Quasi-Newton-Verfahren für nichtdifferenzierbare konvexe Optimierungsaufgaben

Diplomarbeit

vorgelegt von
Anja Rüdiger
aus
Salzgitter

angefertigt im
Institut für Numerische und Angewandte Mathematik
der Georg-August-Universität zu Göttingen
2002

Inhaltsverzeichnis

Kapitel 1

Einleitung

In den letzten Jahrzehnten beschäftigte man sich immer mehr mit dem Problem der nichtdifferenzierbaren Optimierungaufgaben, nachdem viele aktuelle Problemstellungen dieser Art in Industrie und Wissenschaft vorkamen. Hierbei handelt es sich um diejenigen Optimierungsaufgaben, welche einen Punkt suchen, in dem eine nicht notwendig differenzierbare Zielfunktion minimal ist. Man möchte einen Algorithmus finden, mit dem man Problemstellungen dieser Art schnell und effizient lösen kann. Die Anfänge dieser Entwicklung lagen unter anderem in Russland, wo insbesondere N. Z. Shor [18] die „Methoden der Subgradienten“ vorstellte. Für Optimierungsaufgaben mit differenzierbaren Zielfunktionen gibt es zahlreiche Lösungsvorschläge (siehe z.B. [14] oder [6]). Bei denen wird oft neben der Zielfunktion auch der Gradient der Zielfunktion zur Berechnung der Lösung verwendet. Analog hierzu wurden von N. Z. Shor die Subgradienten – auf die in dieser Arbeit noch genauer eingegangen werden soll – als Ersatz für die nicht unbedingt vorhandenen Gradienten eingeführt. Dazu kamen viele Lösungsvorschläge, allerdings waren die dazugehörenden Algorithmen sehr komplex und die Konvergenzeigenschaften gewöhnlich schwächer, als bei den klassischen Methoden mit differenzierbaren Zielfunktionen (genaueres siehe in [9]).
Diese Arbeit beschäftigt sich intensiv mit dem 1998 von R. Mifflin, D. Sun und L. Qi erschienenen Paper [13] über einen Quasi-Newton-Bündel-Algorithmus zur Lösung nichtdifferenzierbarer konvexer Optimierungsaufgaben. Es handelt sich hierbei um ein unrestringiertes Optimierungsproblem, also eine Aufgabe ohne Nebenbedingungen beziehungsweise Restriktionen. Da nichtdifferenzierbare Optimierungsprobleme ärmere analytische Eigenschaften als differenzierbare haben, bedienen sie sich oft der Hilfe von differenzierbaren Optimierungsproblemen. So auch in dieser Arbeit. Unsere nichtdifferenzierbare Zielfunktion soll durch die sogenannte Moreau-Yosida-Regularisierung angenähert werden, eine Funktion die neben ihrer Differenzierbarkeit auch noch andere nützliche Eigenschaften besitzt. Hierauf soll in Kapitel 3 ein-

gegangen werden. Die Moreau-Yosida-Regularisierung stellt sozusagen eine Verbindung zwischen der klassischen und der nichtdifferenzierbaren Optimierung dar. Da man diese Regularisierung und ihren Gradienten nicht exakt berechnen kann – wie wir später sehen werden – verwendet die hier dargestellte Methode stückweise lineare Approximationen, welche als die sogenannte „Bündel-Idee“ in Kapitel 4 beschrieben werden. Der Vorteil dieser Bündel-Methode liegt darin, dass nun für die Berechnung ein quadratisches Programm gelöst werden kann. Auf diese Approximation lässt sich das bewährte Quasi-Newton-Verfahren anwenden.
Wir wollen zuerst die Grundlagen in Kapitel 2 und 3 schaffen, die wir für diese Arbeit benötigen, um dann für die eben kurz erläuterte Problemstellung in Kapitel 4, 5 und 7 einen Algorithmus aufzustellen, mit dem unter weiteren Voraussetzungen globale und superlineare Konvergenz in Kapitel 6 und 8 nachgewiesen werden kann. Zum Schluss soll noch numerisch die Konvergenzgeschwindigkeit an zwei Beispielen getestet werden.

Um die Problemstellung etwas mathematischer auszudrücken, beschreiben wir nun die Aufgabenstellung genauer. Wir betrachten die unrestringierte Optimierungsaufgabe

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

wobei die Zielfunktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ eine nicht notwendig differenzierbare, konvexe Funktion sei.
Diese Funktion soll durch die sogenannte *Moreau-Yosida-Regularisierung*

$$F_M(x) := \min_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{1}{2}\|y - x\|_M^2 \right\}$$

angenähert werden. Hierbei ist die transformierte Norm $\|\cdot\|_M$ durch

$$\|x\|_M := \|M^{\frac{1}{2}}x\|_2 = (x^T M x)^{\frac{1}{2}}, \quad x \in \mathbb{R}^n$$

mit einer symmetrischen, positiv definiten Matrix $M \in \mathbb{R}^{n\times n}$ gegeben. Diese Norm wird auch als elliptische Norm bezeichnet. Es handelt sich bei F_M um eine stetig differenzierbare Funktion, die die gleichen Minima wie die eigentliche Funktion f hat, wie wir gleich sehen werden. Deshalb versucht man die Aufgabe (P) mit Hilfe der neuen Aufgabe

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

zu lösen.

R. T. Rockafellar [17] betrachtete mit als Erster dieses Problem und legte die Grundlage in seinem *proximal point* -Algorithmus. Dort konstruierte er eine Folge $\{x_k\}$, die aus $x_{k+1} \approx \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2}\|x - x_k\|_M^2 \right\}$ berechnet wurde. Darauf aufbauend erforschten J. Bonnans, J. Gilbert, C. Lemaréchal und C. Sagastizábal [1] Optimierungsmethoden, welche die Moreau-Yosida-Regularisierung und das Quasi-Newton-Verfahren miteinander verbinden. C. Zhu [23] wies lineare Konvergenz der aus dem Algorithmus gelieferten Folge gegen eine Lösung x^* nach. C. Lemaréchal und C. Sagastizábal [12] untersuchten implementierbare Versionen der Methoden und zeigten auch für komplexere Problemstellungen numerisch ihr gutes Verhalten. M. Fukushima und L. Qi [5] schlugen danach einen Algorithmus vor, mit dem globale und superlineare Konvergenz nachgewiesen werden konnte. In diesem Aufsatz wird das Thema erneut aufgegriffen und genau analysiert, wobei einige Veränderungen gegenüber den vorherigen Arbeiten durchgeführt werden.

Kapitel 2

Konvexe Funktionen

Konvexe Funktionen spielen in der unrestringierten Optimierung eine wichtige Rolle. Dies werden wir gleich näher begründen. Wie erwähnt betrachten wir in dieser Arbeit ein Optimierungsproblem, bei dem die Zielfunktion konvex ist. Für Konvergenzaussagen werden wir gleichmäßige Konvexität voraussetzen, da so die Existenz einer eindeutigen Lösung gesichert ist. Dieses und die wichtigsten Eigenschaften konvexer und gleichmäßig konvexer Funktionen sollen in diesem Kapitel aufgezeigt werden. Außerdem werden die auch schon erwähnten Subgradienten eingeführt.

2.1 Eigenschaften

Wir beginnen mit den Definitionen und Eigenschaften konvexer und gleichmäßig konvexer Funktionen. Die Beweise werden in diesem Abschnitt nicht durchgeführt; diese kann man aber zum Beispiel in [16] und [7] finden.

Bemerkung: Im ganzen Aufsatz sei $\|\cdot\|$ stets die euklidische Norm auf dem $\mathbb{R}^n$ beziehungsweise die zugeordnete Matrixnorm.

- Eine Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ heißt *konvex*, wenn

$$(1-t)f(x) + tf(y) - f((1-t)x + ty) \geq 0$$

  für alle $x, y \in \mathbb{R}^n, t \in [0,1]$ gilt.

- Eine Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ heißt *gleichmäßig konvex*, wenn eine Konstante $\alpha > 0$ mit

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

  für alle $x, y \in \mathbb{R}^n, t \in [0,1]$ existiert.

- Sei nun $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ konvex. Dann gilt:

  1. f ist auf dem $\mathbb{R}^n$ stetig.
  2. f besitzt in jedem $x \in \mathbb{R}^n$ in jede Richtung $h \in \mathbb{R}^n$ eine *Richtungsableitung* $f'(x;h)$, das heißt

     $$f'(x;h) := \lim_{t \to 0+} \frac{f(x+th) - f(x)}{t}$$

     existiert für alle $x, h \in \mathbb{R}^n$.
  3. Die sogenannte *Gateaux-Variation* $f'(x;\cdot) : \mathbb{R}^n \longrightarrow \mathbb{R}$ ist bei festem $x \in \mathbb{R}^n$ in der zweiten Komponente konvex und es gilt $f'(x;h) \leq f(x+h) - f(x)$ für alle $h \in \mathbb{R}^n$.
  4. Gilt $f'(x;h) = g(x)^T h$ mit einem $g(x) \in \mathbb{R}^n$ für alle $h \in \mathbb{R}^n$, dann existieren alle partiellen Ableitungen $\partial f / \partial x_j, j = 1, ..., n$. Dann heißt f in x *partiell differenzierbar* und es ist auch $\nabla f(x) = g(x)$.

- Ist $F : \mathbb{R}^n \longrightarrow \mathbb{R}$ in $x \in \mathbb{R}^n$ stetig partiell differenzierbar, sind also alle partiellen Ableitungen in x stetig, so heißt F in x *stetig differenzierbar*.

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

Wie schon in der Einleitung beschrieben, wollen wir mit Hilfe von Subgradienten arbeiten, da der Gradient einer nicht notwendig differenzierbaren Funktion nicht unbedingt existiert. Die Subgradienten sollen nun kurz eingeführt werden.

- Bei vorgegebenem $x \in \mathbb{R}^n$ heißt für eine konvexe Funktion $f : \mathbb{R}^n \longrightarrow \mathbb{R}$

  $$\partial f(x) := \{z \in \mathbb{R}^n : z^T(y-x) \leq f(y) - f(x) \quad \text{für alle } y \in \mathbb{R}^n\}$$

  das *Subdifferential* von f in x, Elemente von $\partial f(x)$ heißen *Subgradienten*. Dann gilt:

  1. Für jedes $x \in \mathbb{R}^n$ ist das Subdifferential $\partial f(x)$ nichtleer, konvex und kompakt.
  2. Es ist

     $$f'(x;h) = \max_{z \in \partial f(x)} z^T h \quad \text{für alle } h \in \mathbb{R}^n.$$

  3. Für $x, y \in \mathbb{R}^n$ gilt $(z_x - z_y)^T(x-y) \geq 0$ für alle $z_x \in \partial f(x), z_y \in \partial f(y)$. Dies wird auch mit *Monotonie der Subgradienten* bezeichnet.

4. Ist $F : \mathbb{R}^n \longrightarrow \mathbb{R}$ differenzierbar, so ist der Gradient der Funktion das einzige Element in $\partial F(x)$.

□

Das folgende Lemma wird in einigen Beweisen nützlich sein können. Wir werden es in dieser Arbeit häufig auch für differenzierbare Funktionen verwenden.

Lemma 2.1 *Die drei folgenden Aussagen sind für alle $x, y \in \mathbb{R}^n$ äquivalent:*

- *$f : \mathbb{R}^n \longrightarrow \mathbb{R}$ ist gleichmäßig konvex mit einer Konstanten $\alpha > 0$.*
- *Es gilt $f(y) \geq f(x) + z^T(y - x) + \frac{\alpha}{2}\|y - x\|^2$ für alle $z \in \partial f(x)$.*
- *Es gilt $(z_y - z_x)^T(y - x) \geq \alpha\|y - x\|^2$ für alle $z_x \in \partial f(x), z_y \in \partial f(y)$.*

Diese Eigenschaften werden im Folgenden immer wieder verwendet.

Bemerkung: Ein Grund warum konvexe Funktionen eine wichtige Rolle in der Optimierung spielen ist der, dass ein stationärer Punkt der Zielfunktion f eine globale Lösung der Aufgabe (P), f auf dem $\mathbb{R}^n$ zu minimieren, ist. Hierbei heißt ein Punkt stationär, wenn $f'(x^*; h) \geq 0$ für alle $h \in \mathbb{R}^n$ ist und eine Lösung global, wenn $f(x^*) \leq f(x)$ für alle $x \in \mathbb{R}^n$ ist. Diese Behauptung folgt ganz einfach aus der gerade beschriebenen Tatsache

$$0 \leq f'(x^*; h) \leq f(x^* + h) - f(x^*)$$

für alle $h \in \mathbb{R}^n$. Nun setzen wir $h := x - x^*$ und haben das gewünschte Ergebnis.

□

2.2 Wichtige Sätze

Der nächste Satz zeigt, warum es nützlich sein kann, bei Konvergenzaussagen gleichmäßige Konvexität für die Zielfunktion vorauszusetzen. Die Beweisidee haben wir von einem ähnlichen Beweis aus [20] (Lemma 2.6, S.171) übernommen.

Satz 2.2 *Ist $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ auf dem $\mathbb{R}^n$ gleichmäßig konvex, so hat die unrestringierte Optimierungsaufgabe , f auf dem $\mathbb{R}^n$ zu minimieren, genau eine Lösung.*

Beweis: Man wähle $x_0 \in \mathbb{R}^n$ beliebig und definiere hierzu die Niveaumenge $L_0 := \{x \in \mathbb{R}^n | f(x) \leq f(x_0)\}$. Die Menge ist wegen der Konvexität von f auch selbst konvex. Außerdem ist sie wegen der aus der Konvexität von f folgenden Stetigkeit auch abgeschlossen. Die Beschränktheit soll nun gezeigt werden.
Sei $x \in L_0$ beliebig. Für $t \in (0, 1]$ ist

$$f(x) - f(x_0) - \frac{f(x_0 + t(x - x_0)) - f(x_0)}{t} \geq \frac{\alpha}{2}(1 - t)\|x - x_0\|^2,$$

wie man nach Division durch t aus der gleichmäßigen Konvexität mit einer positiven Konstanten α erhält. Durch den Grenzübergang $t \to 0+$ erhalten wir

$$\begin{aligned} 0 &\geq f(x) - f(x_0) \\ &\geq \frac{\alpha}{2}\|x - x_0\|^2 + f'(x_0; x - x_0) \\ &\geq \frac{\alpha}{2}\|x - x_0\|^2 + z^T(x - x_0) \\ &\geq \frac{\alpha}{2}\|x - x_0\|^2 - \|z\|\|x - x_0\|, \end{aligned}$$

mit einem beliebigen $z \in \partial f(x_0)$ für alle $x \in L_0$, unter Benutzung des in Kapitel 2.1 beschriebenen Zusammenhanges zwischen der Richtungsableitung und dem Subdifferential. Nun folgt, dass $\|x - x_0\| \leq 2\|z\|/\alpha$ für alle $x \in L_0, x \neq x_0$ gilt. Daher ist die Niveaumenge beschränkt, insgesamt also kompakt. Da jede stetige Funktion auf einer kompakten Menge ihr Extremum annimmt, nimmt auch die konvexe Funktion f auf L_0 ihr Minimum an. Um die Eindeutigkeit zu zeigen, wählen wir x^* und x^{**} als zwei verschiedene Lösungen. Da f konvex ist, muss auch $\frac{1}{2}(x^* + x^{**})$ eine Lösung sein, das heißt also $f(x^*) = f(x^{**}) = f(\frac{1}{2}(x^* + x^{**}))$. Aus der gleichmäßigen Konvexität bekommen wir für $t = \frac{1}{2}$

$$0 = \frac{1}{2}f(x^*) + \frac{1}{2}f(x^{**}) - f\left(\frac{1}{2}(x^* + x^{**})\right) \geq \frac{\alpha}{8}\|x^* - x^{**}\|^2$$

und daher $x^* = x^{**}$. Insgesamt ist der Satz bewiesen.

