschaften dieses Verfahrens müssen wir verzichten, da bislang noch keine Konvergenzsätze hierfür bekannt sind. Eine etwas ausführlichere Darstellung erfolgt später.

5.1 Implementation der Verfahren

Bei allen drei zu implementierenden Verfahren handelt es sich um Quasi-Newton-Verfahren der Broydenklasse. Die Algorithmen setzen sich, wie in der Einleitung schon beschrieben, aus einer Richtungs- und einer Schrittweitenstrategie zusammen, wobei die Richtung aus einem linearen Gleichungssystem mit der Update-Matrix bestimmt wird. Im Unterschied zu dem bereits angegebenen Modellalgorithmus brechen wir das Verfahren ab, falls der Zielfunktionsgradient hinreichend nahe bei Null ist oder die Anzahl der Iterationen einen Maximalwert überschreitet. Genauer ist unser Abbruchkriterium

$$\|\nabla f(x_k)\| \leq 10^{-6} \qquad \text{oder} \qquad k > 1000.$$

Die Verfahren haben damit die folgende Gestalt:

- Gegeben sei ein Startvektor $x_0 \in \mathbb{R}^n$, eine symmetrische und positiv definite Startmatrix $B_0 \in \mathbb{R}^{n\times n}$ sowie eine Toleranz für die Gradientennorm $\epsilon > 0$ und eine maximale Iterationszahl $k_{\max} \in \mathbb{N}$. Berechne $g_0 := \nabla f(x_0)$.
- Für $k = 0, 1, \ldots$:
  - Falls $\|g_k\| \leq \epsilon$ oder $k > k_{\max}$, dann STOP.
  - Andernfalls:
    1. Bestimme die Abstiegsrichtung $p_k \in \mathbb{R}^n$ aus $B_k p_k = -g_k$.
    2. Bestimme eine Schrittweite $t_k > 0$ mit $f(x_k + t_k p_k) < f(x_k)$.
    3. Bestimme die neue Näherung $x_{k+1} := x_k + t_k p_k$ und den zugehörigen Zielfunktionsgradienten $g_{k+1} := \nabla f(x_{k+1})$.
    4. Bestimme $B_{k+1} \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit.

In unserem Fall werden wir t_k stets als Wolfe-Schrittweite und B_{k+1} als BFGS-, DW- oder "optimal ϕ"-Update von B_k berechnen. Wir werden neben $\epsilon := 10^{-6}$ und $k_{\max} := 1000$ stets

$$B_0 := |f(x_0)|I$$

mit der $n \times n$ Einheitsmatrix I setzen. Dies ist im allgemeinen eine geschicktere Wahl als die sonst übliche $B_0 = I$. Der Vorschlag hierzu stammt von J. E.

Dennis Jr. und R. Schnabel (siehe [9], S. 209f.) und wird von ihnen an zwei Beispielen demonstriert. Wir haben dies auch an den noch zu nennenden Zielfunktionen getestet und sind zu dem selben Ergebnis gekommen. Nun gehen wir näher auf die Implementation der einzelnen Updates ein. Als Plattform hierfür werden wir Matlab benutzen.

5.1.1 Das BFGS-Verfahren

Das BFGS-Update ist gegeben durch

$$B_{k+1} := B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}$$

und ist symmetrisch und positiv definit, falls B_k es ist und $y_k^T s_k > 0$ gilt. Dies wird in den Tests durch Verwendung der Wolfe-Schrittweite sichergestellt. Da die Abstiegsrichtung p_k aus einem linearen Gleichungssystem mit positiv definitem B_k als Koeffizientenmatrix bestimmt wird, ist es von Vorteil, wenn für diese Matrix eine Choleskyzerlegung

$$B_k = L_k L_k^T$$

mit einer unteren Dreiecksmatix mit positiven Diagonalelementen $L_k \in \mathbb{R}^n$ vorliegt. Dann reduziert sich nämlich der Aufwand zur Lösung des Systems von $O(n^3)$ auf $n(n+1)/2$ beziehungsweise $O(n^2)$ Operationen, da es durch einfaches Vorwärts- und Rückwärtseinsetzen gelöst werden kann. Nun ist aber die Berechnung einer Cholesky-Zerlegung von B_k in jedem Iterationsschritt mit einem Aufwand von $O(n^3)$ Operationen verbunden, was den eben beschriebenen numerischen Vorteil wieder zunichte machen würde. Daher liegt es nahe, anstelle der Update-Matrizen selbst nur ihre Cholesky-Faktoren upzudaten, also den neuen Cholesky-Faktor L_{k+1} für B_{k+1} aus dem alten L_k für B_k zu berechnen. Eine genaue Beschreibung eines solchen Vorgehens ist bei J. Werner ([23], S. 201f.) zu finden. Der Aufwand für ein derartiges Cholesky-Update beträgt im wesentlichen $O(n^2)$ Operationen, so daß der gesamte Aufwand für ein Matrix-Update und anschließende Richtungsberechnung auch nur $O(n^2)$ beträgt. Ein alternatives Vorgehen mit etwa gleichem Rechenaufwand ist die Implementierung des Updates von B_k^{-1} auf B_{k+1}^{-1}. Dies bringt den Vorteil mit sich, daß p_k durch eine Matrix-Vektor-Multiplikation berechnet werden kann. Nachteilig ist jedoch, daß man keine Kontrolle über den Erhalt der positiven Definitheit der Update-Matrizen wie beim Cholesky-Update hat. Wir entscheiden uns daher für eine Implementation mit Cholesky-Update. Matlab stellt dafür den Befehl `cholupdate` zur Verfügung. Dieser berechnet aus einer gegebenen Cholesky-Zerlegung einer symmetrischen, positiv definiten Matrix $A \in \mathbb{R}^{n \times n}$ und einem Vektor $v \in \mathbb{R}^n$ einen Cholesky-Faktor für das Rang-1-Update

$$A_+ := A + vv^T$$

in $O(n^2)$ Operationen. Dabei ist zu beachten, daß MATLAB anstelle der sonst üblichen Darstellung der Cholesky-Zerlegung $A = LL^T$ mit einer unteren Dreiecksmatrix mit positiven Diagonalelementen die Darstellung $A = R^T R$ mit $R = L^T$ wählt. Ein Cholesky-Update von

$$A_+ := A - vv^T$$

ist mit `cholupdate` ebenso möglich, falls die neue Matrix A_+ positiv definit ist. Andernfalls wird die Ausführung des Befehls mit einer Fehlermeldung abgebrochen. Wir können also ein BFGS-Update durch zweifache Anwendung von `cholupdate` durchführen. Bei gegebenem Cholesky-Faktor R_k von B_k hat dies die formale Gestalt

$$\begin{aligned} R_{k+\frac{1}{2}}^T R_{k+\frac{1}{2}} &= R_k^T R_k + \frac{y_k y_k^T}{y_k^T s_k}, \\ R_{k+1}^T R_{k+1} &= R_{k+\frac{1}{2}}^T R_{k+\frac{1}{2}} - \frac{(R_k^T R_k s_k)(R_k^T R_k s_k)^T}{\|R_k s_k\|^2}. \end{aligned}$$

Für die Benutzung von `cholupdate` werden die positiven skalaren Faktoren der Rang-1-Matrizen jeweils in zwei Faktoren für die erzeugenden Vektoren dieser Matrizen zerlegt. Bei Verwendung der Wolfe-Schrittweite ist nach Satz 2.1 der Erhalt der positiven Definitheit beim BFGS-Update gesichert und obige Vorgehensweise für das Update möglich. Wir führen die Addition der Rang-1-Matrix zuerst aus, um dies nicht zu gefährden. Eine elementare Implementation des Cholesky-Updates mit Givens-Rotationen gemäß J. WERNER [23], S. 201f. ist ebenfalls möglich. Da jedoch nicht feststellbar ist, ob ein solches Vorgehen effizienter ist, verlassen wir uns auf die numerischen Qualitäten von MATLAB und verwenden die beschriebene Variante mit `cholupdate`.

5.1.2 Das Dennis-Wolkowicz-Verfahren

Die Vorgehensweise beim Dennis-Wolkowicz-Verfahren ist weitgehend analog zu der beim BFGS-Verfahren. Da sowohl beim inversen schwachen Greenstadt-Update

$$B_{k+\frac{1}{2}} := B_k + \frac{y_k^T B_k^{-1} y_k - y_k^T s_k}{(y_k^T B_k^{-1} y_k)(y_k^T s_k)} y_k y_k^T$$

als auch beim hierauf angewandten BFGS-Update

$$B_{k+1} := B_{k+\frac{1}{2}} - \frac{(B_{k+\frac{1}{2}} s_k)(B_{k+\frac{1}{2}} s_k)^T}{s_k^T B_{k+\frac{1}{2}} s_k} + \frac{y_k y_k^T}{y_k^T s_k}$$

unter der Bedingung $y_k^T s_k > 0$ die positive Definitheit erhalten bleibt, können wir auch hier `cholupdate` einsetzen. Der wesentliche Aufwand zur Berechnung

der drei benötigenten Updates und der Abstiegsrichtung beträgt dann wieder $O(n^2)$ Operationen. Formal läßt sich das Update wie folgt darstellen:

$$\begin{aligned} R_{k+\frac{1}{3}}^T R_{k+\frac{1}{3}} &= R_k^T R_k + \frac{a_k - b_k}{a_k b_k} y_k y_k^T, \\ R_{k+\frac{2}{3}}^T R_{k+\frac{2}{3}} &= R_{k+\frac{1}{3}}^T R_{k+\frac{1}{3}} + \frac{y_k y_k^T}{b_k}, \\ R_{k+1}^T R_{k+1} &= R_{k+\frac{2}{3}}^T R_{k+\frac{2}{3}} - \frac{(R_{k+\frac{1}{3}}^T R_{k+\frac{1}{3}} s_k)(R_{k+\frac{1}{3}}^T R_{k+\frac{1}{3}} s_k)^T}{\| R_{k+\frac{1}{3}} s_k \|^2}. \end{aligned}$$

Zur Abkürzung haben wir wieder

$$a_k = y_k^T (R_k^T R_k)^{-1} y_k \quad \text{und} \quad b_k = y_k^T s_k$$

benutzt. Bei der Umsetzung in MATLAB werden wir selbstverständlich auf die Berechnung einer Inversen verzichten und stattdessen das lineare Gleichungssystem $R_k^T x = y_k$ durch Vorwärtseinsetzen lösen. Mit der Lösung $x \in \mathbb{R}^n$ ist dann $a_k = \|x\|^2$. Die Faktoren vor den Rang-1-Matrizen werden wieder zerlegt. Da der erste Faktor aber auch negativ werden kann, ist hier besondere Vorsicht geboten. Wir verwenden daher nur seinen Absolutbetrag und entscheiden anhand des Vorzeichens ob die Matrix addiert oder subtrahiert wird. Vor dem ersten Cholesky-Update wird ein initial inverse sizing ausgeführt. Die ursprüngliche Startmatrix beziehungsweise ihr Cholesky-Faktor

$$R_0 := \sqrt{|f(x_0)|} I$$

wird dabei vor dem ersten Update durch

$$R_0 := \sqrt{\frac{a_0}{b_0}} R_0$$

ersetzt. Nun kommen wir zum "optimal ϕ"-Verfahren.

5.1.3 Das "optimal ϕ"-Verfahren

Das "optimal ϕ"-Verfahren wurde ebenfalls von DENNIS und WOLKOWICZ in [9] vorgestellt und als Verfahren Nr. 6 numerisch getestet. Es ist ein Broydenklasse-Verfahren, dessen Parameter ϕ_k als Lösung der Aufgabe

$$\text{Minimiere} \quad \omega(B_k^{-1/2} B_{k+1}(\phi) B_k^{-1/2}), \quad \phi \in \mathbb{R}$$

bestimmt wird. Hierbei ist die auf der Menge der symmetrischen und positiv definiten Matrizen definierte Funktion $\omega : \mathbb{R}^{n \times n} \longrightarrow \mathbb{R}$ gegeben durch

$$\omega(A) := \frac{\frac{1}{n} \mathrm{tr}\,(A)}{\det(A)^{1/n}}.$$

Ferner bezeichet $B_{k+1}(\phi)$ das Bryodenklasse-Update mit dem Parameter ϕ von der gegebenen Matrix B_k. Mit den üblichen Abkürzungen ist die Lösung des obigen Problems nach Satz 3.1 aus [9] gegeben durch

$$\phi_k^* := 1 + \frac{(a_k - b_k)b_k}{(1-n)(a_k c_k - b_k^2)}.$$

Für $n \geq 2$ gilt dann

$$\phi_k^* < \frac{a_k c_k}{a_k c_k - b_k^2}.$$

Daher ist nach einer Bemerkung von R. H. Byrd, D. C. Liu und J. Nocedal aus [2] das "optimal ϕ"-Update

$$B_{k+1} := B_k - \frac{(B_k s_k)(B_k s_k)^T}{c_k} + \frac{y_k y_k^T}{b_k} - \frac{(a_k - b_k)b_k c_k}{(1-n)(a_k c_k - b_k^2)} v_k v_k^T$$

mit $v_k := y_k/b_k - B_k s_k/c_k$ symmetrisch und positiv definit, falls B_k dies ist und $y_k^T s_k > 0$ gilt. Unter Verwendung der Wolfe-Schrittweite können wir also wieder `cholupdate` benutzen. Bei den durchgeführten Tests war jedoch zu beobachten, daß der Nenner des Broydenklasse-Parameters verschwinden kann. Dieser Fall sollte in der Implementation berücksichtigt werden, um eine Division durch Null zu vermeiden. Der Gesamtaufwand für die dreifache Anwendung von `cholupdate` und die Richtungsberechnung ist dann wieder $O(n^2)$ Operationen. Die einzelnen Updates haben die formale Gestalt

$$\begin{aligned}
R_{k+\frac{1}{3}}^T R_{k+\frac{1}{3}} &= R_k^T R_k + \frac{y_k y_k^T}{b_k}, \\
R_{k+\frac{2}{3}}^T R_{k+\frac{2}{3}} &= R_{k+\frac{1}{3}}^T R_{k+\frac{1}{3}} - \frac{(R_k^T R_k s_k)(R_k^T R_k s_k)^T}{c_k}, \\
R_{k+1}^T R_{k+1} &= R_{k+\frac{2}{3}}^T R_{k+\frac{2}{3}} - \frac{(a_k - b_k)b_k c_k}{(1-n)(a_k c_k - b_k^2)} v_k v_k^T.
\end{aligned}$$

Dabei benutzen wir die Darstellungen

$$a_k = y_k^T (R_k^T R_k)^{-1} y_k, \quad c_k = \|R_k s_k\|^2, \quad v_k = \frac{y_k}{y_k^T s_k} - \frac{R_k^T R_k s_k}{\|R_k s_k\|^2}.$$

Da der Broydenklasse-Parameter ϕ_k^* auch negativ werden kann, erfolgt die Zerlegung der Faktoren der Rang-1-Matrizen wie beim DW-Verfahren. Die Berechnung von a_k geschieht wie im vorherigen Unterabschnitt ohne Benutzung einer Inversen. Außerdem wird im ersten Iterationsschritt ein initial inverse sizing ausgeführt. Die Matlab-Quelltexte zu den drei beschriebenen Verfahren sind im Anhang zu finden.

5.1.4 Die Wolfe-Schrittweite

Als Schrittweitenstrategie wählen wir für alle Verfahren die Wolfe-Schrittweite, da, wie schon oft erwähnt, hierdurch ohne eine gleichmäßige Konvexitätsvoraussetzung $y_k^T s_k > 0$ für alle k gilt. Für die Berechnung einer Wolfe-Schrittweite sind zahlreiche Algorithmen bekannt. Es soll bei gegebener aktueller Näherung $x_k \in \mathbb{R}^n$ und zugehöhriger Abstiegsrichtung $p_k \in \mathbb{R}^n$ ein $t_k > 0$ mit

$$\text{(I)} \qquad f(x_k + t_k p_k) \leq f(x_k) + \alpha t_k \nabla f(x_k)^T p_k$$

und

$$\text{(II)} \qquad \nabla f(x_k + t_k p_k)^T p_k \geq \beta \nabla f(x_k)^T p_k$$

berechnet werden. Dabei sind $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$ gegebene Konstanten, für die wir später $\alpha := 0.001$ und $\beta := 0.9$ setzen werden. Unser Programm für die Berechnung einer solchen Schrittweite benutzt den Algorithmus von J. E. Dennis Jr. und R. Schnabel aus [8], S. 328ff. und den dort angegebenen Pseudocode. Aus Gründen der Übersichtlichkeit verzichten wir auf eine Vorstellung des Pseudocodes und geben nur die Grundstruktur des Algorithmus wieder.

0. Input: Gegeben seien Konstanten $\alpha \in (0, \frac{1}{2})$ und $\beta \in (\alpha, 1)$, eine aktuelle Näherung $x_k \in L_0$ sowie ein $p \in \mathbb{R}^n$ mit $\nabla f(x_k)^T p_k < 0$.

1. Setze $t := 1$.

2. Falls t die Bedingungen (I) und (II) erfüllt, dann STOP: $t_k := t$ ist eine Wolfe-Schtittweite. Andernfalls:

3. Falls t nur (I) erfüllt und $t \geq 1$ gilt, dann setze $t := 2t$ und gehe zu 2.

4. Falls t nur (I) erfüllt und $t < 1$ gilt oder t die Bedingung (I) nicht erfüllt und $t > 1$ gilt, dann:

   - Sei t_{prev} der letzte zuvor getestete Wert von t.
     Falls $t < 1$ ist, setze $t_{lo} := t$ und $t_{hi} := t_{prev}$.
     Falls $t > 1$ ist, setze $t_{hi} := t$ und $t_{lo} := t_{prev}$.
   - Berechne $t \in (t_{lo}, t_{hi})$ durch sukzessive quadratische Interpolation so, daß t beide Bedinungen (I) und (II) erfüllt. STOP $t_k := t$ ist eine Wolfe-Schrittweite.

5. Falls t die Bedingung (I) nicht erfüllt und $t \leq 1$ gilt, dann verringere t um einen Faktor zwischen 0.1 und 0.5 durch:

   - Im ersten Durchlauf: Wähle $t \geq 0.1$ so, daß t das Minimum des quadratischen hermiteschen Interpolationspolynoms zu den Stützstellen und -werten $(0, f(x_k))$, $(0, \nabla f(x_k)^T p_k)$ und $(1, f(x_k + p_k))$ ist.

- In allen weiteren Durchläufen: Wähle neues $t \in [0.1t_{prev}, 0.5t_{prev}]$ so, daß t das lokale Minimum des kubischen hermiteschen Interploationspolynoms zu den Daten $(0, f(x_k))$, $(0, \nabla f(x_k)^T p_k)$, $(t, f(x_k + tp_k))$ und $(t_{prev}, f(x_k + t_{prev}p_k))$ ist.
- Gehe zu 2.

Als Output des obigen Algorithmus erhalten wir eine Wolfe-Schrittweite $t_k > 0$. Bei genauerer Betrachtung des zugehörigen Pseudocodes in [9] erkennt man, daß dies nicht immer der Fall sein muss, da ein nicht erfolgreicher Ausstieg möglich ist. Bei den von uns benuzten Testfunktionen war dies jedoch nicht der Fall. Wir geben die getesten Zielfunktionen nun an.

5.2 Testfunktionen

Für die numerischen Tests benötigen wir natürlich auch mehrere Zielfunktionen, auf welche die drei Verfahren angewandt werden. Zu diesem Zweck veröffentlichten J. J. Moré, B. S. Garbow und K. E. Hillstrom in [16] eine Sammlung von 35 verschiedenen Zielfunktionen zum Test von Verfahren zur Lösung nichtlinearer Gleichungssysteme, nichtlinearer Ausgleichsprobleme und unrestringierter Optimierungsaufgaben. Für letzteres werden 18 der Zielfunktionen empfohlen. Wie C. Geiger und C. Kanzow, siehe [11] S. 333ff. werden wir dieser Empfehlung folgen. Alle Testfunktionen haben die Gestalt

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

mit $F : \mathbb{R}^n \longrightarrow \mathbb{R}^m$. Für jede der Testfunktionen geben wir an:

(a) die Dimensionen $n \in \mathbb{N}$ und die Komplexität $m \in \mathbb{N}$ des Problems,

(b) die Funktionen $F_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, $i = 1, \ldots, m$ und

(c) den Standard-Startpunkt des Problems $x_0 \in \mathbb{R}^n$.