□

Wenn wir also gleichmäßige Konvexität für die Zielfunktion voraussetzen, ist uns eine eindeutige Lösung unserer Minimierungsaufgabe sicher.
Zur Erinnerung soll nun noch ein Lemma angegeben werden, welches sehr häufig in den folgenden Beweisen genutzt wird und dessen Beweis man zum Beispiel in [19] (Lemma 2.6, S.22) findet:

Lemma 2.3 *Seien $\lambda_{min}(A)$ ein minimaler und $\lambda_{max}(A)$ ein maximaler Eigenwert einer symmetrischen Matrix $A \in \mathbb{R}^{n\times n}$. Dann ist*

$$\lambda_{min}(A)\|x\|^2 \leq x^T Ax \leq \lambda_{max}(A)\|x\|^2 \quad \textit{für alle } x \in \mathbb{R}^n.$$

Damit haben wir die Grundlagen gelegt und können nun im nächsten Kapitel die in der Einleitung erwähnte Moreau-Yosida-Regularisierung einführen.

Kapitel 3

Die Moreau - Yosida - Regularisierung

Um mit der Moreau-Yosida-Regularisierung arbeiten zu können, werden wir hier die wichtigsten Eigenschaften zusammenfassen und beweisen. Danach soll die Regularisierung anhand eines Beispiels verdeutlicht werden.

3.1 Eigenschaften

Wir gehen vor allem der Frage nach, warum statt des Ausgangsproblems (P) das neue Problem

$$(\tilde{P}) \qquad \text{Minimiere} \quad F_M(x) := \min_{y \in \mathbb{R}^n} \left\{ f(y) + \tfrac{1}{2}\|y - x\|_M^2 \right\}, \quad x \in \mathbb{R}^n$$

gelöst werden kann. Die Beweise lassen sich zum Teil auch in [8] (ab S.317) nachvollziehen.

Satz 3.1 *Sei $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ konvex und $M \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit. Wir betrachten die Aufgaben*

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

und

$$(P_{M,x}) \qquad \textit{Minimiere} \quad f_{M,x}(y) := f(y) + \tfrac{1}{2}\|y - x\|_M^2, \quad y \in \mathbb{R}^n.$$

Seien mit $\lambda_{min}(M)$ der kleinste und $\lambda_{max}(M)$ der größte Eigenwert von M bezeichnet. Dann gilt:

1. *Die Aufgabe $(P_{M,x})$ besitzt für jedes $x \in \mathbb{R}^n$ eine eindeutige Lösung*

$p(x) \in \mathbb{R}^n$ und es ist

$$0 \in \partial f_{M,x}(p(x)) = \partial f(p(x)) + M(p(x) - x).$$

2. *Es ist $p : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ global lipschitzstetig mit der Lipschitzkonstanten $\lambda_{max}(M)/\lambda_{min}(M)$ bezüglich der euklidischen Norm, das heißt*

$$\|p(x_1) - p(x_2)\| \leq \frac{\lambda_{max}(M)}{\lambda_{min}(M)} \|x_1 - x_2\|$$

gilt für alle $x_1, x_2 \in \mathbb{R}^n$.

3. *Die Moreau-Yosida-Regularisierung F_M ist konvex und stetig differenzierbar und es ist $\nabla F_M(x) = M(x - p(x)) \in \partial f(p(x))$.*

4. *$\nabla F_M(x)$ ist global lipschitzstetig mit der Lipschitzkonstanten $\|M\|$ bezüglich der euklidischen Norm, das heißt*

$$\|\nabla F_M(x_1) - \nabla F_M(x_2)\| \leq \|M\| \|x_1 - x_2\|$$

gilt für alle $x_1, x_2 \in \mathbb{R}^n$.

5. *Die Menge der Lösungen von (P) stimmt mit der Menge der Lösungen von ($\tilde{P}$) überein. Daher ist $x^* \in \mathbb{R}^n$ genau dann Lösung von (P), wenn $\nabla F_M(x^*) = 0$ beziehungsweise $x^* = p(x^*)$ ist.*

Beweis:

1. Um Satz 2.2 anwenden zu können, brauchen wir nur zu zeigen, dass $f_{M,x}$ gleichmäßig konvex ist.
Für $y, w \in \mathbb{R}^n$ und $t \in [0, 1]$ ist

$$\begin{aligned}
& (1-t) f_{M,x}(y) + t f_{M,x}(w) - f_{M,x}((1-t)y + tw) \\
= \; & (1-t) f(y) + (1-t)\frac{1}{2}\|y - x\|_M^2 + t f(w) + t\frac{1}{2}\|w - x\|_M^2 \\
& - f((1-t)y + tw) - \frac{1}{2}\|(1-t)y + tw - x\|_M^2 \\
= \; & \underbrace{(1-t) f(y) + t f(w) - f((1-t)y + tw)}_{\geq 0} + \frac{1}{2} t(1-t)\|y - w\|_M^2 \\
\geq \; & \frac{1}{2}\lambda_{min}(M) t(1-t)\|y - w\|^2,
\end{aligned}$$

also ist $f_{M,x}$ gleichmäßig konvex mit der positiven Konstanten $\lambda_{min}(M)$ und somit ist die Aufgabe $(P_{M,x})$ für jedes $x \in \mathbb{R}^n$ eindeutig lösbar.

Sei nun $p(x)$ die eindeutige Lösung. Da $f_{M,x}(p(x)) \leq f_{M,x}(w)$ für alle $w \in \mathbb{R}^n$ gilt, gilt auch $0^T(w - p(x)) \leq f_{M,x}(w) - f_{M,x}(p(x))$ und daher $0 \in \partial f_{M,x}(p(x))$.
Zum Schluss zeigen wir, dass $\partial f_{M,x}(y) = \partial f(y) + M(y-x)$ für beliebige $x, y \in \mathbb{R}^n$ ist. Hierfür zeigen wir zwei Richtungen. Als Erstes nehmen wir ein $z \in \partial f(y)$ und weisen $(z + M(y - x)) \in \partial f_{M,x}(y)$ nach. Für jedes $u \in \mathbb{R}^n$ ist dann

$$\begin{aligned}
[z + M(y-x)]^T(u-y) &= z^T(u-y) + (y-x)^T M(u-y) \\
&\leq f(u) - f(y) + (y-x)^T M(u-y) \\
&= f(u) - f(y) + \frac{1}{2}\|u-x\|_M^2 - \frac{1}{2}\|y-x\|_M^2 \\
&\quad -\frac{1}{2}(y-u)^T M(y-u) \\
&\leq f(u) - f(y) + \frac{1}{2}\|u-x\|_M^2 - \frac{1}{2}\|y-x\|_M^2 \\
&= f_{M,x}(u) - f_{M,x}(y).
\end{aligned}$$

Für die umgekehrte Richtung geben wir ein $z \in \partial f_{M,x}(y)$ beliebig vor und weisen $z \in \partial f(y) + M(y-x)$ nach. Für jedes $v \in \mathbb{R}^n$ ist dann

$$\begin{aligned}
[z - M(y-x)]^T(v-y) &\leq f_{M,x}(v) - f_{M,x}(y) - (y-x)^T M(v-y) \\
&= f(v) - f(y) + \frac{1}{2}\|v-x\|_M^2 - \frac{1}{2}\|y-x\|_M^2 \\
&\quad -(y-x)^T M(v-y) \\
&= f(v) - f(y) + \frac{1}{2}(v-y)^T M(v-y).
\end{aligned}$$

Setzt man hier $v := y + t(u-y)$ mit beliebigen $u \in \mathbb{R}^n, t \in (0,1]$, so erhält man, da f konvex ist,

$$\begin{aligned}
t[z - M(y-x)]^T(u-y) &\leq f(y + t(u-y)) - f(y) \\
&\quad +\frac{t^2}{2}(u-y)^T M(u-y) \\
&\leq t[f(u) - f(y)] + \frac{t^2}{2}(u-y)^T M(u-y).
\end{aligned}$$

Nach Division durch t und anschließendem Grenzübergang $t \to 0+$ ist also

$$[z - M(y-x)]^T(u-y) \leq f(u) - f(y) \quad \text{für alle } u \in \mathbb{R}^n.$$

Mit $y = p(x)$ haben wir den ersten Teil des Satzes bewiesen.

2. Seien nun $x_1, x_2 \in \mathbb{R}^n$. Wir verwenden, dass $M(x_1 - p(x_1)) \in \partial f(p(x_1))$, $M(x_2 - p(x_2)) \in \partial f(p(x_2))$ ist. Wegen der Monotonie der Subgradienten bei konvexen Funktionen gilt

$$(M(x_1 - p(x_1)) - M(x_2 - p(x_2)))^T (p(x_1) - p(x_2)) \geq 0$$

und daher

$$(x_1 - x_2)^T M(p(x_1) - p(x_2)) \geq (p(x_1) - p(x_2))^T M(p(x_1) - p(x_2)).$$

Nun bekommen wir

$$\begin{aligned} \lambda_{min}(M)\|p(x_1) - p(x_2)\|^2 &\leq (p(x_1) - p(x_2))^T M(p(x_1) - p(x_2)) \\ &\leq (x_1 - x_2)^T M(p(x_1) - p(x_2)) \\ &\leq \|M(x_1 - x_2)\|\|p(x_1) - p(x_2)\|, \end{aligned}$$

woraus wir

$$\begin{aligned} \lambda_{min}(M)\|p(x_1) - p(x_2)\| &\leq \sqrt{(x_1 - x_2)^T M^2 (x_1 - x_2)} \\ &\leq \sqrt{\lambda_{max}(M^2)\|x_1 - x_2\|^2} \\ &= \sqrt{(\lambda_{max}(M))^2\|x_1 - x_2\|^2} \\ &= \lambda_{max}(M)\|x_1 - x_2\|. \end{aligned}$$

erhalten. Jetzt folgt direkt die Lipschitzstetigkeit durch

$$\|p(x_1) - p(x_2)\| \leq \frac{\lambda_{max}(M)}{\lambda_{min}(M)}\|x_1 - x_2\|.$$

Somit haben wir den zweiten Teil des Satzes bewiesen.

3. Es ist

$$F_M(x) = f(p(x)) + \frac{1}{2}\|p(x) - x\|_M^2,$$

wobei $p(x)$ die eindeutige Lösung von $(P_{M,x})$ ist. Die Konvexität von F_M folgt durch einfaches nachrechnen. Seien hierzu $u, v \in \mathbb{R}^n, t \in [0,1]$ beliebig. Dann ist

$$\begin{aligned} F_M((1-t)u + tv) &= f(p((1-t)u + tv)) \\ &\quad + \frac{1}{2}\|p((1-t)u + tv) - [(1-t)u + tv]\|_M^2 \\ &\leq f((1-t)p(u) + tp(v)) \\ &\quad + \frac{1}{2}\|(1-t)p(u) + tp(v) - [(1-t)u + tv]\|_M^2 \\ &\leq (1-t)f(p(u)) + tf(p(v)) \\ &\quad + \frac{1}{2}\|(1-t)[p(u) - u] + t[p(v) - v]\|_M^2 \\ &\leq (1-t)F_M(u) + tF_M(v) \end{aligned}$$

und daher folgt die Behauptung. Da $F'_M(x;h)$ für alle $h \in \mathbb{R}^n$ existiert, reicht es zu zeigen, dass $F'_M(x;h) = [M(x-p(x))]^T h$ für alle $h \in \mathbb{R}^n$ gilt, da dann – wie in Kapitel 2.1 beschrieben – $\nabla F_M(x) = M(x-p(x))$ gilt. Seien also $h \in \mathbb{R}^n$ und $t \in (0,1]$ vorgegeben. Dann ist

$$\begin{aligned} F_M(x+th) - F_M(x) &\leq f(p(x)) + \frac{1}{2}\|p(x) - (x+th)\|_M^2 - f(p(x)) \\ &\quad -\frac{1}{2}\|p(x) - x\|_M^2 \\ &= \frac{1}{2}\|p(x) - (x+th)\|_M^2 - \frac{1}{2}\|p(x) - x\|_M^2 \\ &= t[M(x-p(x))]^T h + \frac{t^2}{2}\|h\|_M^2. \end{aligned}$$

Division durch t und anschließender Grenzübergang $t \to 0+$ liefern $F'_M(x;h) \leq [M(x-p(x))]^T h$. Entsprechend ist

$$\begin{aligned} F_M(x+th) - F_M(x) &\geq \frac{1}{2}\|p(x+th) - (x+th)\|_M^2 \\ &\quad -\frac{1}{2}\|p(x+th) - x\|_M^2 \\ &= t[M(x-p(x+th))]^T h + \frac{t^2}{2}\|h\|_M^2. \end{aligned}$$

Division durch t und anschließender Grenzübergang $t \to 0+$ liefern unter Benutzung der Stetigkeit von $p(\cdot)$, dass auch $F'_M(x;h) \geq [M(x-p(x))]^T h$ gilt. Damit ist $F'_M(x;h) = [M(x-p(x))]^T h$ und daher folgt die Behauptung. Wegen der aus der Lipschitzstetigkeit folgenden Stetigkeit von $p(\cdot)$ ist F_M stetig partiell differenzierbar beziehungsweise stetig differenzierbar.
Damit haben wir den dritten Teil des Satzes bewiesen.

4. Um die globale Lipschitzstetigkeit von ∇F_M zu zeigen, wählen wir $x_1, x_2 \in \mathbb{R}^n$ beliebig. Wir wissen, dass

$$\begin{aligned} \nabla F_M(x_1) - \nabla F_M(x_2) &= M(x_1 - p(x_1)) - M(x_2 - p(x_2)) \\ &= M(x_1 - x_2) - M(p(x_1) - p(x_2)) \end{aligned}$$

gilt. Eine Skalarproduktbildung beider Seiten mit

$$M^{-1}(\nabla F_M(x_1) - \nabla F_M(x_2))$$

und eine erneute Ausnutzung der Monotonie der Subgradienten $\nabla F_M(x) \in \partial f(p(x))$ für alle $x \in \mathbb{R}^n$ ergibt

$$
\begin{aligned}
& (\nabla F_M(x_1) - \nabla F_M(x_2))^T M^{-1}(\nabla F_M(x_1) - \nabla F_M(x_2)) \\
= \; & (M^{-1}(\nabla F_M(x_1) - \nabla F_M(x_2)))^T (M(x_1 - x_2) - M(p(x_1) - p(x_2))) \\
= \; & (\nabla F_M(x_1) - \nabla F_M(x_2))^T (x_1 - x_2) \\
& - \underbrace{(\nabla F_M(x_1) - \nabla F_M(x_2))^T (p(x_1) - p(x_2))}_{\geq 0} \\
\leq \; & (\nabla F_M(x_1) - \nabla F_M(x_2))^T (x_1 - x_2).
\end{aligned}
$$

Daraus folgt nun

$$
\begin{aligned}
& \frac{1}{\lambda_{max}(M)} \|\nabla F_M(x_1) - \nabla F_M(x_2)\|^2 \\
= \; & \lambda_{min}(M^{-1}) \|\nabla F_M(x_1) - \nabla F_M(x_2)\|^2 \\
\leq \; & (\nabla F_M(x_1) - \nabla F_M(x_2))^T M^{-1}(\nabla F_M(x_1) - \nabla F_M(x_2)) \\
\leq \; & (\nabla F_M(x_1) - \nabla F_M(x_2))^T (x_1 - x_2) \\
\leq \; & \|\nabla F_M(x_1) - \nabla F_M(x_2)\| \|x_1 - x_2\|,
\end{aligned}
$$

also insgesamt

$$\|\nabla F_M(x_1) - \nabla F_M(x_2)\| \leq \|M\| \|x_1 - x_2\|.$$

Damit ist der vierte Teil des Satzes bewiesen.

5. Für den letzten Teil des Satzes sind zwei Richtungen zu zeigen. Sei zunächst x^* eine Lösung von (P). Dann ist $0 \in \partial f(x^*) = \partial f_{M,x}(x^*) - M(x^* - x)$ für alle $x \in \mathbb{R}^n$, also auch $0 \in \partial f(x^*) = \partial f_{M,x^*}(x^*)$. Also nimmt die Funktion f_{M,x^*} in x^* ihr Minimum an, das heißt es gilt $x^* = p(x^*)$. Dann ist aber $\nabla F_M(x^*) = M(x^* - p(x^*)) = 0$. Da F_M konvex ist, nimmt F_M auf dem $\mathbb{R}^n$ ein Minimum an.
   Sei nun umgekehrt x^* eine Lösung von $(\tilde{P})$. Dann ist $\nabla F_M(x^*) = M(x^* - p(x^*)) = 0$, also nimmt die Funktion f_{M,x^*} in $x^* = p(x^*)$ ihr eindeutiges Minimum auf dem $\mathbb{R}^n$ an, da M positiv definit ist. Hieraus folgt $0 \in \partial f(x^*)$, so dass x^* auch eine Lösung von (P) ist.

Damit ist der Satz vollständig bewiesen.

□

3.2 Beispiel

Wir wollen nun ein einfaches Beispiel aus [7] (S.13) zur Veranschaulichung der Moreau-Yosida-Regularisierung angeben. Dazu sei $n = 1$ und $f(x) := |x|$. Das Problem

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

soll durch das Problem

$$(\tilde{P}) \qquad \text{Minimiere} \quad F_M(x) := \min_{y \in \mathbb{R}} \left\{ |y| + \tfrac{1}{2} M (y - x)^2 \right\}, \quad x \in \mathbb{R}$$

gelöst werden.

Hierfür berechnen wir zuerst das Minimum über alle $y \in \mathbb{R}$ von

$$f_{M,x}(y) := |y| + \frac{1}{2} M (y - x)^2$$

und betrachten davon die Lösung $p(x)$ in Abhängigkeit von drei Fällen

$$p(x) = \begin{cases} x - \dfrac{1}{M} & : \quad x > \dfrac{1}{M} \\ 0 & : \quad x \in \left[-\dfrac{1}{M}, \dfrac{1}{M} \right] \\ x + \dfrac{1}{M} & : \quad x < -\dfrac{1}{M}. \end{cases}$$

Nun erhalten wir für die Regularisierung

$$F_M(x) = \begin{cases} x - \dfrac{1}{2M} & : \quad x > \dfrac{1}{M} \\ (M/2) x^2 & : \quad x \in \left[-\dfrac{1}{M}, \dfrac{1}{M} \right] \\ -x - \dfrac{1}{2M} & : \quad x < -\dfrac{1}{M}. \end{cases}$$

In den nächsten Abbildungen machen wir uns die beiden Funktionen anschaulich klar. Die erste Abbildung zeigt uns die Betragsfunktion, die bekanntlich im Nullpunkt nicht differenzierbar ist.

Abbildung 3.1: Die Funktion $f(x) = |x|$.

Dagegen machen wir uns nun die Moreau-Yosida-Regularisierung klar.

Abbildung 3.2: Die Regularisierung $F_1(x)$.

Man sieht, dass die Regularisierung im Gegensatz zu der Ausgangsfunktion überall differenzierbar ist.

Das eindeutige Minimum liegt bei beiden Funktionen in $x^* = 0$.

Kapitel 4

Das Bündel-Konzept

Mit Kapitel 2 und 3 wurden die Grundlagen über die Eigenschaften unserer Zielfunktion und über ihre Regularisierung gelegt. Wir können uns nun langsam dem Algorithmus zuwenden. Unser Ziel ist es, einen Algorithmus zu finden, mit dem man das Minimum der Zielfunktion beziehungsweise der Regularisierung erhält. Somit wäre die uns gestellte Aufgabe gelöst. Bevor wir uns damit im nächsten Kapitel genauer beschäftigen, führen wir hier zuerst die Bündel-Idee ein, um den späteren Algorithmus zu vereinfachen. Man kann sowohl die Regularisierung F_M als auch ∇F_M meist nicht exakt berechnen, da man für die Berechnung von $F_M(x)$ und $\nabla F_M(x)$ bei gegebenem $x \in \mathbb{R}^n$ wieder ein Minimierungsproblem lösen muss. Dieses enthält die Zielfunktion f, wofür wir gerade eine Alternative suchen. Die beiden Funktionen F_M und ∇F_M werden aber bekanntlich im Algorithmus eine wichtige Rolle spielen. Daher ist die Idee, die beiden Funktionen anzunähern und mit diesen Approximationen weiterzuarbeiten. Dies soll in diesem Kapitel beschrieben werden. Dazu wird zuerst die Idee erklärt und danach in einem Bündel-Algorithmus umgesetzt. Dieser wird dann genauer analysiert.

4.1 Die Bündel-Idee

Nun werden die Approximationen für F_M und ∇F_M hergeleitet. Für $x, y \in \mathbb{R}^n$ sei hierzu $d := y - x$. Man erhält für F_M

$$F_M(x) = \min_{d \in \mathbb{R}^n} \left\{ f(x+d) + \frac{1}{2} d^T M d \right\}.$$

Nun approximieren wir $f(x+d)$ durch

$$\max_{i=1,\ldots,j} \left\{ f(u^i) + z_{u^i}^T (x + d - u^i) \right\},$$

wobei $\{u^i\}$ eine Folge ist, auf die später noch genauer eingegangen werden soll und $z_{u^i} \in \partial f(u^i)$. Da f konvex ist, haben wir aus der Definition der

Subgradienten

$$f(x+d) \geq \max_{i=1,\ldots,j} \left\{ f(u^i) + z_{u^i}^T (x+d-u^i) \right\}.$$

Wir wollen nun ein einfaches, aber in dieser Arbeit noch häufig verwendetes Lemma angeben.

Lemma 4.1 *Seien $x \in \mathbb{R}^n$ und $u^i, z_{u^i} \in \mathbb{R}^n$ mit $z_{u^i} \in \partial f(u^i), i = 1, ..., j$ gegeben. Wir definieren die Funktionen*

$$\check{F}_M^j(x) := \min_{d \in \mathbb{R}^n} \left\{ \max_{i=1,\ldots,j} \left\{ f(u^i) + z_{u^i}^T (x+d-u^i) \right\} + \frac{1}{2} d^T M d \right\},$$

und

$$\hat{F}_M^j(x) := f(x + d^j(x)) + \frac{1}{2} d^j(x)^T M d^j(x),$$

wobei $d^j(x)$ die Lösung des Minimierungsproblemes ist, welches man für die Berechnung des Funktionswertes von $\check{F}_M^j(x)$ an der Stelle x zu lösen hat. Dann gilt:

- $\check{F}_M^j(x) \leq F_M(x) \leq \hat{F}_M^j(x)$ *und*
- $F_M(x) = \hat{F}_M^j(x)$ *genau dann, wenn $x + d^j(x) = p(x)$ gilt,*

wobei $p(x)$ die Lösung der Minimierungsaufgabe

$(P_{M,x})$ *Minimiere* $f_{M,x}(y) := f(y) + \frac{1}{2}\|y - x\|_M^2, \quad y \in \mathbb{R}^n$

ist.

Beweis: Die erste Ungleichung folgt aus der Konvexität von f. Da $p(x)$ die eindeutige Lösung von $(P_{M,x})$ ist, folgt auch die zweite Ungleichung. Daher gilt Gleichheit genau dann, wenn $x + d^j(x) = p(x)$ gilt.

□

Hiermit haben wir also eine untere und eine obere Schranke für F_M und eine Annäherung an $p(x)$ durch $x + d^j(x)$ gefunden. Wir definieren die Differenz durch

$$\epsilon^j(x) := \hat{F}_M^j(x) - \check{F}_M^j(x).$$

4.2 Der Bündel-Algorithmus

In Form eines Algorithmus lassen sich unsere Überlegungen wie folgt aufschreiben:

- Setze $u^1 := x$, mit einem Startwert $x \in \mathbb{R}^n$ und wähle $z_{u^1} \in \partial f(u^1)$.
- Für $j = 1, 2, \ldots$:
  - Berechne die Lösung $d^j(x)$ von

    $$\text{Min.} \quad \check{f}^j_{M,x}(x+d) := \max_{i=1,\ldots,j} \left\{ f(u^i) + z_{u^i}^T (x + d - u^i) \right\} + \frac{1}{2}\|d\|^2_M,$$

    $d \in \mathbb{R}^n$.
  - Berechne

    $$\begin{aligned} \check{F}^j_M(x) &:= \check{f}^j_{M,x}(x + d^j(x)) \\ \hat{F}^j_M(x) &:= f(x + d^j(x)) + \frac{1}{2}\|d^j(x)\|^2_M \\ \epsilon^j(x) &:= \hat{F}^j_M(x) - \check{F}^j_M(x). \end{aligned}$$

  - Setze $u^{j+1} := x + d^j(x)$, wähle $z_{u^{j+1}} \in \partial f(u^{j+1})$ und beginne den Algorithmus erneut.

Der Vorteil dieses Algorithmus ist, dass nun ein quadratisches Programm gelöst werden kann, was unseren eigentlichen Algorithmus im nächsten Kapitel vereinfacht. Eine Abbruchbedingung wollen wir erst nach den nächsten beiden Lemmata angeben. Den Beweis des nächsten Lemmas übernehmen wir in ähnlicher Form von Proposition 3 aus [4].

Lemma 4.2 *Die Bezeichnungen seien wie im Unteralgorithmus beschrieben. Dann erhalten wir:*

1. $\lim_{j\to\infty} \check{F}^j_M(x) = F_M(x)$,
2. $\lim_{j\to\infty} d^j(x) = p(x) - x$, *wobei* $p(x)$ *die eindeutige Lösung der in Satz 3.1 formulierten Minimierungsaufgabe* $(P_{M,x})$ *ist,*
3. $\lim_{j\to\infty} \hat{F}^j_M(x) = F_M(x)$.

Insgesamt gilt daher

$$\epsilon^j(x) \longrightarrow 0 \quad \text{für } j \to \infty.$$

Beweis: Im Beweis gehen wir auf die im Unteralgorithmus festgelegten Funktionen zurück. Dabei sind für $x, d, u^i, z_{u^i} \in \mathbb{R}^n$ mit $z_{u^i} \in \partial f(u^i)$

$$\check{f}^j_{M,x}(x+d) := \max_{i=1,\ldots,j} \left\{ f(u^i) + z_{u^i}^T(x+d-u^i) \right\} + \frac{1}{2}\|d\|_M^2$$

und

$$f_{M,x}(x+d) := f(x+d) + \frac{1}{2}\|d\|_M^2$$

definiert. Wir zeigen zuerst, dass $\check{f}^j_{M,x}(u^{j+1})$ für $j \to \infty$ gegen $f_{M,x}(u^*)$ konvergiert, wobei die Funktion $\check{f}^j_{M,x}$ in u^{j+1} und die Funktion $f_{M,x}$ in u^* ihr eindeutiges Minimum annimmt. Daraus würde der erste Punkt folgen.
Dazu betrachten wir die beiden Funktionen genauer. Es gilt

$$\check{f}^j_{M,x}(x+d) \overset{(1)}{\leq} \check{f}^{j+1}_{M,x}(x+d) \overset{(2)}{\leq} f_{M,x}(x+d) \quad \text{für alle } j \text{ und } d,$$

da f konvex ist, also $f(x+d) \geq \max_{i=1,\ldots,j+1} \left\{ f(u^i) + z_{u^i}^T(x+d-u^i) \right\}$ für $z_{u^i} \in \partial f(u^i)$ gilt.
Insbesondere gilt daher auch

$$\check{f}^j_{M,x}(u^{j+1}) \leq \check{f}^j_{M,x}(u^{j+2}) \overset{(3)}{\leq} \check{f}^{j+1}_{M,x}(u^{j+2}) \leq \check{f}^{j+1}_{M,x}(u^*) \overset{(4)}{\leq} f_{M,x}(u^*),$$

woraus

$$\check{f}^{j+1}_{M,x}(u^{j+2}) - \check{f}^j_{M,x}(u^{j+1}) \overset{(5)}{\longrightarrow} 0 \quad \text{für } j \to \infty$$

folgt. Aus Satz 3.1 wissen wir, dass $f_{M,x}$ gleichmäßig konvex ist. Aus dem selben Grund ist auch $\check{f}^j_{M,x}$ gleichmäßig konvex und es gilt

$$\begin{aligned} \check{f}^j_{M,x}(tu^{j+2} + (1-t)u^{j+1}) \quad &\leq \quad t\check{f}^j_{M,x}(u^{j+2}) + (1-t)\check{f}^j_{M,x}(u^{j+1}) \\ &\quad -\frac{1}{2}t(1-t)\|u^{j+2} - u^{j+1}\|_M^2 \end{aligned}$$

für alle $t \in (0,1]$. Nach Division durch t und anschließendem Grenzübergang $t \to 0+$ folgt wegen der in Kapitel 2.1 beschriebenen Beziehung zwischen Richtungsableitung und Subgradienten

$$\begin{aligned} \check{f}^j_{M,x}(u^{j+2}) \quad &\geq \quad \check{f}^j_{M,x}(u^{j+1}) + \frac{\check{f}^j_{M,x}(u^{j+1} + t(u^{j+2} - u^{j+1})) - \check{f}^j_{M,x}(u^{j+1})}{t} \\ &\quad +\frac{1}{2}(1-t)\|u^{j+2} - u^{j+1}\|_M^2 \\ &= \quad \check{f}^j_{M,x}(u^{j+1}) + \check{f}'^j_{M,x}(u^{j+1}; u^{j+2} - u^{j+1}) + \frac{1}{2}\|u^{j+2} - u^{j+1}\|_M^2 \\ &\geq \quad \check{f}^j_{M,x}(u^{j+1}) + z_{u^{j+1}}^T(u^{j+2} - u^{j+1}) + \frac{1}{2}\|u^{j+2} - u^{j+1}\|_M^2 \end{aligned}$$

für alle $z_{u^{j+1}} \in \partial \check{f}^j_{M,x}(u^{j+1})$. Da die Funktion $\check{f}^j_{M,x}$ in u^{j+1} ihr eindeutiges Minimum annimmt, gilt

$$\check{f}^j_{M,x}(u^{j+2}) \geq \check{f}^j_{M,x}(u^{j+1}) + \frac{1}{2}\|u^{j+2} - u^{j+1}\|^2_M.$$