Auf die in der Literatur sonst üblichen Angaben über bekannte Minima der Testfunktionen verzichten wir, da dies besonders bei beliebig dimensionalen Zielfunktionen sehr unübersichtlich wird.

1. **Helical valley function**

   (a) $n = 3, \quad m = 3$

   (b) $F_1(x) = 10[x_3 - 10\,\theta(x_1, x_2)]$
   $F_2(x) = 10[(x_1^2 + x_2^2)^{1/2} - 1]$

$F_3(x) = x_3$

mit

$$\theta(x_1, x_2) = \begin{cases} \frac{1}{2\pi} \arctan\left(\frac{x_2}{x_1}\right), & \text{falls } x_1 > 0 \\ \frac{1}{2\pi} \arctan\left(\frac{x_2}{x_1}\right) + 0.5, & \text{falls } x_1 < 0 \end{cases}$$

(c) $x_0 = (-1, 0, 0)^T$

2. **Biggs' EXP6 function**

(a) $n = 6, \quad m \geq n$ variabel

(b) $F_i(x) = x_3 \exp(-t_i x_1) - x_4 \exp(-t_i x_2) + x_6 \exp(-t_i x_5) - y_i$

mit

$y_i = \exp(-t_i) - 5\exp(-10t_i) + 3\exp(-4t_i)$ und $t_i = i/10$

(c) $x_0 = (1, 2, 1, 1, 1, 1)^T$

3. **Gaussian function**

(a) $n = 3, \quad m = 15$

(b) $F_i(x) = x_1 \exp\left(\frac{-x_2(t_i - x_3)^2}{2}\right) - y_i$

mit $t_i = (8 - i)/2$ und

i	y_i
1,15	0.0009
2,14	0.0044
3,13	0.0175
4,12	0.0540
5,11	0.1295
6,10	0.2420
7,9	0.3521
8	0.3989

(c) $x_0 = (0.4, 1, 0)^T$

4. **Powell badly scaled function**

(a) $n = 2, \quad m = 2$

(b) $F_1(x) = 10^4 x_1 x_2 - 1$

$F_2(x) = \exp(-x_1) + \exp(-x_2) - 1.0001$

(c) $x_0 = (0, 1)^T$

5. **Box three-dimensional function**

(a) $n = 3, \quad m \geq n$ variabel

(b) $F_i(x) = \exp(-t_i x_1) - \exp(-t_i x_2) - x_3(\exp(-t_i) - \exp(-10t_i))$
mit $t_i = i/10$

(c) $x_0 = (0, 10, 20)^T$

6. **Variably dimensioned function**

(a) n variabel, $\quad m = n + 2$

(b) $F_i(x) = x_i - 1, \quad 1 \le i \le n$
$F_{n+1}(x) = \sum_{j=1}^{n} j(x_j - 1)$
$F_{n+2}(x) = \left(\sum_{j=1}^{n} j(x_j - 1)\right)^2$

(c) $x_0 = (\xi_j)_{j=1,\ldots,n}^T$ mit $\xi_j = 1 - (j/n)$

7. **Watson function**

(a) $2 \le n \le 31, \quad m = 31$

(b) $F_i(x) = \sum_{j=2}^{n}(j-1)x_j t_i^{j-2} - \left(\sum_{j=1}^{n} x_j t_i^{j-1}\right)^2 - 1$
mit $t_i = i/29$

(c) $x_0 = (0, \ldots, 0)^T$

8. **Penalty function I**

(a) n variabel, $\quad m = n + 1$

(b) $F_i(x) = 10^{-5/2}(x_i - 1), \quad 1 \le i \le n$
$F_{n+1}(x) = \left(\sum_{j=1}^{n} x_j^2\right) - \frac{1}{4}$

(c) $x_0 = (1, 2, \ldots, n)^T$

9. **Penalty function II**

(a) n variabel, $\quad m = 2n$

(b) $F_1(x) = x_1 - 0.2$
$F_i(x) = 10^{-5/2}\left(\exp(0.1x_i) + \exp(0.1x_{i-1}) - y_i\right), \quad 2 \le i \le n$
$F_i(x) = 10^{-5/2}\left(\exp(0.1x_{i-n+1}) + \exp(-0.1)\right), \quad n < i \le 2n - 1$
$F_{2n}(x) = \left(\sum_{j=1}^{n}(n - j + 1)x_j^2\right) - 1$
mit $y_i = \exp(i/10) + \exp((i-1)/10)$

(c) $x_0 = (0.5, \ldots, 0.5)^T$

10. **Brown badly scaled function**

(a) $n = 2, \quad m = 3$

(b) $F_1(x) = x_1 - 10^6$
$F_2(x) = x_2 - 2 \cdot 10^-6$
$F_3(x) = x_1 x_2 - 2$

(c) $x_0 = (1, 1)^T$

11. **Brown and Dennis function**

(a) $n = 4, \quad m \geq n$ variabel

(b) $F_i(x) = (x_1 + t_i x_2 - \exp(t_i))^2 + (x_3 + x_4 \sin t_i - \cos t_i)^2$
mit $t_i = i/5$

(c) $x_0 = (25, 5, -5, -1)^T$

12. **Gulf research and development function**

(a) $n = 3, \quad n \leq m \leq 100$

(b) $F_i(x) = \exp\left(-\frac{|y_i - x_2|^{x_3}}{x_1}\right) - t_i$
mit $y_i = 25 + (-50 \ln t_i)^{2/3}$ und $t_i = i/100$

(c) $x_0 = (5, 2.5, 0.15)^T$

13. **Trigonometric function**

(a) n variabel, $\quad m = n$

(b) $F_i(x) = n - \sum_{j=1}^{n} \cos x_j + i(1 - \cos x_i) - \sin x_i$

(c) $x_0 = (1/n, \ldots, 1/n)^T$

14. **Extended Rosenbrock function**

(a) n gerade, $\quad m = n$

(b) $F_{2i-1}(x) = 10(x_{2i} - x_{2i-1}^2)$
$F_{2i}(x) = 1 - x_{2i-1}$

(c) $x_0 = (\xi_j)_{j=1,\ldots,n}^T$ mit $\xi_{2j-1} = -1.2$, $\xi_{2j} = 1$

15. **Extended Powell singular function**

(a) n ein Vielfaches von 4, $\quad m = n$

(b) $F_{4i-3}(x) = x_{4i-3} + 10x_{4i-2}$
$F_{4i-2}(x) = 5^{1/2}(x_{4i-1} - x_{4i})$
$F_{4i-1}(x) = (x_{4i-2} - 2x_{4i-1})^2$
$F_{4i}(x) = 10^{1/2} (x_{4i-3} - x_{4i})^2$

(c) $x_0 = (\xi_j)_{j=1,\ldots,n}^T$ mit $\xi_{4j-3} = 3$, $\xi_{4j-2} = -1$, $\xi_{4j-1} = 0$, $\xi_{4j} = 1$

16. **Beale function**

    (a) $n = 2, \quad m = 3$

    (b) $F_i(x) = y_i - x_1(1 - x_2^i)$
    mit $y_1 = 1.5$, $y_2 = 2.25$, $y_3 = 2.625$

    (c) $x_0 = (1, 1)^T$

17. **Wood function**

    (a) $n = 4, \quad m = 6$

    (b) $F_1(x) = 10(x_2 - x_1^2)$
    $F_2(x) = 1 - x_1$
    $F_3(x) = 90^{1/2}(x_4 - x_3^2)$
    $F_4(x) = 1 - x_3$
    $F_5(x) = 10^{1/2}(x_2 + x_4 - 2)$
    $F_6(x) = 10^{-1/2}(x_2 - x_4)$

    (c) $x_0 = (-3, -1, -3, -1)^T$

18. **Chebyquad function**

    (a) n variabel, $\quad m \geq n$

    (b) $F_i(x) = \frac{1}{n} \sum_{j=1}^{n} T_i(x_j) - \int_0^1 T_i(\tau) d\tau$
    wobei T_i das auf das Intervall $[0, 1]$ transformierte i-te Tschebyscheff-Polynom ist.

    (c) $x_0 = (\xi_j)_{j=1,\ldots,n}^T$ mit $\xi_j = j/(n + 1)$

Die letzte dieser 18 Zielfunktionen können wir in unseren Tests leider nicht verwenden. Dies hat den folgenden Grund: Für jede Komponente x_j von x benötigen wir ein Intervall $[a, b]$, in welchem sich x_j befindet und welches wir dann auf $[0, 1]$ transformieren. Da sich die Komponenten mit jeder Iteration der Verfahren aber ändern, müssen wir die Intervalle vorher sehr groß wählen. Bedingt durch die großen Intervalle sind die transformierten Werte für den Gradienten so klein, daß das Abbruchkriterium jeweils nach nur einer Iteration erreicht ist, ohne daß ein stationärer Punkt vorliegt.

Desweiteren müssen wir den Startvektor bei der helical valley function ändern, da der ursprünglich angegebene Vektor bereits stationärer Punkt der Zielfunktion ist. Als neuen Startvektor wählen wir $x_0 = (1, 1, 1)^T$ und erläutern jetzt die Testergebnisse.

5.3 Testergebnisse

Wir testen nun die numerischen Eigenschaften der drei Verfahren, indem wir sie auf die 17 zur Verfügung stehenden Testfunktionen anwenden. Einige dieser Testfunktionen erlauben es, entweder die Dimension n oder die Komplexität m frei oder mit gewissen Einschränkungen zu wählen. In diesen Fällen benutzen wir mehrfach dieselbe Testfunktion mit unterschiedlichen Werten für n oder m. Da die Verwendung von Broydenklasse-Verfahren nur bis zu einer Dimension von etwa 100 Variablen empfohlen wird, wählen wir falls möglich $n = 3, 10, 20, 50, 100$, beziehungsweise $m = 3, 10, 20, 50, 100$. Falls wir diese Werte aufgrund von Beschränkungen nicht wählen dürfen, wählen wir zulässige Werte ähnlicher Größe. Als Testkriterien werden wir die Anzahl der Iterationen bis zum Erreichen des Abbruchkriteriums sowie die Anzahl der hierfür benötigten Zielfunktionsauswertungen benutzen. Unter einer Auswertung verstehen wir dabei eine Auswertung von Zielfunktion und Gradient. Eine getrennte Angabe der Gradientenauswertungen ist bei der gewählten Implementierung nicht nötig, da die Zielfunktion und ihr Gradient immer gleichzeitig ausgewertet werden. Die Präsentation der Ergebnisse erfolgt in tabellarischer Form. In jeder Zeile werden angegeben:

- Die verwendete Zielfunktion (func) durch ihre Nummer im vorangegangenen Abschnitt,
- die verwendete Zahl an Variablen (n) der Zielfunktion,
- die verwendete Komplexität (m) der Zielfunktion,
- die Anzahl der benötigten Iterationen (iter) und
- die Anzahl der benötigten Funktionsauswertungen (feval).