Wegen (1) und (5) haben wir daher auch

$$\check{f}^{j+1}_{M,x}(u^{j+2}) - \check{f}^j_{M,x}(u^{j+1}) \geq \frac{1}{2}\|u^{j+2} - u^{j+1}\|^2_M \longrightarrow 0 \quad \text{für } j \to \infty,$$

und insbesondere

$$\|u^{j+1} - u^j\| \xrightarrow{(6)} 0 \quad \text{für } j \to \infty.$$

Andererseits bekommen wir für alle j und $z_{u^j} \in \partial f(u^j)$ aus der Definition

$$\check{f}^j_{M,x}(u^{j+1}) \overset{(7)}{\geq} f(u^j) + z^T_{u^j}(u^{j+1} - u^j) + \frac{1}{2}\|u^{j+1} - x\|^2_M.$$

Nach (4), (7) und der Definition von $f_{M,x}(u^j)$ haben wir für alle j

$$\begin{aligned}
f_{M,x}(u^*) &\geq \check{f}^j_{M,x}(u^{j+1}) \\
&\geq f(u^j) + z^T_{u^j}(u^{j+1} - u^j) + \frac{1}{2}\|u^{j+1} - x\|^2_M \\
&= f_{M,x}(u^j) + z^T_{u^j}(u^{j+1} - u^j) + \frac{1}{2}\|u^{j+1} - x\|^2_M - \frac{1}{2}\|u^j - x\|^2_M \\
&= f_{M,x}(u^j) + z^T_{u^j}(u^{j+1} - u^j) + (\tfrac{1}{2}u^{j+1} + \tfrac{1}{2}u^j - x)^T M(u^{j+1} - u^j) \\
&\geq f_{M,x}(u^j) - \left(\|M(\tfrac{1}{2}u^{j+1} + \tfrac{1}{2}u^j - x)\| + \|z_{u^j}\|\right)\|u^{j+1} - u^j\|.
\end{aligned}$$

Wir bekommen aus (6) und der gerade hergeleiteten Ungleichung

$$f_{M,x}(u^*) \geq \limsup_{j\to\infty} f_{M,x}(u^j).$$

Da das Minimum der Funktion $f_{M,x}$ eindeutig ist, gilt aber auch $f_{M,x}(u^*) < f_{M,x}(u^j)$ für alle j. Daraus erhalten wir die Konvergenz von $f_{M,x}(u^j)$ gegen $f_{M,x}(u^*)$, denn sowohl die Folge $\{u^j\}$ als auch die Folge $\{z_{u^j}\}$ ist beschränkt. Erstere, da die Folge $\{u^j\}$ in $L_* := \left\{y \in \mathbb{R}^n \,|\, \check{f}^1_{M,x}(y) \leq f_{M,x}(u^*)\right\}$ enthalten ist und da L_* wegen der gleichmäßigen Konvexität von $\check{f}^j_{M,x}$ kompakt ist (siehe Beweis von Satz 2.2) und letztere, da der folgende Satz gilt:

- Sei $\{u^j\}$ eine Folge und $z_{u^j} \in \partial f(u^j)$. Dann folgt aus der Beschränktheit der $\{u^j\}$ die Beschränktheit der $\{z_{u^j}\}$.

Denn für jedes $h \in \mathbb{R}^n$ ist $G(h) := \sup_{j\in\mathbb{N}} f'(u^j; h) < \infty$, da wir aus Kapitel 2.1 wissen, dass $f'(u^j; h) \leq f(u^j + h) - f(u^j)$ ist, wobei die Folge $\{u^j\}$ beschränkt und f stetig ist. Weiterhin wissen wir auch aus Kapitel 2.1, dass

$f'(u^j;h)$ für alle $u^j, h \in \mathbb{R}^n$ immer existiert und konvex ist und daher die Abbildung $G : \mathbb{R}^n \longrightarrow \mathbb{R}$ ebenfalls als das Supremum von konvexen Funktionen wieder konvex ist (siehe z.B. S.35 in [16]), also insbesondere auch stetig. Daher ist

$$\beta := \max_{\|h\|=1} G(h) < \infty.$$

Für alle j mit $z_{u^j} \neq 0$ ist

$$\|z_{u^j}\| = z_{u^j}^T \frac{z_{u^j}}{\|z_{u^j}\|} \leq \max_{z \in \partial f(u^j)} z^T \frac{z_{u^j}}{\|z_{u^j}\|} = f'\left(u^j; \frac{z_{u^j}}{\|z_{u^j}\|}\right) \leq \beta$$

und insbesondere ist daher $\{z_{u^j}\}$ beschränkt.
Aus der gleichmäßigen Konvexität folgt – wie oben im Beweis beschrieben –

$$f_{M,x}(u^j) \geq f_{M,x}(u^*) + \frac{1}{2}\|u^j - u^*\|_M^2.$$

Nun haben wir

$$\|u^j - u^*\| \xrightarrow{(8)} 0,$$

womit schon die zweite Behauptung bewiesen ist. Schließlich gilt

$$\check{f}_{M,x}^j(u^{j+1}) \longrightarrow f_{M,x}(u^*),$$

da wegen (7), (8) und der Stetigkeit von f

$$\begin{aligned} \lim_{j\to\infty} \check{f}_{M,x}^j(u^{j+1}) &\geq \lim_{j\to\infty} \left(f(u^j) + z_{u^j}^T(u^{j+1} - u^j) + \frac{1}{2}\|u^{j+1} - x\|_M^2 \right) \\ &= f(u^*) + \frac{1}{2}\|u^* - x\|_M^2 \\ &= f_{M,x}(u^*) \end{aligned}$$

gilt, aber nach (4) $\check{f}_{M,x}^j(u^{j+1}) \leq f_{M,x}(u^*)$ ist.
Wir zeigen als nächstes die dritte Behauptung.
Diese folgt aber sofort aus (8) und der Stetigkeit von f . Schließlich folgt aus der Definition von $\epsilon^j(x)$ auch die letzte Behauptung des Lemmas.

□

Dieses Lemma zeigt, dass sowohl die untere als auch die obere Schranke von F_M gegen F_M konvergiert. Im nächsten Lemma wollen wir eine geeignete Abschätzung für ∇F_M finden.

Lemma 4.3 *Für alle j und für alle $x \in \mathbb{R}^n$ gilt*

- $\|\nabla F_M(x) + M d^j(x)\|_{M^{-1}} = \|p(x) - (x + d^j(x))\|_M \leq \sqrt{2\epsilon^j(x)}$

- $\|\nabla F_M(x) + Md^j(x)\| \leq \sqrt{2\epsilon^j(x)\|M\|}$.

Beweis: Wir wissen, wie schon im letzten Beweis beschrieben, dass

$$f_{M,x}(u) \geq f_{M,x}(w)+z^T(u-w)+\frac{1}{2}\|u-w\|_M^2 \quad \text{für alle } u,w \in \mathbb{R}^n, z \in \partial f_{M,x}(w)$$

gilt und dass $0 \in \partial f_{M,x}(p(x))$. Mit $u = x + d^j(x), w = p(x)$ und $z = 0$ haben wir

$$f_{M,x}(x + d^j(x)) \geq f_{M,x}(p(x)) + \frac{1}{2}\|x + d^j(x) - p(x)\|_M^2,$$

also insbesondere

$$\hat{F}_M(x) \geq F_M(x) + \frac{1}{2}\|x + d^j(x) - p(x)\|_M^2.$$

Jetzt lässt sich der erste Teil des Lemmas beweisen durch

$$\begin{aligned}
\|\nabla F_M(x) + Md^j(x)\|_{M^{-1}} &= \|M(x - p(x)) + Md^j(x)\|_{M^{-1}} \\
&= \|M(x + d^j(x) - p(x))\|_{M^{-1}} \\
&= \|p(x) - x - d^j(x)\|_M \\
&\leq \sqrt{2\left(\hat{F}_M(x) - F_M(x)\right)} \\
&\leq \sqrt{2\left(\hat{F}_M(x) - \check{F}_M(x)\right)} \\
&= \sqrt{2\epsilon^j(x)},
\end{aligned}$$

mit Hilfe von Lemma 4.1 und der Definition von $\epsilon^j(x)$. Der zweite Teil folgt direkt aus

$$\begin{aligned}
& \|\nabla F_M(x) + Md^j(x)\|^2 \\
= \; & \|M(p(x) - x - d^j(x))\|^2 \\
= \; & \left(M^{\frac{1}{2}}(p(x) - x - d^j(x))\right)^T M \left(M^{\frac{1}{2}}(p(x) - x - d^j(x))\right) \\
\leq \; & \lambda_{max}(M)\|M^{\frac{1}{2}}(p(x) - x - d^j(x))\|^2 \\
= \; & \|M\|\|p(x) - x - d^j(x)\|_M^2
\end{aligned}$$

und dem ersten Teil, wobei $\lambda_{max}(M)$ der größte Eigenwert von M ist.

□

Da man den Gradienten $\nabla F_M = M(x - p(x))$ nicht exakt berechnen kann – wie am Anfang des Kapitels beschrieben – versucht man diesen durch $-Md^j(x)$ anzunähern. Das Lemma lässt erkennen, dass die Genauigkeit von $\epsilon^j(x)$ bestimmt wird. Die Differenz $\epsilon^j(x)$ soll klein gehalten werden, damit

eine möglichst genaue Annäherung sowohl an F_M als auch an den Gradienten ∇F_M gewährleistet werden kann. Ist die Differenz hinreichend klein, so kann der Unteralgorithmus gestoppt werden. Dazu wollen wir eine geeignete Abbruchbedingung formulieren. Im folgenden Bündel-Prozess wird ein $\delta(x) \in \mathbb{R}_{>0}$ eingeführt, welches als eine obere Schranke den Wert von $\epsilon^j(x)$ bestimmen kann. Genauer soll die Ungleichung

$$\epsilon^j(x) \leq \delta(x) \min\{d^j(x)^T M d^j(x), N\} \qquad (*)$$

erfüllt sein, wobei $N \in \mathbb{R}_{>0}$ gegeben ist und $\delta(x)$ während des gesamten Bündel-Prozesses fest bleibt. Falls $(*)$ nicht erfüllt ist, setzen wir $u^{j+1} := x + d^j(x)$, ersetzen j durch $j+1$, lösen das Minimierungsproblem in $\check{F}_M^{j+1}$ und testen $(*)$ mit den neu erhaltenen $d^{j+1}(x)$ und $\epsilon^{j+1}(x)$. Genauer erhalten wir den folgenden Unteralgorithmus beziehungsweise auch *Bündel-Algorithmus* genannt:

- Setze $u^1 := x$, mit einem Startwert $x \in \mathbb{R}^n$ und wähle $z_{u^1} \in \partial f(u^1)$.
- Für $j = 1, 2, \ldots$:
  - Berechne die Lösung $d^j(x)$ von

    $$\text{Min.} \quad \check{f}_{M,x}^j(x+d) := \max_{i=1,\ldots,j} \left\{ f(u^i) + z_{u^i}^T (x + d - u^i) \right\} + \frac{1}{2} \|d\|_M^2,$$

    $d \in \mathbb{R}^n$.
  - Berechne

    $$\begin{aligned} \check{F}_M^j(x) &:= \check{f}_{M,x}^j(x + d^j(x)) \\ \hat{F}_M^j(x) &:= f(x + d^j(x)) + \frac{1}{2} \|d^j(x)\|_M^2 \\ \epsilon^j(x) &:= \hat{F}_M^j(x) - \check{F}_M^j(x). \end{aligned}$$

  - Falls $\epsilon^j(x) \leq \delta(x) \min\{d^j(x)^T M d^j(x), N\}$ erfüllt ist, dann setze

    $$\check{F}_M(x) := \check{F}_M^j(x), \hat{F}_M(x) := \hat{F}_M^j(x), d(x) := d^j(x), \epsilon(x) := \epsilon^j(x)$$

    und beende den Algorithmus,

    sonst setze $u^{j+1} := x + d^j(x)$, wähle $z_{u^{j+1}} \in \partial f(u^{j+1})$ und beginne den Algorithmus erneut.

Das nächste Lemma zeigt, dass dieser Algorithmus nach einer endlichen Anzahl von Schritten abbricht.

Lemma 4.4 *Falls f nicht durch x minimiert wird, bricht nach einer endlichen Anzahl von Schritten der Bündel-Algorithmus mit einer Lösung $d^j(x)$ ab, so dass $(*)$ erfüllt wird.*

Beweis: Wir nehmen an, der Bündel-Algorithmus breche nicht ab. Dann konvergiert für $j \to \infty$ nach Lemma 4.2 $\epsilon^j(x)$ gegen 0. Daher konvergiert nach Lemma 4.3 auch $\|Md^j(x) + \nabla F_M(x)\|$ gegen 0. Da x keine optimale Lösung ist, gilt $\nabla F_M(x) \neq 0$, also existiert ein $\delta_0 \in \mathbb{R}_{>0}$, so dass $\|Md^j(x)\| \geq \delta_0$ für alle hinreichend großen j gilt. Da

$$\begin{aligned} d^j(x)^T M d^j(x) &= (Md^j(x))^T M^{-1} M d^j(x) \\ &\geq \lambda_{min}(M^{-1}) \|Md^j(x)\|^2 \\ &= \frac{1}{\lambda_{max}(M)} \|Md^j(x)\|^2 \\ &= \frac{1}{\|M\|} \|Md^j(x)\|^2 \end{aligned}$$

ist, wobei $\lambda_{max}(M)$ der größte Eigenwert von M und $\lambda_{min}(M^{-1})$ der kleinste Eigenwert von M^{-1} ist und da $(*)$ nicht erfüllt wird, gilt

$$\epsilon^j(x) > \delta(x) \min\left\{\frac{\delta_0^2}{\|M\|}, N\right\}$$

für alle hinreichend großen j. Da $\delta(x)$ während des gesamten Bündel-Prozesses fest bleibt ist dies aber ein Widerspruch zu $\epsilon^j(x) \longrightarrow 0$ für $j \to \infty$ und damit ist das Lemma bewiesen.