Die Angabe der letzten beiden Größen erfolgt für alle drei Verfahren getrennt. Ein Überschreiten von $k_{max} = 1000$ kennzeichenen wir durch ein "∞" für iter und feval. Einen Abbruch des "optimal ϕ"-Verfahrens aufgund von Auslöschung im Nenner machen wir durch ein "$\times$" kenntlich.

Tabelle 5.1: Testergebnisse

			BFGS		DW		opt ϕ	
func	n	m	iter	feval	iter	feval	iter	feval
1	3	3	26	32	19	23	23	25
2	6	6	43	49	54	59	54	57
2	6	10	39	42	44	48	51	58
2	6	20	36	40	46	49	45	48
2	6	50	44	49	47	52	48	53

			BFGS		DW		opt ϕ	
func	n	m	iter	feval	iter	feval	iter	feval
2	6	100	45	50	47	53	51	58
3	3	15	4	15	6	16	6	16
4	2	2	259	375	308	471	294	338
5	3	3	42	54	42	51	43	53
5	3	10	44	54	38	46	37	45
5	3	20	39	47	32	39	35	42
5	3	50	39	40	38	41	38	39
5	3	100	42	48	41	47	45	51
6	3	5	7	10	7	10	7	10
6	10	12	10	16	10	16	×	×
6	20	22	23	24	23	24	×	×
6	50	52	31	32	31	32	×	×
6	100	102	37	38	37	38	×	×
7	3	31	14	16	17	19	16	18
7	10	31	84	89	116	121	121	124
7	20	31	79	84	133	140	144	149
7	31	31	81	89	157	166	165	172
8	3	4	79	94	92	116	90	94
8	10	11	101	125	99	132	×	×
8	20	21	104	131	83	105	×	×
8	50	51	96	126	90	117	×	×
8	100	101	104	137	96	125	×	×
9	3	6	10	16	12	17	13	18
9	10	20	175	221	205	273	320	368
9	20	40	537	581	498	552	766	861
9	50	100	617	622	627	642	615	626
9	100	200	794	795	722	723	755	756
10	2	3	14	51	17	55	×	×
11	4	4	99	112	85	95	83	89
11	4	10	79	82	62	67	59	64
11	4	20	48	50	36	38	34	36
11	4	50	∞	∞	∞	∞	∞	∞
11	4	100	∞	∞	∞	∞	∞	∞
12	3	3	1	2	1	2	×	×
12	3	10	1	2	1	2	×	×
12	3	20	1	2	1	2	×	×
12	3	50	1	2	1	2	1	2
12	3	100	56	64	63	76	72	75
13	3	3	15	22	16	19	16	19
13	10	10	40	77	33	36	34	37

			BFGS		DW		opt ϕ	
func	n	m	iter	feval	iter	feval	iter	feval
13	20	20	53	127	63	69	65	73
13	50	50	116	339	57	62	56	61
13	100	100	180	703	61	69	63	74
14	2	2	57	73	48	61	61	65
14	10	10	76	87	59	72	47	53
14	20	20	61	79	61	76	58	66
14	50	50	57	84	60	73	64	78
14	100	100	52	65	51	57	55	64
15	4	4	41	42	40	41	42	43
15	12	12	56	57	44	45	49	50
15	20	20	55	56	40	41	40	41
15	52	52	66	68	48	49	45	46
15	100	100	71	74	56	57	60	61
16	2	3	13	14	16	17	17	18
17	4	6	43	46	35	47	40	46

Wenn wir nun die erhaltene Datenflut betrachten, fällt vor allem eines auf: Keines der drei Verfahren ist den anderen auffällig überlegen. Daher müssen wir mit statistischen Methoden arbeiten. Weil das "optimal ϕ"-Verfahren bei einigen der Zielfunktionen mit einer versuchten Nulldivison abbricht, müssen wir bei einem Vergleich aller drei Verfahren auf die Daten zur entsprechenden Zielfunktion verzichten. Deshalb führen wir ab jetzt zwei getrennte Untersuchungen durch. In der ersten Untersuchung vergleichen wir nur das BFGS-Verfahren mit dem Dennis-Wolkowicz-Verfahren und verzichten dabei lediglich auf die Daten zu den Zielfunktionen, bei denen $k_{\max}$ überschritten wird. In der zweiten Untersuchung berücksichtigen wir dann alle drei Verfahren, müssen dafür allerdings zusätzlich auf die Informationen zu den Zielfunktionen mit versuchter Nulldivison verzichten.

Wir beginnen nun mit dem Vergleich von BFGS- und DW-Verfahren. Da nach einer Überprüfung mit SAS für keine der auftretenden Größen eine Verteilungsannahme gerechtfertigt ist und die Werte für die Iterationszahlen und Zielfunktionsauswertungen untereinander abhängig und verbunden sind, können wir leider keine statistischen Tests durchführen. Dies beschränkt uns auf deskriptive, insbesondere graphische Methoden. Um uns einen groben Überblick über die gegebenen Iterationszahlen zu verschaffen, erstellen wir mit STATISTICA einen Scatterplot. Dabei wird für jede Zeile in obiger Tabelle die Iterationszahl des BFGS-Verfahrens gegen die Iterationszahl des DW-Verfahrens aufgetragen. Die Lage des dadurch entstandenen Punktes im Koordinatensystem bezüglich der Winkelhalbierenden zeigt dann, welches der Verfahren weniger Iterationen benötigt

hat. Desweiteren lassen wir für alle so entstandenen Punkte eine Regressionsgerade zeichen. Die Lage dieser Geraden zur Winkelhalbierenden gibt wiederum Aufschluss darüber, welches Verfahren durchschnittlich weniger Iterationen benötigt. Bei Betrachtung der zugehörigen Abbildung 5.1 fällt auf, daß ein we-

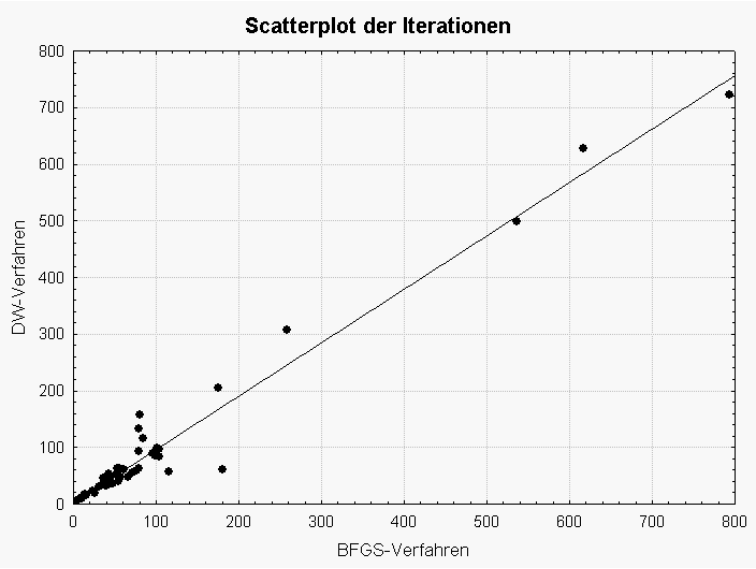

Abbildung 5.1: Iterationen von BFGS- und DW-Verfahren

nig mehr Punkte unterhalb der Winkelhalbierenden liegen als oberhalb. Für die entsprechenden Zielfunktionen benötigt das BFGS-Verfahren also mehr Iterationen als das DW-Verfahren. Die Regressionsgerade liegt nur etwas unterhalb der Winkelhalbierenden. Wir können also höchstens auf eine leichte Überlegenheit des Dennis-Wolkowicz-Verfahrens bei den Iterationszahlen schließen, da die Situation fast ausgeglichen ist. Weil wir uns aber nicht nur auf eventuell subjektive optische Eindrücke bei Scatterplots verlassen wollen, berechnen wir noch einige Kennzahlen. Als erstes berechnen wir für beide Verfahren die durchschnittlich benötigten Iterationen als arithmetisches Mittel der Iterationszahlen. Allerdings gibt uns diese Kennzahl keinen Aufschluß darüber, ob und wie stark ein Verfahren dem anderen überlegen ist. Daher führen wir eine relative Bewertung der Iterationszahlen ein. Für jede Zielfunktion erhält das Verfahren mit geringerer Iterationszahl die relative Bewertung 0. Sind die Iterationszahlen für beide Verfahren gleich, erhalten beide eine 0. Benötigt ein Verfahren mehr Iterationen, so werden diesem Verfahren die benötigten Mehriterationen im Verhältnis zu den Iterationen des besseren Verfahrens als Bewertung zugeordnet. Die relative Bewertung entspricht dann den benötigten prozentualen Mehriterationen im Vergleich zum besseren Verfahren. Als Beispiel betrachten wir die Daten zur helical valley function in Tabelle 5.1. Das DW-Verfahren benötigt 19 Iterationen, das BFGS-Verfahren 7 Iterationen mehr. Als Bewertungen erhält das DW-Verfahren

folglich eine 0 und das BFGS-Verfahren $7/19 \approx 0.3642$. Das BFGS-Verfahren ist hier also ca. 36,42 % schlechter als das DW-Verfahren. Auf diese Art verfahren wir mit allen nutzbaren Daten aus Tabelle 5.1 und erhalten unter Verwendung von SAS die folgenden Mittelwerte der Iterationszahlen und relativen Bewertungen:

Tabelle 5.2: Vergleich der Iterationszahlen

∅	BFGS	DW
Iterationen	86.845	84.000
rel. Bewertung	0.128	0.088

Das BFGS-Verfahren benötigt also durchschnittlich 2.845 Iterationen mehr als das DW-Verfahren und ist durchschnittlich 12.8 % schlechter als das jeweils bessere Verfahren. Das DW-Verfahren ist im Schnitt nur 8.8 % schlechter als das jeweils bessere Verfahren. Anhand dieser Kennzahlen können wir also den Eindruck des Scatterplots bestätigen und auf eine leichte Überlegenheit des DW-Verfahrens bei den benötigen Iterationen schließen.

Für die Anzahl der Zielfunktionsauswertungen gehen wir analog vor und erhalten die Abblidung 5.2. Hier ist die Interpretation des Scatterplots eindeutiger, da die

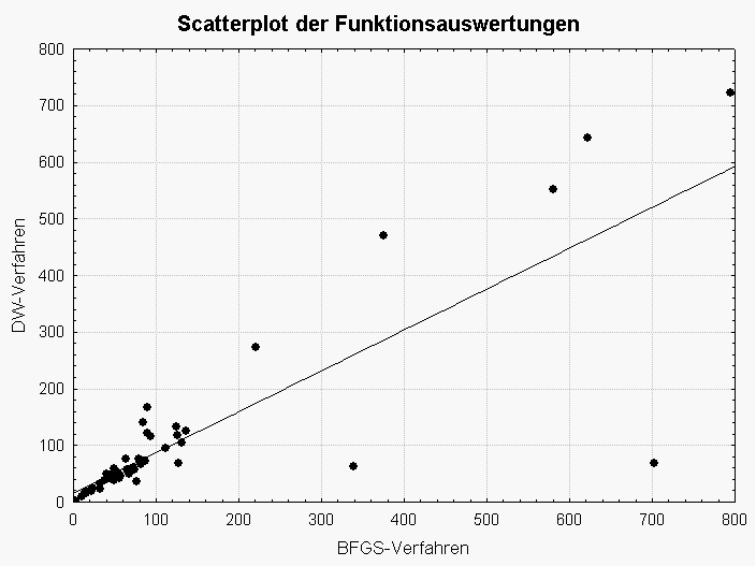

Abbildung 5.2: Funktionsauswertungen von BFGS- und DW-Verfahren

Mehrheit der Punkte unterhalb der Winkelhalbierenden liegt und die berechnete Regressionsgerade weit unterhalb verläuft. Wir können also auf eine deutliche Überlegenheit des DW-Verfahrens bei den benötigten Zielfunktionsauswertungen

schließen. Um dies zu überprüfen, berechnen wir analog zu Tabelle 5.2 die durchschnittlich benötigten Funktionsauswertungen und die durchschnittliche relative Bewertung der Funktionsauswertungen.

Tabelle 5.3: Vergleich der Funktionsauswertungen

∅	BFGS	DW
Funktionsausw.	112.431	96.431
rel. Bewertung	0.345	0.073

Beide Kennzahlen sind für das DW-Verfahren deutlich geringer als für das BFGS-Verfahren. Der durch den Scatterplot entstandene Eindruck ist also gerechtfertigt. Das DW-Verfahren ist dem BFGS-Verfahren bei den benötigten Zielfunktionsauswertungen stark überlegen. Allerdings ist ein sehr großer Anteil des guten Abschneidens des DW-Verfahrens auf die trigonometric function zurückzuführen. Bei dieser Zielfunktion schneidet das BFGS-Verfahren nämlich derart schlecht ab, daß es bei einem Verzicht auf diese Funktion insgesamt im Vorteil wäre. Wir werten diese Daten jedoch nicht als vernachlässigbare Ausreißer, sondern arbeiten auch weiterhin mit ihnen. Eine Zielfunktion von geringer Bedeutung würde schließlich keine Berücksichtigung in der anerkannten Testsammlung aus [16] finden. Also ist das DW-Verfahren bei den benötigten Funktionsauswertungen wesentlich besser als das BFGS-Verfahren.

Betrachtet man nun beide Testkriterien zusammen, so ergibt sich beim Vergleich der beiden Verfahren ein Vorteil für das Dennis-Wolkowicz-Verfahren. Wir können also die Testergebnisse von Dennis und Wolkowicz für diese beiden Verfahren in ihrer Tendenz bestätigen. Es ist jedoch zu bedenken, daß für die Berechnung eines Cholesky-Updates beim DW-Verfahren die Matlab-Funktion `cholupdate` einmal mehr aufgerufen werden muss als beim BFGS-Verfahren. Daher ist bei der gewählten Implementation der numerische Aufwand für die Berechnung eines DW-Updates um 50% höher als für ein BFGS-Update. Dies kann den obigen Vorteil wieder ausgleichen oder sogar umkehren. Bei einer elementaren Implementation des DW-Updates mittels Givens-Rotationen analog zum BFGS-Update von J. Werner aus [23], S. 201f. könnte der Unterschied geringer ausfallen. Da dies aber eine Frage der Implementation des Updates und nicht des Verfahrens an sich ist, überprüfen wir dies nicht genauer. Auf eine Zählung der tatsächlich benötigten Fließkommaoperationen müssen wir ohnehin verzichten, da Matlab dies ab der Version 6 nicht mehr unterstützt.

Wir fahren nun mit dem Vergleich aller drei Verfahren fort und verzichten dabei auf die Zielfunktionen, bei denen $k_{\max}$ überschritten oder eine Nulldivison versucht wird. Da wiederum keine Verteilungsannahmen gerechtfertigt sind und Abhängigkeit der Daten besteht, gehen wir erneut graphisch und deskriptiv vor.

Bei den Scatterplots vergleichen wir die Verfahren jeweils paarweise. Ein dreidimensionaler Scatterplot ist mit STATISTICA zwar möglich, jedoch in gedruckter Form nur schwer zu interpretieren. Die Analyse erfolgt vorerst für beide Testkriterien getrennt. Zunächst betrachten wir wieder BFGS- und DW-Verfahren. Dies ist zwar schon geschehen, aber durch die entfallenen Daten entstehen leicht

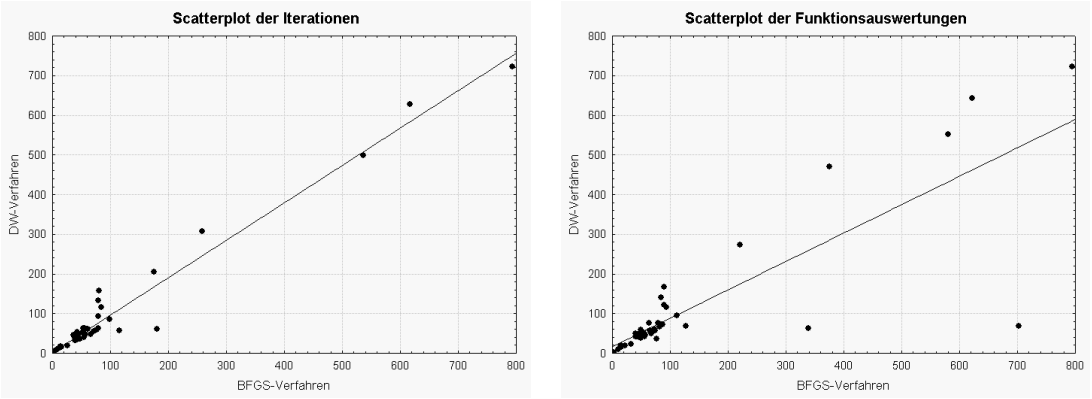

Abbildung 5.3: Vergleich von BFGS- und DW-Verfahren

veränderte Ergebnisse. Der Scatterplot der benötigten Iterationen mit Regressionsgerade links in Abbildung 5.3 verändert sich durch die zusätzlich entfallenen Werte, beziehungsweise Punkte nur minimal. Es liegen immer noch ein wenig mehr Punkte unterhalb der Winkelhalbierenden als oberhalb. Die Regressionsgerade ist nahezu unverändert. Wir können also erneut höchstens auf eine sehr geringe Überlegenheit des DW-Verfahrens bei den Iterationen schließen. Bei den benötigten Zielfunktionsauswertungen rechts in Abbildung 5.3 verändert sich der Scatterplot ebenfalls nur geringfügig. Es liegt also wieder ein deutlicher Vorteil des DW-Verfahrens bei den Zielfunktionsauswertungen vor. Jetzt betrachten wir

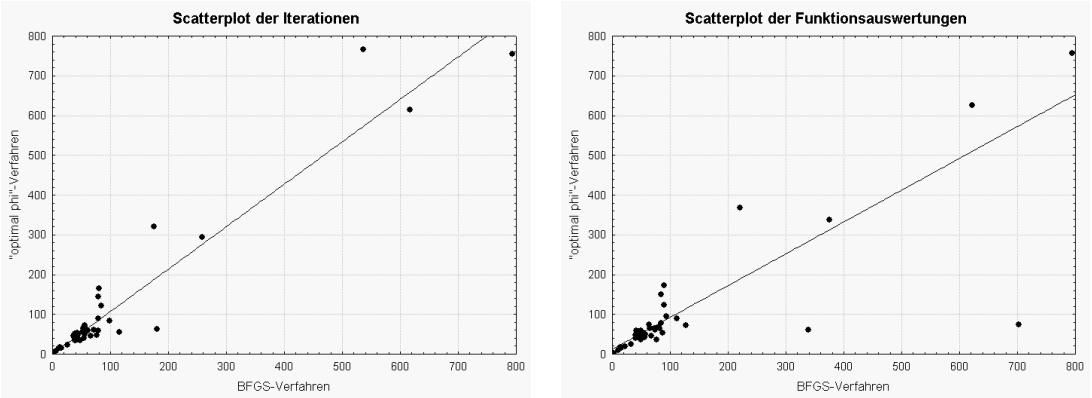

Abbildung 5.4: Vergleich von BFGS- und "optimal ϕ"-Verfahren

BFGS- und "optimal ϕ"-Verfahren. Wie beim Vergleich von BFGS- und DW-Verfahren zeigt der Scatterplot der Iterationen in Abbildung 5.4 links ein fast ausgeglichenes Verhältnis zwischen den Verfahren. Das BFGS-Verfahren ist dabei leicht im Vorteil. Bei den Funktionsauswertungen rechts in Abbildung 5.4 ist das BFGS-Verfahren allerdings deutlich unterlegen. Für den noch fehlenden Vergleich von DW- und "optimal ϕ"-Verfahren betrachten wir zunächst Abbildung 5.5. Der Scatterplot der benötigten Iterationen zeigt eine Überlegenheit des