□

Dieses Lemma ist wesentlich für den Algorithmus, der im nächsten Kapitel beschrieben wird.

Kapitel 5

Der Algorithmus

In diesem Kapitel kommen wir zu dem Algorithmus, der unsere „Ersatz-Optimierungsaufgabe“

$$(\tilde{P}) \qquad \text{Minimiere} \quad F_M(x) := \min_{y\in\mathbb{R}^n} \left\{ f(y) + \tfrac{1}{2}\|y-x\|_M^2 \right\}, \quad x \in \mathbb{R}^n$$

löst und damit auch unser eigentliches Problem (P). Der Algorithmus soll zuerst nur angegeben werden; danach werden die einzelnen Schritte erklärt beziehungsweise ihre Wohldefiniertheit gezeigt. Den allgemeinen Aufbau kann man zum Beispiel auch in [20] nachlesen (ab S.195), wo man ebenfalls Informationen zu der hier in abgewandelter Form verwendeten Armijo-Schrittweite findet (ab S.166). Der in [20] beschriebene Algorithmus für differenzierbare Funktionen lässt sich auch für unseren Fall übertragen. Auch hier werden wir durch die Berechnung der Abstiegsrichtung, durch die Berechnung der Schrittweite und durch die Wahl des Matrix-Updates das Verfahren festlegen. Wie in den vorherigen Kapiteln beschrieben, wird aber die Moreau-Yosida-Regularisierung F_M durch ihre Schranken $\hat{F}_M$ und $\check{F}_M$ und der Gradient ∇F_M durch $-Md$ ersetzt. Die Genauigkeit der Abschätzungen wird durch den Bündel-Prozess festgelegt.

5.1 Der Quasi-Newton-Bündel-Algorithmus

Wir verwenden die Notationen $\epsilon_k := \epsilon(x_k), d_k := d(x_k)$ und so weiter.

Schritt 0 (Initialisierung). Seien σ, ρ und $N \in \mathbb{R}_{>0}$ mit $\sigma < 1$ und $\rho < 1$. Weiterhin sei $\{\delta_k\}$ eine Folge positiver Zahlen mit der Eigenschaft $\sum_{k=0}^{\infty} \delta_k < \infty$. Sei $x_0 \in \mathbb{R}^n$ ein Startwert und $B_0 \in \mathbb{R}^{n\times n}$ eine symmetrische, positiv definite Matrix. Setze $k := 0$ und finde d_0 und ϵ_0, wie im letzten Kapitel beschrieben mit

$$\epsilon_0 \leq \delta_0 \min\{d_0^T M d_0, N\},$$

wobei der Bündel-Prozess mit $j = 1$ und $u^1 = x_0$ gestartet werden kann.
Schritt 1 (Berechnung einer Abstiegsrichtung). Falls $\|Md_k\| = 0$ ist, dann stoppe mit der optimalen Lösung x_k. Ansonsten berechne

$$s_k := B_k^{-1} M d_k.$$

Schritt 2 (Berechnung der Schrittweite). Starte mit $m = 0$ und wähle i_k als die kleinste ganze nichtnegative Zahl m mit

$$\check{F}_M(x_k + \rho^m s_k) \leq \hat{F}_M(x_k) - \sigma \rho^m s_k^T M d_k,$$

wobei

$$\hat{F}_M(x_k + \rho^m s_k) - \check{F}_M(x_k + \rho^m s_k)$$
$$\leq \delta_{k+1} \min\{d(x_k + \rho^m s_k)^T M d(x_k + \rho^m s_k), N\}$$

erfülllt sein soll. Setze $\tau_k := \rho^{i_k}$ und $x_{k+1} := x_k + \tau_k s_k$.
Schritt 3 (Berechnung der Matrix). Berechne eine symmetrische, positiv definite Matrix $B_{k+1} \in \mathbb{R}^{n \times n}$ durch eine sogenannte Update-Formel. Setze $k := k + 1$ und gehe zu Schritt 1.

Falls eine Lösung existiert, so soll dieser Algorithmus entweder mit einer Lösung abbrechen oder eine Folge konstruieren, die gegen eine Lösung konvergiert. Um dies zu klären, betrachten wir den Aufbau genauer. Die einzelnen Schritte sollen etwas erklärt und die Durchführbarkeit des Algorithmus gezeigt werden.

Falls in Schritt 1 $\|Md_k\| = 0$ ist, dann folgt, dass $\|d_k\| = 0$ ist und aus

$$\epsilon_k \leq \delta_k \min\{d_k^T M d_k, N\} \qquad (*),$$

dass $\epsilon_k = 0$ ist. Nun erhalten wir aus Lemma 4.3, dass dann auch $\nabla F_M(x_k) = 0$ gilt, also ist x_k eine optimale Lösung und der Algorithmus bricht an dieser Stelle ab.
Bei s_k handelt es sich offensichtlich um eine Abstiegsrichtung bezüglich des angenäherten Gradienten. Denn ist $\|d_k\| \neq 0$ und B_k positiv definit, so ist $(-Md_k)^T s_k = -(Md_k)^T B_k^{-1}(Md_k) < 0$.
In Schritt 2 berechnen wir als Erstes ein $d(x_k + \rho^m s_k)$, wobei wir schon aus Lemma 4.4 wissen, falls $x_k + \rho^m s_k$ nicht f minimiert, dass wir nach einer endlichen Anzahl von Schritten ein solches $d(x_k + \rho^m s_k)$ finden, so dass $(*)$ erfüllt wird. Danach wird die obere Ungleichung überprüft. Falls diese nicht erfüllt wird, erhöhen wir m um 1 und wiederholen den Vorgang mit einem neuen $x_k + \rho^m s_k$. Diese Form zur Berechnung der Armijo-Schrittweite wird auch mit *backtracking line search* bezeichnet. Falls $x_k + \rho^m s_k$ schon eine optimale Lösung sein sollte, kann es sein, dass der Bündel-Algorithmus nicht

terminiert. Wir setzen hier aber voraus, dass diese Situation nicht eintritt. Um die Wohldefiniertheit von i_k zu zeigen, beweisen wir das nächste Lemma.

Lemma 5.1 *Falls f nicht durch x_k minimiert wird, existiert ein $\bar{\tau}_k > 0$ mit*

$$\check{F}_M(x_k + \tau s_k) \leq \hat{F}_M(x_k) - \sigma\tau s_k^T M d_k$$

für alle $\tau \in (0, \bar{\tau}_k]$, wobei

$$\hat{F}_M(x_k + \tau s_k) - \check{F}_M(x_k + \tau s_k) \leq \delta_{k+1} \min\{d(x_k + \tau s_k)^T M d(x_k + \tau s_k), N\}$$

erfüllt sein soll.

Beweis: Da f nicht durch x_k minimiert wird, existiert eine positive Zahl $\tilde{\tau}_k > 0$, so dass für alle $\tau \in (0, \tilde{\tau}_k]$ ebenso $x_k + \tau s_k$ nicht f minimiert. Daher kann man nach Lemma 4.4 ein $d(x_k + \tau s_k)$ für alle $\tau \in (0, \tilde{\tau}_k]$ finden. Nun betrachten wir zwei Fälle.
Im ersten Fall sei $\hat{F}_M(x_k) = F_M(x_k)$. Dann folgt $x_k + d(x_k) = p(x_k)$ aus Lemma 4.1 und

$$-Md(x_k) = M(x_k - x_k - d(x_k)) = M(x_k - p(x_k)) = \nabla F_M(x_k).$$

Da x_k keine Lösung ist, erhält man aus der Wahl der Abstiegsrichtung, dass

$$-s_k^T M d_k = s_k^T \nabla F_M(x_k) < 0$$

ist. Da F_M eine stetig differenzierbare Funktion und $\sigma < 1$ ist, existiert eine Zahl $\bar{\tau}_k > 0$ ($\bar{\tau}_k \leq \tilde{\tau}_k$), so dass für alle $\tau \in (0, \bar{\tau}_k]$

$$F_M(x_k + \tau s_k) \leq F_M(x_k) + \sigma\tau s_k^T \nabla F_M(x_k)$$

gilt. Mit Lemma 4.1 folgt so die Behauptung.
Im zweiten Fall sei $\hat{F}_M(x_k) > F_M(x_k)$. Aus Stetigkeitsgründen existiert ein $\bar{\tau}_k > 0$ mit

$$F_M(x_k + \tau s_k) \leq \hat{F}_M(x_k) - \sigma\tau s_k^T M d_k \quad \text{für alle } \tau \in (0, \bar{\tau}_k].$$

Wegen $\check{F}_M(x_k + \tau s_k) \leq F_M(x_k + \tau s_k)$ folgt die Behauptung.

□

In Schritt 3 lässt sich die Matrix B_{k+1} durch eine sogenannte Update-Formel berechnen. Eine wichtige Klasse hierbei ist die Broyden-Klasse, auf die aber erst in Kapitel 7 eingegangen werden soll. Daher gilt der nächste Konvergenzsatz unabhängig von dem Verfahren zur Berechnung der Matrix B_{k+1}.

Insgesamt haben wir gezeigt, dass der Algorithmus wohldefiniert ist. Falls dieser abbricht, dann bekommen wir eine optimale Lösung. Falls er nicht abbricht, so sagen die nächsten Kapitel etwas über die Konvergenz der Folge $\{x_k\}$ gegen eine Lösung x^* aus, falls eine existiert.

Kapitel 6

Ein erster Konvergenzsatz

Wir wollen einen ersten Konvergenzsatz beweisen. Da wir diesen noch unabhängig von dem Matrix-Update herleiten, handelt es sich lediglich um einen sehr schwachen Konvergenzsatz.
Um eine gute Approximation für die Moreau-Yosida-Regularisierung und für ihren Gradienten zu sichern, haben wir in Kapitel 4 ein $\delta(x)$ eingeführt. Für den Algorithmus in Kapitel 5 wurde vorausgesetzt, dass $\sum_{k=0}^{\infty} \delta_k < \infty$ ist. Also existiert eine Konstante $C > 0$ mit

$$\sum_{k=0}^{\infty} \delta_k \leq C.$$

Diese Feststellung und eine Aussage aus der Analysis benötigen wir für den Beweis der globalen Konvergenz.

6.1 Ein allgemeiner Konvergenzsatz

Lemma 6.1 *Sei $\{a_k\}$ eine beschränkte Folge in $\mathbb{R}$, ferner existiere eine Folge $\{\gamma_k\}$ nichtnegativer Zahlen mit $\sum_{k=0}^{\infty} \gamma_k < \infty$ und $a_{k+1} \leq a_k + \gamma_k$ für alle k. Dann konvergiert die Folge $\{a_k\}$.*

Beweis: Wir definieren uns eine Folge $\{b_k\}$ durch

$$b_k := a_k - \sum_{j=0}^{k-1} \gamma_j.$$

Um zu zeigen, dass diese Folge konvergiert, weisen wir Monotonie und Beschränktheit nach. Wir betrachten

$$b_{k+1} - b_k = a_{k+1} - \sum_{j=0}^{k} \gamma_j - a_k + \sum_{j=0}^{k-1} \gamma_j = a_{k+1} - a_k - \gamma_k \leq 0$$

und daher ist die Folge monoton fallend. Die Beschränktheit folgt aus

$$|b_k| = \left| a_k - \sum_{j=0}^{k-1} \gamma_j \right| \leq |a_k| + \left| \sum_{j=0}^{k-1} \gamma_j \right| \leq b,$$

wobei $b > 0$ eine Konstante ist. Insgesamt konvergiert die Folge $\{b_k\}$. Wegen $a_k = b_k + \sum_{j=0}^{k-1} \gamma_j$ und der Konvergenz von $\{\sum_{j=0}^{k-1} \gamma_j\}$ folgt die Konvergenz von $\{a_k\}$.

□

Nun folgt der Satz über globale Konvergenz.

Satz 6.2 *Wir betrachten die unrestringierte Optimierungsaufgabe (P). Wir nehmen an, die Zielfunktion f sei von unten beschränkt und es existieren zwei Konstanten $c_1, c_2 \in \mathbb{R}_{>0}$, so dass für die in Schritt 3 des Quasi-Newton-Bündel-Algorithmus berechnete Folge $\{B_k\}$ einmal $\|B_k\| \leq c_1$ und auch $\|B_k^{-1}\| \leq c_2$ für alle k gilt. Bricht das Verfahren nicht vorzeitig mit einer Lösung x^* ab, so wird f durch jeden Häufungspunkt von der aus dem Algorithmus gelieferten Folge $\{x_k\}$ minimiert.*

Beweis: Wir nehmen an, das Verfahren breche nicht ab. Wir wollen zeigen, dass der Gradient der Moreau-Yosida-Regularisierung ∇F_M bei jedem Häufungspunkt verschwindet. Dafür betrachten wir das Konvergenzverhalten der Folge $\{\nabla F_M(x_k)\}$ genauer. Wir analysieren aber zuerst das Konvergenzverhalten von der Folge $\{F_M(x_k)\}$ und schließen dann auf das des Gradienten. Da nach Voraussetzung f von unten beschränkt ist, ist ebenfalls die Folge $\{F_M(x_k)\}$ von unten beschränkt. Sie ist auch von oben beschränkt, denn mit Lemma 4.1 und dem Algorithmus erhalten wir für $k \geq 0$

$$\begin{aligned}
F_M(x_{k+1}) &\leq \hat{F}_M(x_{k+1}) \\
&\leq \check{F}_M(x_{k+1}) + N\delta_{k+1} \\
&\leq \hat{F}_M(x_k) - \sigma\rho^{i_k} s_k^T M d_k + N\delta_{k+1} \\
&= \hat{F}_M(x_k) - \sigma\rho^{i_k} (Md_k)^T B_k^{-1} M d_k + N\delta_{k+1} \\
&\leq \hat{F}_M(x_k) + N\delta_{k+1} \\
&\leq \check{F}_M(x_k) + N\delta_k + N\delta_{k+1} \\
&\leq F_M(x_k) + N(\delta_k + \delta_{k+1}) \\
&\leq F_M(x_{k-1}) + N(\delta_{k-1} + \delta_k) + N(\delta_k + \delta_{k+1}) \\
&\leq \ldots \\
&\leq F_M(x_0) + N(\delta_0 + \delta_1) + \ldots + N(\delta_k + \delta_{k+1}) \\
&\leq F_M(x_0) + 2NC.
\end{aligned}$$

Die Konvergenz der Folge $\{F_M(x_k)\}$ erhalten wir nun aus der gerade hergeleiteten Abschätzung und Lemma 6.1, da $F_M(x_{k+1}) \leq F_M(x_k) + \gamma_k$ gilt, wobei $\gamma_k := N(\delta_k + \delta_{k+1})$ eine Folge positiver Zahlen mit $\sum_{k=0}^{\infty} \gamma_k < \infty$ ist. Aus obigem Lemma haben wir dann die Konvergenz von $\{F_M(x_k)\}$ und können ein F_M^* durch $F_M^* := \lim_{k\to\infty} F_M(x_k)$ definieren. Da eine notwendige Bedingung für die Konvergenz einer Reihe ist, dass die Folgenglieder gegen null gehen, gilt $\{\delta_k\} \to 0$. Dann muss auch $\{\epsilon_k\} \to 0$ gelten und daher ist

$$\lim_{k\to\infty} \check{F}_M(x_k) = \lim_{k\to\infty} \hat{F}_M(x_k) = \lim_{k\to\infty} F_M(x_k) = F_M^*.$$