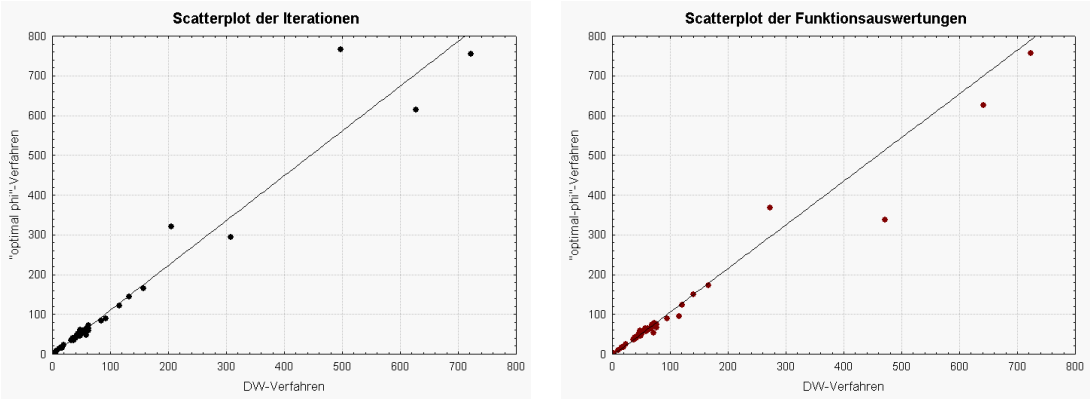

Abbildung 5.5: Vergleich von DW- und "optimal ϕ"-Verfahren

DW-Verfahrens. Diese Überlegenheit ist zwar wieder gering, aber etwas ausgeprägter als bei den vorangegangenen Plots der Iterationszahlen. Der Scatterplot der benötigten Funktionsauswertungen zeigt ebenfalls eine geringe Überlegenheit des DW-Verfahrens. Nach diesem Vergleich der Verfahren mit graphischen Mitteln berechnen wir nun die bereits bekannten Kennzahlen. Die relative Bewertung bezieht sich dann auf das jeweils beste der drei Verfahren.

Tabelle 5.4: Vergleich der Iterationszahlen

$\varnothing$	BFGS	DW	optimal ϕ
Iterationen	98.130	95.283	105.283
rel. Bewertung	0.167	0.119	0.178

Die erhaltenen Werte in Tabelle 5.4 bestätigen wieder die Interpretationen der entsprechenden Scatterplots. Das DW-Verfahren ist dem BFGS- und dem "optimal ϕ"-Verfahren bei beiden Kennzahlen überlegen. Das BFGS-Verfahren ist verglichen mit dem "optimal ϕ"-Verfahren nur geringfügig im Vorteil. Die Rangfolge der Verfahren bezüglich der benötigten Iterationen ist also

$$\text{DW} \quad > \quad \text{BFGS} \quad > \quad \text{optimal } \phi.$$

Hierbei steht das Symbol ">" für eine leichte Überlegenheit. Die Rangfolge von Dennis und Wolkowicz kann bei diesem Testkriterium nur teilweise bestätigt

werden, da die in [9] beobachtete starke Unterlegenheit des BFGS-Verfahrens gegenüber beiden anderen Verfahren hier nicht vorliegt. Jetzt betrachten wir die Kennzahlen zu den Funktionsauswertungen.

Tabelle 5.5: Vergleich der Funktionsauswertungen

$\varnothing$	BFGS	DW	optimal ϕ
Funktionsausw.	126.848	107.457	113.913
rel. Bewertung	0.451	0.110	0.132

Die berechneten Werte bestätigen die starke Unterlegenheit des BFGS-Verfahrens gegenüber den anderen beiden Verfahren. Von diesen ist dann das DW-Verfahren leicht im Vorteil. Die Interpretationen der Scatterplots sind durch die Kennzahlen also bestätigt. Die daraus entstehende Rangfolge bezüglich der benötigten Funktionsauswertungen ist somit

$$\text{DW} \;>\; \text{optimal } \phi \;\gg\; \text{BFGS},$$

wobei "$\gg$" für eine starke Überlegenheit steht. Das überaus schlechte Abschneiden des BFGS-Verfahrens ist wiederum durch die trigonometric function bei hoher Dimension zu erklären. Die Ergebnisse von Dennis und Wolkowicz bezüglich der benötigten Funktionsauswertungen werden also in der Tendenz bestätigt.

Fassen wir die Betrachtungen zu den zwei Testkriterien zusammen, so ist das DW-Verfahren den beiden anderen überlegen. Das zweitbeste ist das "optimal ϕ"-Verfahren gefolgt vom BFGS-Verfahren. Es ist jedoch zu beachten, daß das "optimal ϕ"-Verfahren nicht bei allen Zielfunktionen einsetzbar ist, da sichergestellt werden muss, daß im Broydenklasse-Parameter keine Division durch Null stattfindet. Desweiteren muss das BFGS-Verfahren `cholupdate` nur zweimal je Iteration aufrufen, die beiden anderen Verfahren einmal mehr. Ob sich hierdurch an der Rangfolge der Verfahren etwas ändert, können wir aus besagten Gründen leider nicht feststellen. Eine Zusammenfassung der theoretischen und numerischen Aspekte erfolgt im nächsten Kapitel.

Kapitel 6

Zusammenfassung und Ausblick

Nach den umfangreichen Untersuchungen der vorangegangenen Kapitel stellt sich natürlich die Frage, welche Bedeutung dem Dennis-Wolkowicz-Verfahren für unrestringierte Optimierungsaufgaben zukommt. Dazu wiederholen wir die wichtigsten Ergebnisse in knapper Form und urteilen anschließend über die Qualität des Verfahrens, insbesondere im Vergleich zum BFGS-Verfahren. Wir trennen zunächst nach theoretischen und numerischen Eigenschaften und ziehen dann ein Fazit über die Bedeutung aller in dieser Arbeit untersuchten Verfahren.

Die theoretischen Konvergenzeigenschaften des Dennis-Wolkowicz-Verfahrens, so wie sie in Kapitel 4 ausführlich vorgestellt worden sind, sind identisch mit den entsprechenden Eigenschaften des BFGS-Verfahrens. Viele dieser Eigenschaften sind schon seit längerem bekannt und stammen aus der Arbeit [21] von J. Werner. Dies sind im einzelnen:

- Globale R-lineare Konvergenz des durch eine effiziente Schrittweitenstrategie gedämpften Verfahrens bei gleichmäßig konvexer Zielfunktion,
- globale Q-superlineare Konvergenz des ungedämpften Verfahrens bei strikt konvexer quadratischer Zielfunktion,
- Gültigkeit der Dennis-Moré-Bedingung für Q-superlineare Konvergenz und sich daraus ergebende Folgerungen.

Auf Basis dieser Ergebnisse und den aus [22] stammenden Untersuchungen zum BFGS-Verfahren konnten wir dann weitere Konvergenzeigenschaften auf das DW-Verfahren übertragen und beim globalen Konvergenzsatz 4.1 auch semi-effiziente Schrittweiten zulassen. Die so entstandenen Resultate sind die folgenden:

- Globale R-lineare Konvergenz des durch eine semi-effiziente Schrittweitenstrategie gedämpften Verfahrens bei gleichmäßig konvexer Zielfunktion,
- Gültigkeit eines Bounded-Deterioration-Satzes und daraus folgende lokale Q-lineare Konvergenz des ungedämpften Verfahrens,

- globale Konvergenz des DW-Verfahrens mit "cautious" Update bei semi-effizienter Schrittweitenstrategie und nichtkonvexer Zielfunktion.

Somit gelten die wichtigsten theoretischen Konvergenzeigenschaften des BFGS-Verfahrens auch für das Dennis-Wolkowicz-Verfahren. Lediglich der Satz von Powell konnte nicht übertragen werden. Da dieser jedoch nur Dämpfung durch die Wolfe-Schrittweite erlaubt und keine Angabe über die Konvergenzgeschwindigkeit macht, ist dies nur ein kleinerer Mangel des DW-Verfahrens. Auf der theoretischen Ebene sind das Dennis-Wolkowicz-Verfahren und das BFGS-Verfahren also gleichwertig.

Bei der Untersuchung der numerischen Eigenschaften des Dennis-Wolkowicz-Verfahrens galt unser Interesse hauptsächlich dem Vergleich mit dem BFGS-Verfahren. Die wichtigsten Schlussfolgerungen aus den Testergebnissen sind:

- Das Dennis-Wolkowicz-Verfahren ist mindestens gleichwertig zum BFGS-Verfahren im Testkriterium Anzahl der benötigten Iterationen.

- Das Dennis-Wolkowicz-Verfahren ist dem BFGS-Verfahren überlegen im Testkriterium Anzahl der benötigten Zielfunktionsauswertungen.

Das Dennis-Wolkowicz-Verfahren scheint dem BFGS-Verfahren also in den numerischen Tests überlegen zu sein. In der gewählten Implementation ist jedoch der Aufwand für die Berechnung eines DW-Updates um 50% höher als für ein BFGS-Update, was den oben beschriebenen Vorteil wieder ausgleichen oder umkehren kann. In der Praxis ist das Dennis-Wolkowicz-Verfahren dem BFGS-Verfahren also nicht zwangsweise überlegen.

Daher wird wohl auch in Zukunft das BFGS-Verfahren das Verfahren der Wahl zur numerischen Behandlung unrestringierter Optimierungsaufgaben bei niedriger Dimension bleiben. Das Dennis-Wolkowicz-Verfahren wird trotz seiner guten numerischen und theoretischen Eigenschaften in den praktischen Anwendungen höchstens eine untergeordnete Rolle spielen und nur von theoretischem Interesse sein. Das bei den numerischen Tests kurz erwähnte "optimal ϕ"-Verfahren verfügt zwar über gute numerische Eigenschaften, wird ohne eine umfassende Untersuchung der theoretischen Eigenschaften wohl auch weiterhin ohne Bedeutung bleiben.

Mit diesem Fazit beenden wir die Ausführungen über "Das Dennis-Wolkowicz-Verfahren für unrestringierte Optimierungsaufgaben" und geben in den nachfolgenden Anhängen die noch fehlenden Quelltexte an.

Anhang A

Matlab-Quelltexte

A.1 Das BFGS-Verfahren

```
function[x, iter, f_eval, A]=bfgs(f_name, param, x_c, max_iter, tol);
%
% BFGS-Verfahren
%
% Input:
%
% f_name:   zu minimierende Zielfunktion [f,g,H]=f_name(x)
% param:    Parameter fuer die Zielfunktion
% x_c:      Startvektor des Verfahrens
% max_iter: maximale Anzahl der Iterationen
% tol:      Abbruchgrenze fuer die Gradientennorm
%
% Output:
%
% x:        Loesungsnaeherung, mit der das Verfahren abbricht
% iter:     Anzahl der Iterationen bis zum Abbruch
% f_eval:   Anzahl der Zielfunktionsauswertungen bis zum Abbruch
% A:        Matrix mit Zielfunktionswerten und -Gradientennormen
%
%
[f_c,g_c]=feval(f_name,x_c,param);
n=length(x_c);
R_c=sqrt(abs(f_c))*eye(n);

f_eval=1;
iter=0;
A=[iter,f_c,norm(g_c)];

while(norm(g_c) > tol)&(iter <= max_iter)
```

```
    p=-R_c\(R_c'\g_c);

    [t,f_p,g_p,f_ev]=wolfe3(x_c,p,f_c,g_c,f_name,param);
    f_eval=f_eval+f_ev;

    s=t*p;
    y=g_p-g_c;
    x_p=x_c+s;

    Rs=R_c*s;
    Bs=R_c'*Rs;
    r1=sqrt(Rs'*Rs);
    r2=sqrt(y'*s);

    R_c=cholupdate(R_c, y /r2, '+');
    R_c=cholupdate(R_c, Bs/r1, '-');

    iter=iter+1;
    A=[A;iter,f_p,norm(g_p)];
    g_c=g_p;
    x_c=x_p;
end;
x=x_c;
```

A.2 Das Dennis-Wolkowicz-Verfahren

```
function[x, iter, f_eval, A]=dw(f_name, param, x_c, max_iter, tol);
%
% Dennis-Wolkowicz-Verfahren
%
% Input:
%
% f_name:   zu minimierende Zielfunktion [f,g,H]=f_name(x)
% param:    Parameter fuer die Zielfunktion
% x_c:      Startvektor des Verfahrens
% max_iter: maximale Anzahl der Iterationen
% tol:      Abbruchgrenze fuer die Gradientennorm
%
% Output:
%
% x:        Loesungsnaeherung, mit der das Verfahren abbricht
% iter:     Anzahl der Iterationen bis zum Abbruch
% f_eval:   Anzahl der Zielfunktionsauswertungen bis zum Abbruch
```

```
% A:        Matrix mit Zielfunktionswerten und -Gradientennormen
%
%
[f_c,g_c]=feval(f_name,x_c,param);
n=length(x_c);
R_c=sqrt(abs(f_c))*eye(n);

f_eval=1;
iter=0;
A=[iter,f_c,norm(g_c)];

while(norm(g_c) > tol)&(iter <= max_iter)

     p=-R_c\(R_c'\g_c);

     [t,f_p,g_p,f_ev]=wolfe3(x_c,p,f_c,g_c,f_name,param);
     f_eval=f_eval+f_ev;

     s=t*p;
     y=g_p-g_c;
     b=y'*s;
     x_p=x_c+s;

     if iter==0
         a_1=R_c'\y;
         a=a_1'*a_1;
         R_c=sqrt(a/b)*R_c;
     end;

     a_1=R_c'\y;
     a=a_1'*a_1;
     r0=sqrt(abs((a-b)/(a*b)));

     if (((a-b)/(a*b)) >= 0)
         R_c=cholupdate(R_c, y*r0, '+');
     else
         R_c=cholupdate(R_c, y*r0, '-');
     end;

     Rs=R_c*s;
     Bs=R_c'*Rs;

     r1=sqrt(Rs'*Rs);
     r2=sqrt(b);
```

```
     R_c=cholupdate(R_c, y /r2, '+');
     R_c=cholupdate(R_c, Bs/r1, '-');

     iter=iter+1;
     A=[A;iter,f_p,norm(g_p)];
     g_c=g_p;
     x_c=x_p;
end;
x=x_c;
```

A.3 Das "optimal ϕ"-Verfahren

```
function[x, iter, f_eval, A]=optphi(f_name, param, x_c, max_iter, tol);
%
% optimal-phi-Verfahren
%
% Input:
%
% f_name:   zu minimierende Zielfunktion [f,g,H]=f_name(x)
% param:    Parameter fuer die Zielfunktion
% x_c:      Startvektor des Verfahrens
% max_iter: maximale Anzahl der Iterationen
% tol:      Abbruchgrenze fuer die Gradientennorm
%
% Output:
%
% x:        Loesungsnaeherung, mit der das Verfahren abbricht
% iter:     Anzahl der Iterationen bis zum Abbruch
% f_eval:   Anzahl der Zielfunktionsauswertungen bis zum Abbruch,
%           bei versuchter Nulldivision wird -1 zurueckgegeben
% A:        Matrix mit Zielfunktionswerten und -Gradientennormen
%
%
[f_c,g_c]=feval(f_name,x_c,param);
n=length(x_c);
R_c=sqrt(abs(f_c))*eye(n);

f_eval=1;
iter=0;
A=[iter,f_c,norm(g_c)];

while(norm(g_c) > tol)&(iter <= max_iter)&(f_eval >= 0)

     p=-R_c\(R_c'\g_c);
```

```
[t,f_p,g_p,f_ev]=wolfe3(x_c,p,f_c,g_c,f_name,param);
f_eval=f_eval+f_ev;

s=t*p;
y=g_p-g_c;
b=y'*s;
x_p=x_c +s;

if iter==0
     a_1=R_c'\y;
     a=a_1'*a_1;
     R_c=sqrt(a/b)*R_c;
end;

Rs=R_c*s;
Bs=R_c'*Rs;
c=Rs'*Rs;
a_1=R_c'\y;
a=a_1'*a_1;

r1=sqrt(Rs'*Rs);
r2=sqrt(b);

R_c=cholupdate(R_c, y /r2, '+');
R_c=cholupdate(R_c, Bs/r1, '-');

nenner=(1-n)*(a*c-b^2);

if nenner==0
     fprintf('\n optphi: versuchte Nulldivision \n');
     f_eval=-1;
else
    phi=1+((a-b)*b)/nenner;
    r3=sqrt(abs(1-phi));
    v=y/b -Bs/c;

    if ((1-phi) >= 0)
        R_c=cholupdate(R_c, v*r1*r3, '+');
    else
        R_c=cholupdate(R_c, v*r1*r3, '-');
    end;
end;

iter=iter+1;
```

```
    A=[A;iter,f_p,norm(g_p)];
    g_c=g_p;
    x_c=x_p;
end;
x=x_c;
```

A.4 Die Wolfe-Schrittweite

```
function[t, f_p, g_p, f_eval]=...
        wolfe3(x_c, p, f_c, g_c, f_name, param)
%
% Berechnung einer Wolfe-Schrittweite
%
% Algorithmus siehe Dennis-Schnabel S. 328ff.
%
% Input:
%
% x_c:      aktuelle Naeherung
% p:        aktuelle (Abstiegs-)Suchrichtung
% f_c:      Zielfunktionswert in x_c
% g_c:      Zielfunktionsgradient in x_c
% f_name:   zu minimierende Zielfunktion [f,g,H]=f_name(x)
% param:    Parameter fuer die Zielfunktion
%
% Output:
%
% t:        Wolfe-Schrittweite
% f_p:      Zielfunktionswert in der neuen Naeherung x_p:=x_c+t*p
% g_p:      Zielfunktionsgradient in x_p
% f_eval:   Anzahl der Zielfunktionsauswertungen bis zum Abbruch
%
%
n=length(x_c);
maxstep=10^9;
steptol=10^(-9);
retcode=2;
maxtaken=0;
f_eval=0;
D=eye(n);
newtlen=norm(D*p);
alpha=0.0001;
beta=0.9;

if (newtlen > maxstep)
```

```
     p=p*(maxstep/newtlen);
     newtlen=maxstep;
end;

init_slope=g_c'*p;

if init_slope>=0
     error('wolfe3: p is not a descent direction !')
end;

rellength=max(abs(p)./ max(abs(x_c),diag(D)));
min_t=steptol/rellength;
t=1;

while (retcode==2)
x_p = x_c +t*p;
[f_p,g_p]=feval(f_name,x_p,param);
f_eval=f_eval+1;

if (f_p <= f_c+alpha*t*init_slope)
     new_slope=g_p'*p;

     if(new_slope < beta*init_slope)

          if ((t==1)&(newtlen < maxstep))
               max_t= maxstep/newtlen;
               t_prev=t;
               f_prev=f_p;
               t= min(2*t,max_t);
               x_p=x_c+t*p;
               [f_p,g_p]=feval(f_name,x_p,param);
               f_eval=f_eval+1;

       if(f_p <= f_c+alpha*t*init_slope)
                    new_slope=g_p'*p;
               end;

               while((f_p <= f_c+alpha*t*init_slope)&...
                     (new_slope<beta*init_slope)&(t<max_t))
                    t_prev=t;
                    f_prev=f_p;
                    t=min(2*t,max_t);
                    x_p=x_c+t*p;
                    [f_p,g_p]=feval(f_name,x_p,param);
                    f_eval=f_eval+1;
```

```
            if(f_p <= f_c+alpha*t*init_slope)
                 new_slope=g_p'*p;
            end;
       end;

       if (t >= max_t)
            error('wolfe3: stepsize longer than permitted !');
       end;
  end;

  if ((t < 1)|((t > 1)&(f_p > f_c+alpha*t*init_slope)))
       t_lo=min(t,t_prev);
       t_diff=abs(t_prev -t);

       if (t < t_prev)
            f_lo=f_p;
            f_hi=f_prev;
       else
            f_lo=f_prev;
            f_hi=f_p;
       end;

       t_incr=-new_slope*t_diff^2/...
              (2*(f_hi -(f_lo + new_slope*t_diff)));

       if(t_incr < 0.2*t_diff)
            t_incr=0.2*t_diff;
       end;

       t=t_lo+t_incr;
       x_p=x_c+t*p;
       [f_p,g_p]=feval(f_name,x_p,param);
       f_eval=f_eval+1;

       if(f_p > f_c+alpha*t*init_slope)
            t_diff=t_incr;
            f_hi=f_p;
       else
            new_slope=g_p'*p;

            if (new_slope < beta*init_slope)
                 t_lo=t;
                 t_diff=t_diff-t_incr;
                 f_lo=f_p;
```

```
                end;
            end;

            while( (new_slope < beta*init_slope)&(t_diff >= min_t))
                t_incr= - new_slope*t_diff^2/...
                        (2*(f_hi -(f_lo + new_slope*t_diff)));

                if(t_incr < 0.2*t_diff)
                    t_incr=0.2*t_diff;
                end;

                t=t_lo+t_incr;
                x_p=x_c+t*p;
                [f_p,g_p]=feval(f_name,x_p,param);
                f_eval=f_eval+1;

                if(f_p > f_c+alpha*t*init_slope)
                    t_diff=t_incr;
                    f_hi=f_p;
                else
                    new_slope=g_p'*p;

                    if (new_slope < beta*init_slope)
                        t_lo=t;
                        t_diff=t_diff-t_incr;
                        f_lo=f_p;
                    end;
                end;
            end;

            if(new_slope < beta*init_slope)
                f_p=f_lo;
                x_p=x_c+t_lo*p;
            end;
        end;
    end;

retcode=0;

if (t*newtlen>0.99*maxstep)
    maxtaken=1;
end;

elseif (t< min_t)
    retcode=1;
```

```
     fprintf('\n wolfe3: retcode=1 no wolfe-stepsize computable !\n');
     x_p=x_c;
     f_p=f_c;
     g_p=g_c;
else
     if (t==1)
          t_temp=-init_slope/(2*(f_p -f_c -init_slope ));
     else
          a=(1/(t-t_prev))*((1/t^2)*(f_p -f_c -t*init_slope)-...
            (1/t_prev^2)*(f_prev -f_c -t_prev*init_slope));
          b=(1/(t-t_prev))*((-t_prev/t^2)*(f_p -f_c -t*init_slope)+...
            (t/t_prev^2)*(f_prev -f_c -t_prev*init_slope));
          disc=b^2-3*a*init_slope;

          if a==0
               t_temp=-init_slope/(2*b);
          else
               t_temp=(-b +sqrt(disc))/(3*a);
          end;

          if (t_temp > 0.5*t)
               t_temp=0.5*t;
          end;
     end;

     t_prev=t;
     f_prev=f_p;

     if (t_temp <=0.1*t)
          t=0.1*t;
     else
          t=t_temp;
     end;
end;
end;

if (f_p > f_c +alpha*t*g_c'*p)|(g_p'*p < beta*g_c'*p)
     error('wolfe3: wolfe-conditions not fulfilled !');
end;
```