Die Berechnung der Schrittweite liefert

$$\check{F}_M(x_k + \tau_k s_k) \leq \hat{F}_M(x_k) - \sigma\tau_k s_k^T M d_k.$$

Da $\sigma\tau_k s_k^T M d_k \geq 0$ ist, muss

$$\lim_{k\to\infty} \tau_k s_k^T M d_k = 0$$

gelten. Mit $s_k := B_k^{-1} M d_k$ gilt

$$\begin{aligned}
\tau_k s_k^T M d_k &= \tau_k (M d_k)^T B_k^{-1} M d_k \\
&\geq \tau_k \lambda_{min}(B_k^{-1}) \|M d_k\|^2 \\
&= \tau_k \frac{1}{\lambda_{max}(B_k)} \|M d_k\|^2 \\
&= \tau_k \frac{1}{\|B_k\|} \|M d_k\|^2 \\
&\geq \tau_k \frac{1}{c_1} \|M d_k\|^2,
\end{aligned}$$

wobei $\lambda_{min}(B_k^{-1})$ der kleinste Eigenwert von B_k^{-1} und $\lambda_{max}(B_k)$ der größte Eigenwert von B_k ist. Dies impliziert, dass

$$\lim_{k\to\infty} \tau_k \|M d_k\|^2 = 0$$

ist. Sei x^* ein Häufungspunkt von $\{x_k\}$ und sei $\{x_k\}_{k\in K}$ eine Teilfolge, die gegen x^* konvergiert. Aus Lemma 4.3 und der Stetigkeit von ∇F_M erhalten wir

$$\lim_{k\to\infty, k\in K} -M d_k = \lim_{k\to\infty, k\in K} \nabla F_M(x_k) = \nabla F_M(x^*).$$

Jetzt werden zwei Fälle unterschieden.
Im ersten Fall sei $\liminf_{k\to\infty, k\in K} \tau_k > 0$. Dann haben wir aus den beiden letzten Grenzwerten sofort das Ergebnis

$$\nabla F_M(x^*) = 0.$$

Im zweiten Fall sei $\liminf_{k\to\infty,k\in K}\tau_k = 0$. Dieser Fall ist etwas komplizierter. Dafür nehmen wir , wenn nötig, eine Teilfolge, so dass wir $\tau_k \to 0$ für $k \in K, k \to \infty$ erhalten. Wir bekommen aus dem Schritt 2 des Algorithmus die Abschätzung bei einem Nichtabbruch

$$\check{F}_M(x_k + \rho^{i_k-1}s_k) > \hat{F}_M(x_k) - \sigma\rho^{i_k-1}s_k^T M d_k,$$

wobei $\rho^{i_k-1} = \tau_k/\rho$ ist und nach Lemma 4.1

$$F_M(x_k + \rho^{i_k-1}s_k) > F_M(x_k) - \sigma\rho^{i_k-1}s_k^T M d_k,$$

also auch

$$\frac{F_M(x_k + \rho^{i_k-1}s_k) - F_M(x_k)}{\rho^{i_k-1}} > -\sigma s_k^T M d_k. \quad (**)$$

Da $\lim_{k\to\infty,k\in K} -Md_k = \nabla F_M(x^*)$ ist, ist $\{Md_k\}_{k\in K}$ beschränkt. Zusammen mit der Voraussetzung, dass $\|B_k^{-1}\| \le c_2$ ist, muss auch $\{s_k\}_{k\in K}$ beschränkt sein, also können wir, wenn nötig, eine Teilfolge wählen, so dass

$$\lim_{k\to\infty,k\in K} s_k = s^*$$

gilt. Da $\{\rho^{i_k-1}\}_{k\in K} \to 0$ gilt, erhält man, indem man den Grenzwert auf $(**)$ anwendet

$$(s^*)^T\nabla F_M(x^*) \ge \sigma(s^*)^T\nabla F_M(x^*).$$

Andererseits erhalten wir auch

$$\begin{aligned} -s_k^T M d_k &= -s_k^T B_k s_k \\ &\le -\lambda_{min}(B_k)\|s_k\|^2 \\ &= -\frac{1}{\lambda_{max}(B_k^{-1})}\|s_k\|^2 \\ &= -\frac{1}{\|B_k^{-1}\|}\|s_k\|^2 \\ &\le -\frac{1}{c_2}\|s_k\|^2 \end{aligned}$$

und indem man den Grenzwert anwendet

$$(s^*)^T\nabla F_M(x^*) \le -\frac{1}{c_2}\|s^*\|^2,$$

was zusammen mit dem vorherigen Ergebnis

$$(s^*)^T\nabla F_M(x^*) = 0 \text{ und } s^* = 0$$

liefert. Da

$$\lim_{k\to\infty,k\in K} s_k = \lim_{k\to\infty,k\in K}(B_k^{-1}Md_k) = 0$$

ist und

$$\lim_{k\to\infty, k\in K} -Md_k = \nabla F_M(x^*),$$

ergibt dies zusammen mit den Voraussetzungen an $\{B_k^{-1}\}$

$$\nabla F_M(x^*) = 0$$

und damit ist der Konvergenzsatz bewiesen.

□

Es wurde eine Methode zum Lösen einer nichtdifferenzierbaren, konvexen Optimierungsaufgabe beschrieben, wobei die Moreau-Yosida-Regularisierung F_M und ihr Gradient ∇F_M annähernd berechnet wurden. Bis hierhin wurde lediglich ein sehr schwacher Konvergenzsatz angegeben. Dieser besagt, dass falls Häufungspunkte existieren, jeder Häufungspunkt der erzeugten Folge Lösung der Optimierungsaufgabe (P) ist. Die nächsten Kapitel beschäftigen sich mit der Frage, ob sich nicht noch ein besserer Konvergenzsatz finden lässt.

Kapitel 7

Das Update

Bei dem Quasi-Newton-Bündel-Algorithmus aus Kapitel 5 wird in Schritt 3 von der Berechnung der Matrix durch eine sogenannte Update-Formel gesprochen. Hiermit wollen wir uns in diesem Abschnitt nun genauer beschäftigen.

7.1 Das BFGS-Verfahren

Ein Quasi-Newton-Verfahren ist nicht nur durch die Wahl der Schrittweitenstrategie, sondern insbesondere auch durch die Update-Formel festgelegt. Das Verfahren zeichnet sich dadurch aus, dass anstelle der exakten Hesse-Matrix der zu minimierenden Funktion eine geeignete Approximation an diese verwendet wird. Wie man am besten die neue symmetrische und positiv definite Matrix berechnet, ist nicht eindeutig zu beantworten. Wir wollen hier auf einen Vertreter der Broyden-Klasse eingehen, nämlich auf das BFGS-Verfahren von Broyden, Fletcher, Goldfarb und Shanno (siehe auch ab S.195 in [20]). Dies lässt sich auf glatte, nicht zu hochdimensionale unrestringierte Optimierungsaufgaben anwenden, bei denen neben der Zielfunktion auch der Gradient zur Verfügung steht. Dieses Verfahren stellte sich in der numerischen Praxis als die erfolgreichste aller Quasi-Newton-Formeln heraus. Wir definieren die Vektoren Δx_k und Δy_k durch

$$\Delta x_k := x_{k+1} - x_k \quad \text{und} \quad \Delta y_k := -Md_{k+1} + Md_k.$$

Auch in diesem Kapitel nehmen wir statt des eigentlichen Gradienten der Moreau-Yosida-Regularisierung $\nabla F_M(x) = M(x - p(x))$ die Annäherung $-Md(x)$ an diesen. Ursprünglich wählte man Δy_k als die Differenz der Gradienten an den Stellen x_{k+1} und x_k. Wir verwenden auch hierbei die Annäherung. Nun lässt sich unsere neue Matrix B_{k+1} aus

$$B_{k+1} := BFGS(B_k, \Delta x_k, \Delta y_k) := B_k - \frac{(B_k \Delta x_k)(B_k \Delta x_k)^T}{\Delta x_k^T B_k \Delta x_k} + \frac{\Delta y_k \Delta y_k^T}{\Delta x_k^T \Delta y_k}$$

berechnen. Diese Formel mit $\Delta y_k := \nabla F_M(x_{k+1}) - \nabla F_M(x_k)$ entdeckten die vier Autoren praktisch zeitgleich und unabhängig voneinander, so dass diese Formel auch auf vier unterschiedlichen Wegen hergeleitet wurde. Ausgegangen ist man von der Erfüllung der *Quasi-Newton Gleichung*

$$B_{k+1}\Delta x_k = \Delta y_k$$

beziehungsweise auch *Sekantengleichung* genannt. Hier verweisen wir aber auf [20] und [6]. Dort findet man auch die folgende wichtige Vererbung:

- Falls B_k symmetrisch und positiv definit ist und $\Delta x_k^T \Delta y_k > 0$ gilt, dann ist die neue Matrix $B_{k+1} := BFGS(B_k, \Delta x_k, \Delta y_k)$ ebenfalls symmetrisch und positiv definit.

Diese Eigenschaft lässt sich von dem ursprünglichen BFGS-Update auch für unseren Fall übernehmen.
Wie genau und wann ein BFGS-Update vorgenommen wird, soll jeweils vor den nächsten beiden Konvergenzsätzen kurz beschrieben werden.

Kapitel 8

Weitere Konvergenzsätze

Wie schon in der Einleitung angekündigt, wird in diesem Kapitel nun globale und superlineare Konvergenz nachgewiesen. Um uns eine Lösung zu sichern, setzen wir für unsere Zielfunktion f gleichmäßige Konvexität voraus. Damit lässt sich Satz 2.2 anwenden und wir haben eine eindeutige Lösung für unser Minimierungsproblem. Da wir aber nun ein „Ersatzproblem“

$$(\tilde{P}) \qquad \text{Minimiere} \quad F_M(x) := \min_{y\in\mathbb{R}^n} \left\{ f(y) + \tfrac{1}{2}\|y-x\|_M^2 \right\}, \quad x \in \mathbb{R}^n$$

betrachten, wollen wir im nächsten Satz zeigen, dass aus der gleichmäßigen Konvexität der Zielfunktion auch die gleichmäßige Konvexität der Moreau-Yosida-Regularisierung folgt. Den Beweis haben wir aus [21].

Satz 8.1 *Ist f gleichmäßig konvex, so ist auch die Moreau-Yosida-Regularisierung*

$$F_M(x) := \min_{y\in\mathbb{R}^n} \left\{ f(y) + \frac{1}{2}\|y-x\|_M^2 \right\}$$

gleichmäßig konvex.

Beweis: Wir können $F_M(x)$ als

$$F_M(x) := f(p(x)) + \frac{1}{2}\|p(x)-x\|_M^2$$

schreiben, wobei $p(x)$ die eindeutige Lösung der Minimierungsaufgabe

$$\text{Minimiere} \quad f_{M,x}(y) := f(y) + \frac{1}{2}\|y-x\|_M^2, \quad y \in \mathbb{R}^n$$

ist (siehe Kapitel 3). Aus der gleichmäßigen Konvexität folgt nach Division durch t und anschließendem Grenzübergang $t \to 0+$

$$\frac{\alpha}{2}\|y-p(x)\|_M^2 + f'(p(x); y-p(x)) \le f(y) - f(p(x))$$

für alle $y \in \mathbb{R}^n$ mit einer positiven Konstanten α. Da wir aus Kapitel 2.1 wissen, dass $z^T(y-p(x)) \leq f'(p(x);y-p(x))$ für $z \in \partial f(p(x))$ gilt, bekommen wir wegen $M(x-p(x)) \in \partial f(p(x))$

$$\frac{\alpha}{2}\|y-p(x)\|_M^2 + (M(x-p(x)))^T(y-p(x)) \leq f(y) - f(p(x))$$

für alle $y \in \mathbb{R}^n$. Für beliebige $x, w \in \mathbb{R}^n$ ist nun

$$\begin{aligned}
& F_M(w) = \min_{y\in\mathbb{R}^n}\left\{f(y) + \frac{1}{2}\|y-w\|_M^2\right\} \\
\geq\ & \min_{y\in\mathbb{R}^n}\left\{f(p(x)) + \frac{\alpha}{2}\|y-p(x)\|_M^2 + (M(x-p(x)))^T(y-p(x))\right. \\
& \left. +\frac{1}{2}\|y-w\|_M^2\right\} \\
=\ & f(p(x)) + \min_{y\in\mathbb{R}^n}\left\{\frac{\alpha}{2}\|y-p(x)\|_M^2 + (M(x-p(x)))^T(y-p(x))\right. \\
& \left. +\frac{1}{2}\|y-w\|_M^2\right\} \\
=\ & F_M(x) - \frac{1}{2}\|p(x)-x\|_M^2 + \frac{\alpha}{2(1+\alpha)^2}\|w-x\|_M^2 \\
& +\frac{1}{1+\alpha}(M(x-p(x)))^T(w-x) + \frac{1}{2}\left\|p(x)-w+\frac{1}{1+\alpha}(w-x)\right\|_M^2 \\
=\ & F_M(x) - \frac{1}{2}\|p(x)-x\|_M^2 + \frac{\alpha}{2(1+\alpha)^2}\|w-x\|_M^2 \\
& +\frac{1}{1+\alpha}(M(x-p(x)))^T(w-x) + \frac{1}{2}\|p(x)-w\|_M^2 \\
& +\frac{1}{1+\alpha}(M(p(x)-w))^T(w-x) + \frac{1}{2(1+\alpha)^2}\|w-x\|_M^2 \\
=\ & F_M(x) - \frac{1}{2}\|p(x)-x\|_M^2 + \frac{1}{2}\|p(x)-w\|_M^2 - \frac{1}{2(1+\alpha)}\|w-x\|_M^2 \\
=\ & F_M(x) + \nabla F_M(x)^T(w-x) - (M(x-p(x)))^T(w-x) \\
& -\frac{1}{2}\|p(x)-x\|_M^2 + \frac{1}{2}\|p(x)-w\|_M^2 - \frac{1}{2(1+\alpha)}\|w-x\|_M^2 \\
=\ & F_M(x) + \nabla F_M(x)^T(w-x) + \frac{\alpha}{2(1+\alpha)}\|w-x\|_M^2.
\end{aligned}$$

Da M positiv definit ist, konnten wir das Minimum von $\min_{y\in\mathbb{R}^n}\{\frac{\alpha}{2}\|y-p(x)\|_M^2 + (M(x-p(x)))^T(y-p(x)) + \frac{1}{2}\|y-w\|_M^2\}$ bilden und erhalten $y = ((1+\alpha)p(x) - x + w)/(1+\alpha)$. Insgesamt ist

$$F_M(w) \geq F_M(x) + \nabla F_M(x)^T(w-x) + \frac{\alpha}{2(1+\alpha)}\|w-x\|_M^2$$

für alle $x, w \in \mathbb{R}^n$, woraus nach Lemma 2.1 die gleichmäßige Konvexität folgt.

□

Die Rückrichtung dieses Satzes gilt ebenfalls. Einen Beweis findet man zum Beispiel auch in [11].

8.1 Globale Konvergenz

In diesem Kapitel möchten wir R-lineare Konvergenz der aus dem Quasi-Newton-Bündel-Algorithmus entstandenen Folge $\{x_k\}$ gegen die Lösung x^* nachweisen. Das heißt, es existiert eine Konstante $c > 0$ und ein $q \in (0,1)$ mit

$$\|x_k - x^*\| \leq cq^k$$

für alle k.
Wir wollen zuerst ein Lemma beweisen, in dem eine Art Semi-Effizienz unserer Schrittweite gezeigt wird. Dabei heißt für eine differenzierbare Funktion F eine Schrittweite *semi-effizient*, wenn eine von x_k und der Richtung s_k unabhängige Konstante $\bar{\eta} > 0$ mit

$$F(x_k) - F(x_k + \tau_k s_k) \geq \bar{\eta} \min \left[-\nabla F(x_k)^T s_k, \left(\frac{\nabla F(x_k)^T s_k}{\|s_k\|} \right)^2 \right]$$

für alle $k \geq 0$ existiert. Es wird die Verminderung der Funktion F, verursacht durch die Armijo-Schrittweite, nach unten abgeschätzt. Diese Ungleichung stellte sich schon als fundamental bei der Konvergenzanalyse heraus (siehe z.B. ab S.168 in [20]). Deshalb wollen wir eine solche Beziehung auch für unseren Fall beweisen. Auch hier nehmen wir sowohl für F_M als auch für ∇F_M wieder ihre Approximationen.

Lemma 8.2 *Es existieren zwei positive Konstanten η_1, η_2, so dass entweder*

$$\begin{aligned} \check{F}_M(x_k + \tau_k s_k) \quad \leq \quad & \hat{F}_M(x_k) - \eta_1 \frac{(s_k^T M d_k)^2}{\|s_k\|^2} \\ & + \frac{\eta_1}{1-\sigma} \frac{s_k^T (\nabla F_M(x_k) + M d_k)(s_k^T M d_k)}{\|s_k\|^2} \end{aligned}$$

oder

$$\check{F}_M(x_k + \tau_k s_k) \leq \hat{F}_M(x_k) - \eta_2 s_k^T M d_k$$

für alle $k \geq 0$ gilt.

Beweis: Falls in Schritt 2 des Algorithmus die Ungleichung bei der Schrittweitenbestimmung

$$\check{F}_M(x_k + \rho^m s_k) \leq \hat{F}_M(x_k) - \sigma \rho^m s_k^T M d_k$$

schon für $m = 0$ erfüllt ist, haben wir für $\eta_2 := \sigma$ die zweite Behauptung. Für ein $m > 0$ wählen wir $i_k > 0$ als die kleinste Zahl, bei der die Ungleichung der Schrittweitenbestimmung erfüllt ist, das heißt für $\tau_k/\rho = \rho^{i_k - 1}$ ist sie es noch nicht. Also gilt

$$\check{F}_M(x_k + (\tau_k/\rho)s_k) > \hat{F}_M(x_k) - \sigma(\tau_k/\rho)s_k^T M d_k$$

beziehungsweise nach Lemma 4.1

$$F_M(x_k + (\tau_k/\rho)s_k) > F_M(x_k) - \sigma(\tau_k/\rho)s_k^T M d_k.$$

Dann folgt aus dem Mittelwertsatz

$$(\tau_k/\rho)s_k^T \nabla F_M(x_k + \theta(\tau_k/\rho)s_k) > -\sigma(\tau_k/\rho)s_k^T M d_k,$$

wobei $\theta \in (0,1)$ ist. Daraus folgt aus der Lipschitzstetigkeit von ∇F_M

$$\begin{aligned} & (\tau_k/\rho)(-\sigma s_k^T M d_k - s_k^T \nabla F_M(x_k)) \\ < \; & (\tau_k/\rho)s_k^T(\nabla F_M(x_k + \theta(\tau_k/\rho)s_k) - \nabla F_M(x_k)) \\ \leq \; & \|M\|(\tau_k/\rho)s_k^T(\theta(\tau_k/\rho)s_k) \\ < \; & \|M\|((\tau_k/\rho)\|s_k\|)^2, \end{aligned}$$

woraus sich die Abschätzung

$$\tau_k > -\rho \frac{s_k^T \nabla F_M(x_k) + \sigma s_k^T M d_k}{\|M\|\|s_k\|^2}$$

für τ_k ergibt. Verwendet man diese Abschätzung nun in der Ungleichung der Schrittweitenbestimmung, so ergibt dies

$$\check{F}_M(x_k + \tau_k s_k) \leq \hat{F}_M(x_k) + \frac{\rho\sigma}{\|M\|} \frac{(s_k^T \nabla F_M(x_k) + \sigma s_k^T M d_k)(s_k^T M d_k)}{\|s_k\|^2}.$$

Daraus folgt für $\eta_1 := \rho\sigma(1-\sigma)/\|M\|$ die erste Behauptung.

□

Bemerkung: Die Ähnlichkeit zu der Semi-Effizienz der Schrittweite wird klar. Entweder gilt für alle $k \geq 0$ und für von x_k und der Richtung s_k unabhängige Konstanten $\eta_1, \eta_2 > 0$

$$\begin{aligned} \hat{F}_M(x_k) \; - \; & \check{F}_M(x_k + \tau_k s_k) \geq \eta_1 \left(\frac{(M d_k)^T s_k}{\|s_k\|} \right)^2 \\ & - \frac{\eta_1}{1-\sigma} \frac{(\nabla F_M(x_k) + M d_k)^T s_k (M d_k)^T s_k}{\|s_k\|^2} \end{aligned}$$

oder

$$\hat{F}_M(x_k) - \check{F}_M(x_k + \tau_k s_k) \geq \eta_2 (M d_k)^T s_k.$$

□

Bevor wir zu der globalen Konvergenz kommen, soll nur noch ein einfaches Lemma aus der Analysis bewiesen werden.

Lemma 8.3 *Gilt für eine nichtnegative Folge* $\{\delta_k\}_{k\geq 0}$, *dass* $\sum_{k=0}^{\infty}\delta_k < \infty$ *ist, dann gilt auch*

$$\prod_{k=0}^{\infty}(1+\delta_k) < \infty.$$

Beweis: Wir wissen, dass $1+\delta_k \leq \exp(\delta_k)$ gilt. Da $\sum_{k=0}^{\infty}\delta_k < \infty$ ist, existiert eine Konstante $C > 0$ mit $\sum_{k=0}^{\infty}\delta_k \leq C$. Daher gilt

$$\prod_{k=0}^{\infty}(1+\delta_k) \leq \prod_{k=0}^{\infty}\exp(\delta_k) = \exp\left(\sum_{k=0}^{\infty}\delta_k\right) \leq \exp(C) < \infty,$$

womit die Behauptung bewiesen ist.

□

Wir verwenden in unserem Algorithmus – wie in Kapitel 7 angekündigt – für Schritt 3 einen BFGS-Unteralgorithmus. Die BFGS-Unteralgorithmen zum Nachweis der R-linearen Konvergenz und zum Nachweis der superlinearen Konvergenz weichen leicht voneinander ab. Daher nennen wir den folgenden den BFGS1-Unteralgorithmus und beschreiben ihn durch:

- Setze $B_0 := M$, wobei bekanntlich $M \in \mathbb{R}^{n\times n}$ eine symmetrische positiv definite Matrix ist.
- Für $k = 0, 1, \ldots$:
  - Falls für gegebene Konstanten $c_3 \in (0,\infty)$ und $c_4 \in (0,1)$

$$\|\Delta x_k\|_M(\sqrt{2\epsilon_k} + \sqrt{2\epsilon_{k+1}}) \leq c_3 \Delta x_k^T \Delta y_k \qquad (\star)$$

    und

$$2\|\Delta y_k\|_M(\sqrt{2\epsilon_k} + \sqrt{2\epsilon_{k+1}}) \leq c_4\|\Delta y_k\|^2 \qquad (\star\star)$$

    erfüllt sind, dann setze $B_{k+1} := BFGS(B_k, \Delta x_k, \Delta y_k)$,
    sonst setze $B_{k+1} := M$.

Dabei sind $\Delta x_k := x_{k+1} - x_k, \Delta y_k := -Md_{k+1} + Md_k$ und $\epsilon_k := \hat{F}_M(x_k) - \check{F}_M(x_k)$. Was es mit diesen beiden Ungleichungen auf sich hat, wird im nächsten Beweis klar. Auf jeden Fall ist durch die erste Ungleichung schon die positive Definitheit der BFGS-Matrizen gesichert. Denn wir haben in Kapitel 7 beschrieben, dass $\Delta x_k^T \Delta y_k > 0$ eine Bedingung für die positive Definitheit der neuen Matrix ist. Da wir mit einer symmetrischen positiv definiten Matrix

M beginnen, ist durch die Ungleichung $(\star)$ in jedem Schritt eine neue symmetrische positiv definite Matrix gegeben. Diesen BFGS1-Unteralgorithmus wollen wir in Schritt 3 für jede Iteration verwenden und nun analysieren, zu welchen Konvergenzergebnissen man damit kommen kann.
Wie schon beschrieben genügt es, gleichmäßige Konvexität der Zielfunktion vorauszusetzen und wir erhalten den Konvergenzsatz.

Satz 8.4 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Sei f gleichmäßig konvex auf einer kompakten Menge D, $\{B_k\}$ eine Folge aus dem BFGS1-Unteralgorithmus. Bricht der Quasi-Newton-Bündel-Algorithmus nicht vorzeitig mit einer Lösung ab, dann konvergiert die daraus gelieferte Folge $\{x_k\}$ gegen die eindeutige Lösung x^* R-linear.*

Beweis: Wir nehmen an, das Verfahren breche nicht ab. Für eine Lösung x^* gilt also $x_k \neq x^*$ für alle $k \geq 0$. Da F_M nach Satz 8.1 auf einer kompakten Menge D gleichmäßig konvex ist, existiert eine eindeutige Lösung. Unser Ziel ist es, $\|x_k - x^*\|$ nach oben durch cq^k mit einer Konstanten $c > 0$ und $q \in (0,1)$ abzuschätzen, um so R-lineare Konvergenz zu erhalten. Wegen der gleichmäßigen Konvexität und da $\nabla F_M(x^*) = 0$ ist, gilt für alle $k \geq 0$

$$F_M(x_k) - F_M(x^*) \geq \frac{\alpha}{2}\|x_k - x^*\|^2.$$

Daher können wir auch $F_M(x_k) - F_M(x^*)$ abschätzen und so das gewünschte Ergebnis erhalten.
Es sei eine Menge K definiert durch

$$K := \{j \in \mathbb{N} | \text{ die Bedingungen } (\star) \text{ und } (\star\star) \text{ sind für } k = j-1 \text{ erfüllt}\}.$$

Die Elemente, die nicht in K liegen, sollen mit

$$k_0 := 0, k_1, k_2, ..., k_i, ...$$

bezeichnet werden. Dies bedeutet, dass für alle $j \in K$ die neue Matrix B_{k+1} durch ein BFGS-Update erzeugt wird und für alle anderen $B_{k+1} := M$ gesetzt wird.
Wir führen den Beweis für eine endliche Menge K. Dies ist der allgemeinere Teil. Wäre K unendlich, so existiert ein $\bar{k}$, so dass für alle $k \geq \bar{k}$ immer ein BFGS-Update vorgenommen wird. Dieser Fall wird indirekt hier mitbehandelt.
Wir beginnen den Beweis so, wie viele andere Konvergenzbeweise auch beginnen, nämlich mit einer Analyse der Hilfsfunktion

$$\psi(B_k) := \text{Spur}(B_k) - \ln(\det(B_k)),$$

woraus sich wichtige Abschätzungen gewinnen lassen. Eingeführt wurde diese Technik in einem Beweis von J. Nocedal und S. J. Wright [14] und seitdem

oft verwendet. Es wird einem Beweis von R. H. Byrd und J. Nocedal (siehe [2]) gefolgt. Wir verwenden die Abkürzungen $\Delta x_k := x_{k+1} - x_k$, $\Delta y := -Md_{k+1} + Md_k$ und $\Delta\bar{y}_k := \nabla F_M(x_{k+1}) - \nabla F_M(x_k)$. Jetzt beginnen wir, die Hilfsfunktion zu analysieren. Diese ist für alle symmetrischen und positiv definiten Matrizen auch selbst positiv, da

$$\psi(B_k) = \sum_{i=1}^{n} \lambda_i^{(k)} - \ln\left(\prod_{i=1}^{n} \lambda_i^{(k)}\right) = \sum_{i=1}^{n} \underbrace{(\lambda_i^{(k)} - \ln \lambda_i^{(k)})}_{\geq 1} \geq n > 0$$

ist, wobei $0 < \lambda_n^{(k)} \leq ... \leq \lambda_1^{(k)}$ die Eigenwerte der positiv definiten Matrix B_k sind. Außerdem wissen wir, dass für BFGS-Matrizen

- $\text{Spur}(B_{k+1}) = \text{Spur}(B_k) - \dfrac{\|B_k\Delta x_k\|^2}{\Delta x_k^T B_k \Delta x_k} + \dfrac{\|\Delta y_k\|^2}{\Delta y_k^T \Delta x_k}$
- $\det(B_{k+1}) = \det(B_k)\dfrac{\Delta y_k^T \Delta x_k}{\Delta x_k^T B_k \Delta x_k}$

gilt (siehe S.197 in [20]). Damit erhalten wir

$$\begin{aligned}
& \psi(B_{k+1}) \\
= \;& \text{Spur}(B_k) - \frac{\|B_k\Delta x_k\|^2}{\Delta x_k^T B_k \Delta x_k} + \frac{\|\Delta y_k\|^2}{\Delta y_k^T \Delta x_k} - \ln\left(\det(B_k)\frac{\Delta y_k^T \Delta x_k}{\Delta x_k^T B_k \Delta x_k}\right) \\
= \;& \psi(B_k) - \frac{\|B_k\Delta x_k\|^2}{\Delta x_k^T B_k \Delta x_k} + \frac{\|\Delta y_k\|^2}{\Delta y_k^T \Delta x_k} - \ln\left(\frac{\Delta y_k^T \Delta x_k}{\Delta x_k^T B_k \Delta x_k}\frac{\Delta x_k^T \Delta x_k}{\Delta x_k^T \Delta x_k}\right) \\
= \;& \psi(B_k) + \ln\left(\frac{\Delta x_k^T B_k \Delta x_k}{\|\Delta x_k\|\|B_k\Delta x_k\|}\right)^2 + \left[1 - \frac{\|B_k\Delta x_k\|^2}{\Delta x_k^T B_k \Delta x_k} + \ln\frac{\|B_k\Delta x_k\|^2}{\Delta x_k^T B_k \Delta x_k}\right] \\
& + \left[\frac{\|\Delta y_k\|^2}{\Delta y_k^T \Delta x_k} - 1 - \ln\frac{\Delta y_k^T \Delta x_k}{\Delta x_k^T \Delta x_k}\right].
\end{aligned}$$