Anhang B

SAS-Quelltexte

B.1 Vergleich von BFGS- und DW-Verfahren

```
data alles;
infile 'q:\sas-opt\ergebnisse.dat' expandtabs;
input func n m iter_b feval_b iter_d feval_d iter_p feval_p @@;
run;

data ohne_phi;
set alles;
drop iter_p feval_p;
if iter_b > 1000 then delete;
run;

data iter;
set ohne_phi;
drop feval_b feval_d;
if iter_b < iter_d
then do
score_iter_d=(iter_d-iter_b)/iter_b;
score_iter_b=0;
end;
if iter_b > iter_d
then do
score_iter_b=(iter_b-iter_d)/iter_d;
score_iter_d=0;
end;
if iter_b eq iter_d
then do
score_iter_b=0;
score_iter_d=0;
end;
```

```
proc print;
run;

proc means data=iter;
var iter_b score_iter_b iter_d score_iter_d;
run;

data feval;
set ohne_phi;
drop iter_b iter_d;
if feval_b < feval_d
then do
score_feval_d=(feval_d-feval_b)/feval_b;
score_feval_b=0;
end;
if feval_b > feval_d
then do
score_feval_b=(feval_b-feval_d)/feval_d;
score_feval_d=0;
end;
if feval_b eq feval_d
then do
score_feval_b=0;
score_feval_d=0;
end;

proc print;
run;

proc means data=feval;
var feval_b score_feval_b feval_d score_feval_d;
run;
```

B.2 Vergleich aller drei Verfahren

```
data alles;
infile 'q:\sas-opt\ergebnisse.dat' expandtabs;
input func n m iter_b feval_b iter_d feval_d iter_p feval_p @@;
run;

data mit_phi;
set alles;
if iter_b > 1000 then delete;
```

```
if feval_p=0 then delete;
run;

data iter;
set mit_phi;
drop feval_b feval_d feval_p;
if iter_b eq min(iter_b,iter_d,iter_p)
then do
score_iter_d=(iter_d-iter_b)/iter_b;
score_iter_P=(iter_P-iter_b)/iter_b;
score_iter_b=0;
end;
if iter_d eq min(iter_b,iter_d,iter_p)
then do
score_iter_b=(iter_b-iter_d)/iter_d;
score_iter_p=(iter_p-iter_d)/iter_d;
score_iter_d=0;
end;
if iter_p eq min(iter_b,iter_d,iter_p)
then do
score_iter_b=(iter_b-iter_p)/iter_p;
score_iter_d=(iter_d-iter_p)/iter_p;
score_iter_p=0;
end;

proc print;
run;

proc means data=iter;
var iter_b score_iter_b iter_d score_iter_d iter_p score_iter_p;
run;

data feval;
set mit_phi;
drop iter_b iter_d iter_p;
if feval_b eq min(feval_b,feval_d,feval_p)
then do
score_feval_d=(feval_d-feval_b)/feval_b;
score_feval_P=(feval_P-feval_b)/feval_b;
score_feval_b=0;
end;
if feval_d eq min(feval_b,feval_d,feval_p)
then do
score_feval_b=(feval_b-feval_d)/feval_d;
score_feval_p=(feval_p-feval_d)/feval_d;
```

```
score_feval_d=0;
end;
if feval_p eq min(feval_b,feval_d,feval_p)
then do
score_feval_b=(feval_b-feval_p)/feval_p;
score_feval_d=(feval_d-feval_p)/feval_p;
score_feval_p=0;
end;

proc print;
run;

proc means data=feval;
var feval_b score_feval_b feval_d score_feval_d feval_p score_feval_p;
run;
```

Literaturverzeichnis

[1] C. G. Broyden, J. E. Dennis Jr. and J. J. Moré, *On the local and superlinear convergence of quasi-Newton methods*, J. Inst. Maths Applics 12 (1973), pp. 223-246.

[2] R. H. Byrd, D. C. Liu and J. Nocedal, *On the behavior of Broyden's class of quasi-Newton methods*, SIAM J. Optim. 2 (1992), pp. 533-557.

[3] R. H. Byrd and J. Nocedal, *A tool for the analysis of quasi-Newton methods*, SIAM J. Numer. Anal. 26 (1989), pp. 727-739.

[4] Y.-H. Dai, *Convergence properties of the BFGS algorithm*, SIAM J. Optim. 13 (2002), pp. 693-701.

[5] J. E. Dennis Jr., J. Martínez and R. A. Tapia, *Convergence theory for the unstructured BFGS secant method with an application to nonlinear least squares*, JOTA 61 (1989), pp. 161-178.

[6] J. E. Dennis Jr. and J. J. Moré, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp. 28 (1974), pp. 549-560.

[7] J. E. Dennis Jr. and J. J. Moré, *Quasi-Newton methods, motivation and theory*, SIAM Review 19 (1977), pp. 46-89.

[8] J. E. Dennis Jr. and R. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*, Prentice-Hall, Englewood Cliffs (1983).

[9] J. E. Dennis Jr. and H. Wolkowicz, *Sizing and least-change secant methods*, SIAM J. Numer. Anal. 30 (1993), pp. 1291-1314.

[10] R. Fletcher, *Practical Methods of Optimization*, John Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore (1987).

[11] C. Geiger und C. Kanzow, *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*, Springer-Verlag, Berlin-Heidelberg-New York (1999).

[12] L. Han and G. Liu, *Global analysis of the Dennis-Wolkowicz least-change secant algorithm*, SIAM J. Optim. 8 (1998), pp. 813-832.

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

[14] D.-H. Li and M. Fukushima, *A modified BFGS method and its global convergence in nonconvex minimization*, J. Comput. Appl. Math. 129 (2001), pp. 15-35.

[15] D.-H. Li and M. Fukushima, *On the global convergence of the BFGS method for nonconvex optimization problems*, SIAM J. Optim. 11 (2001), pp. 1054-1064.

[16] J. J. Moré, B. S. Garbow and K. E. Hillstrom, *Testing unconstrained optimization software*, ACM Transactions on mathematical software 7 (1981), pp. 17-41.

[17] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-Verlag, Berlin-Heidelberg-New-York (1999).

[18] J. D. Pearson, *Variable metric methods of minimization*, Comput. J. 12 (1969), pp. 171-178.

[19] M. J. D. Powell, *Some global convergence properties of a variable metric alogrithm for minimization without exact line searches*, SIAM-AMS Proceedings 9 (1976), R. W. Cottle and C. E. Lemke eds., pp. 53-72.

[20] W. Warth und J. Werner, *Effiziente Schrittweitenfunktionen bei unrestringierten Optimierungsaufgaben*, Computing 19 (1977), S. 59-72.

[21] J. Werner, *Global and superlinear convergence of the Dennis-Wolkowicz Quasi-Newton Method*, Göttingen (1998).

[22] J. Werner, *Globale und lokale Konvergenz des BFGS-Verfahrens*, Göttingen (2002).

[23] J. Werner, *Numerische Mathematik II*, Vieweg-Verlag, Braunschweig-Wiesbaden (1992).

[24] J. Werner, *Über die globale Konvergenz von Variable-Metrik-Verfahren bei nicht-exakter Schrittweitenbestimmung*, Numer. Math. 31 (1978), S. 321-334.

Danksagung

Zum Abschluß dieser Arbeit möchte ich noch kurz die Gelegenheit nutzen, allen Personen zu danken, die mich während meines Studiums und der Entstehung dieser Arbeit moralisch, fachlich und nicht zuletzt auch finanziell unterstützt haben. Besonderer Dank gilt dabei Herrn Prof. Dr. Jochen Werner für die interessante Themenstellung sowie die gute und stets freundliche Betreuung. Ferner danke ich Herrn dipl.-math. Stefan Härtel für die Hilfe beim Layout und weiteren TEXnischen Fragen.