Wir wollen uns nun davon überzeugen, dass die Hilfsfunktion beschränkt ist. Dazu betrachten wir die einzelnen Terme genauer. Beginnen wir mit dem letzten Klammerausdruck. Dort sehen wir uns den Ausdruck $\|\Delta y_k\|^2 / \Delta y_k^T \Delta x_k$ an. Nach Definition von K gelten für jedes $k \in [k_{i-1}, k_i - 1)$ die Ungleichungen $(\star)$ und $(\star\star)$ für $i \geq 1$. Wir verwenden lediglich Lemma 4.3 und Bedingung $(\star)$ aus dem BFGS1-Unteralgorithmus und erhalten aus

$$\begin{aligned}
\Delta x_k^T \Delta y_k &= \Delta x_k^T \Delta\bar{y}_k + \Delta x_k^T(\Delta y_k - \Delta\bar{y}_k) \\
&= \Delta x_k^T \Delta\bar{y}_k + \Delta x_k^T M^{\frac{1}{2}} M^{-\frac{1}{2}}(\Delta y_k - \Delta\bar{y}_k) \\
&\geq \Delta x_k^T \Delta\bar{y}_k - \|\Delta x_k\|_M \|\Delta y_k - \Delta\bar{y}_k\|_{M^{-1}} \\
&\geq \Delta x_k^T \Delta\bar{y}_k - \|\Delta x_k\|_M (\|\nabla F_M(x_k) + Md_k\|_{M^{-1}} \\
&\quad + \|\nabla F_M(x_{k+1}) + Md_{k+1}\|_{M^{-1}}) \\
&\geq \Delta x_k^T \Delta\bar{y}_k - \|\Delta x_k\|_M (\sqrt{2\epsilon_k} + \sqrt{2\epsilon_{k+1}}) \\
&\geq \Delta x_k^T \Delta\bar{y}_k - c_3 \Delta x_k^T \Delta y_k
\end{aligned}$$

direkt eine erste Abschätzung durch

$$\begin{aligned}
\Delta x_k^T \Delta y_k &\geq \frac{1}{1+c_3} \Delta x_k^T \Delta \bar{y}_k \\
&\geq \frac{\alpha}{1+c_3} \|\Delta x_k\|^2 \\
&\geq \frac{\alpha}{(1+c_3)\|M\|^2} \|\Delta \bar{y}_k\|^2 \\
&= \frac{\alpha}{(1+c_3)\|M\|^2} \left(\|\Delta y_k\|^2 + \|\Delta \bar{y}_k - \Delta y_k\|^2 + 2\Delta y_k^T (\Delta \bar{y}_k - \Delta y_k) \right) \\
&\geq \frac{\alpha}{(1+c_3)\|M\|^2} \left(\|\Delta y_k\|^2 - 2\|\Delta y_k\|_M \|\Delta \bar{y}_k - \Delta y_k\|_{M^{-1}} \right) \\
&\geq \frac{\alpha}{(1+c_3)\|M\|^2} \left(\|\Delta y_k\|^2 - 2\|\Delta y_k\|_M (\sqrt{2\epsilon_k} + \sqrt{2\epsilon_{k+1}}) \right) \\
&\geq \frac{\alpha(1-c_4)}{(1+c_3)\|M\|^2} \|\Delta y_k\|^2 =: \alpha_1 \|\Delta y_k\|^2.
\end{aligned}$$

Hierbei wurde nur die gleichmäßige Konvexität mit einer Konstanten $\alpha > 0$ (siehe Lemma 2.1), die Lipschitzstetigkeit des Gradienten mit Lipschitzkonstanten $\|M\|$ und auch Bedingung $(\star\star)$ ausgenutzt. Sehen wir uns nun den zweiten Ausdruck $\Delta y_k^T \Delta x_k / \|\Delta x_k\|^2$ in der letzten Klammer an. Diesen haben wir gerade schon mitabgeschätzt, denn es gilt

$$\Delta y_k^T \Delta x_k \geq \frac{\alpha}{1+c_3} \|\Delta x_k\|^2 =: \alpha_2 \|\Delta x_k\|^2.$$

Insgesamt ist daher der Klammerausdruck beschränkt mit

$$\left[\frac{\|\Delta y_k\|^2}{\Delta y_k^T \Delta x_k} - 1 - \ln \frac{\Delta y_k^T \Delta x_k}{\Delta x_k^T \Delta x_k} \right] \leq \frac{1}{\alpha_1} - 1 - \ln \alpha_2 =: \hat{C}.$$

Mit $C := \psi(M) + \hat{C}$ erhalten wir daher rekursiv bis $k_{i-1} \notin K$

$$\begin{aligned}
& \psi(B_{k+1}) \\
\leq\ & \psi(B_k) + \left[\frac{1}{\alpha_1} - 1 - \ln \alpha_2 \right] + \ln \left(\frac{\Delta x_k^T B_k \Delta x_k}{\|\Delta x_k\| \|B_k \Delta x_k\|} \right)^2 \\
& + \left[1 - \frac{\|B_k \Delta x_k\|^2}{\Delta x_k^T B_k \Delta x_k} + \ln \frac{\|B_k \Delta x_k\|^2}{\Delta x_k^T B_k \Delta x_k} \right] \\
\leq\ & \sum_{j=k_{i-1}}^{k} \left\{ \ln \left(\frac{\Delta x_k^T B_k \Delta x_k}{\|\Delta x_k\| \|B_k \Delta x_k\|} \right)^2 + \left[1 - \frac{\|B_k \Delta x_k\|^2}{\Delta x_k^T B_k \Delta x_k} + \ln \frac{\|B_k \Delta x_k\|^2}{\Delta x_k^T B_k \Delta x_k} \right] \right\} \\
& + \psi(B_{k_{i-1}}) + \hat{C}(k - k_{i-1} + 1) \\
\leq\ & \sum_{j=k_{i-1}}^{k} \left\{ \ln \left(\frac{\Delta x_k^T B_k \Delta x_k}{\|\Delta x_k\| \|B_k \Delta x_k\|} \right)^2 + \left[1 - \frac{\|B_k \Delta x_k\|^2}{\Delta x_k^T B_k \Delta x_k} + \ln \frac{\|B_k \Delta x_k\|^2}{\Delta x_k^T B_k \Delta x_k} \right] \right\} \\
& + C(k - k_{i-1} + 1).
\end{aligned}$$

Wir wollen den in der Summe stehenden Ausdruck genauer untersuchen. Dazu definieren wir κ_j durch

$$\kappa_j := \underbrace{-\ln \overbrace{\left(\frac{\Delta x_j^T B_j \Delta x_j}{\|\Delta x_j\| \|B_j \Delta x_j\|} \right)^2}^{\in (0,1]}}_{\leq 0} \underbrace{- \left[1 - \frac{\|B_j \Delta x_j\|^2}{\Delta x_j^T B_j \Delta x_j} + \ln \frac{\|B_j \Delta x_j\|^2}{\Delta x_j^T B_j \Delta x_j} \right]}_{\leq 0}.$$

Daher ist $\kappa_j \geq 0$, da die Funktion $u(\xi) := 1 + \ln(\xi) - \xi \leq 0$ für alle $\xi \in \mathbb{R}_{>0}$ ist. Dies wollen wir kurz anschaulich klar machen, da wir diese Funktion auch später noch weiter verwenden werden.

Abbildung 8.1: Die Funktion $u(\xi) := 1 + \ln(\xi) - \xi$

Die Hilfsfunktion ist daher beschränkt. Durch die genauere Betrachtung von κ_j lassen sich zwei wichtige Abschätzungen für den eigentlichen Konvergenzbeweis gewinnen. Da auch $\psi(B_{k+1}) > 0$ ist, erhalten wir

$$\frac{1}{k - k_{i-1} + 1} \sum_{j=k_{i-1}}^{k} \kappa_j < C.$$

Nun definieren wir für die j mit $k_{i-1} \leq j \leq k$ eine Menge J_k, bei denen κ_j die $\lceil 1/2(k - k_{i-1} + 1) \rceil$ kleinsten Werte annimmt. Hierbei ist $\lceil \cdot \rceil$ so definiert, dass $\lceil t \rceil = i$ ist, wenn $i - 1 < t \leq i$ für $i \in \{1, 2, ..\}$. Sei κ_{max_k} der größte

Wert der κ_j für $j \in J_k$. Dann ist

$$\begin{aligned} & \frac{1}{k-k_{i-1}+1}\sum_{j=k_{i-1}}^{k}\kappa_j \geq \frac{1}{k-k_{i-1}+1}\left(\kappa_{max_k}+\sum_{j=k_{i-1},j\notin J_k}^{k}\kappa_j\right) \\ \geq & \frac{1}{k-k_{i-1}+1}\left(\kappa_{max_k}+\left(1-\frac{1}{2}\right)(k-k_{i-1}+1)\kappa_{max_k}-\kappa_{max_k}\right) \\ = & \frac{1}{2}\kappa_{max_k}. \end{aligned}$$

Also haben wir für alle $j \in J_k$

$$\begin{aligned} \kappa_j & \leq 2\left(\frac{1}{k-k_{i-1}+1}\sum_{j=k_{i-1}}^{k}\kappa_j\right) \\ & < 2C. \end{aligned}$$

Da außerdem

$$-\ln\left(\frac{\Delta x_j^T B_j \Delta x_j}{\|\Delta x_j\|\|B_j\Delta x_j\|}\right)^2 \leq \kappa_j$$

gilt, muss auch

$$-\ln\left(\frac{\Delta x_j^T B_j \Delta x_j}{\|\Delta x_j\|\|B_j\Delta x_j\|}\right)^2 < 2C$$

gelten. Daher ist

$$\frac{\Delta x_j^T B_j \Delta x_j}{\|\Delta x_j\|\|B_j\Delta x_j\|} > e^{-C} =: \beta_1$$

und somit haben wir die erste für unseren Beweis wichtige Beziehung hergeleitet. Andersherum gilt auch für alle $j \in J_k$

$$-\left[1-\frac{\|B_j\Delta x_j\|^2}{\Delta x_j^T B_j \Delta x_j}+\ln\frac{\|B_j\Delta x_j\|^2}{\Delta x_j^T B_j \Delta x_j}\right] \leq \kappa_j,$$

daher auch

$$1-\frac{\|B_j\Delta x_j\|^2}{\Delta x_j^T B_j \Delta x_j}+\ln\frac{\|B_j\Delta x_j\|^2}{\Delta x_j^T B_j \Delta x_j} > -2C.$$

Da die Funktion $u(\xi) := 1-\xi+ln(\xi)$, wie wir in Abbildung 8.1 gesehen haben, für alle $\xi > 0$ nicht positiv ist und sowohl für $t \to 0$ als auch für $t \to \infty$ gegen $-\infty$ divergiert, folgt für alle $j \in J_k$

$$0 < \frac{\|B_j\Delta x_j\|^2}{\Delta x_j^T B_j \Delta x_j} \leq \beta_2$$

mit einer Konstanten $\beta_2 > 0$, um die obere Ungleichung erfüllen zu können. Damit erhalten wir unsere zweite wichtige Beziehung

$$\frac{\|B_j \Delta x_j\|}{\|\Delta x_j\|} = \frac{\|B_j \Delta x_j\|}{\|\Delta x_j\|} \frac{\|B_j \Delta x_j\|}{\|B_j \Delta x_j\|} \frac{\Delta x_j^T B_j \Delta x_j}{\Delta x_j^T B_j \Delta x_j} \leq 1 \cdot \beta_2 = \beta_2.$$

Zusammenfassend existieren also Konstanten $\beta_1, \beta_2 > 0$, so dass für alle $k \in [k_{i-1}, k_i - 1)$, wobei $k_{i-1}, k_i \notin K$ für $i \geq 1$, die Beziehungen

$$\frac{\Delta x_j^T B_j \Delta x_j}{\|\Delta x_j\| \|B_j \Delta x_j\|} \geq \beta_1$$

und

$$\frac{\|B_j \Delta x_j\|}{\|\Delta x_j\|} \leq \beta_2$$

für mindestens $\left\lceil \frac{1}{2}(k - k_{i-1} + 1) \right\rceil$ Werte von $j \in [k_{i-1}, k]$ gelten. Da die Beziehungen auch für $j \notin K$ gelten, sollen die Konstanten so gewählt sein, dass die Beziehungen auch für $B_j := M$ gelten. Wir definieren eine weitere Menge I, die alle j enthält, bei denen die beiden Beziehungen erfüllt sind. Also sind $k_0, k_1, ..., k_i, ... \in I$. In Lemma 8.2 haben wir schon den Begriff der Semi-Effizienz bei Schrittweiten eingeführt. Ein noch besseres Ergebnis lässt sich erzielen, wenn man die Abschätzung

$$\hat{F}_M(x_j) - \check{F}_M(x_j + \tau_j s_j) \geq \bar{\eta} \|\nabla F_M(x_j)\|^2$$

mit einer Konstanten $\bar{\eta} > 0$ für alle hinreichend großen $j \in I$ herleiten kann. Dies soll jetzt geschehen. Aus Lemma 8.2 kennen wir diese Abschätzung der Differenz $\hat{F}_M(x_j) - \check{F}_M(x_j + \tau_j s_j)$ mit Hilfe von Md_k. Daher suchen wir nach einem Zusammenhang zwischen $\nabla F_M(x_k)$ und Md_k. Da die Menge D beschränkt ist, folgt aus der Lipschitzstetigkeit des Gradienten ∇F_M und der gleichmäßigen Konvexität der Regularisierung F (siehe Lemma 2.1), dass auch $\|\nabla F_M(u) - \nabla F_M(v)\|$ für alle $u, v \in \mathbb{R}^n$ beschränkt ist und somit auch die Folge $\{\|\nabla F_M(x_k)\|\}$. Außerdem gilt nach Lemma 4.3 und der ϵ_k-Abschätzung $(*)$ aus dem Bündel-Algorithmus für alle $k \geq 0$

$$\begin{aligned} \|\nabla F_M(x_k) + Md_k\| &\leq \sqrt{2\epsilon_k \|M\|} \\ &\leq \sqrt{2\delta_k \min\{d_k^T M d_k, N\} \|M\|} \\ &\leq \sqrt{2\delta_k (Md_k)^T M^{-1} M d_k \|M\|} \\ &= \sqrt{2\|M\|\delta_k} \|Md_k\|_{M^{-1}}. \end{aligned}$$

Da $\lim_{k\to\infty} \delta_k = 0$ ist, ergibt eine Grenzwertbildung

$$\lim_{k\to\infty} \frac{\|\nabla F_M(x_k) + Md_k\|}{\|Md_k\|} = 0.$$

Daher existiert ein $\bar{k}$, so dass sich $\|Md_k\|$ von oben und von unten für alle $k \geq \bar{k}$ abschätzen lässt durch

- $2\|\nabla F_M(x_k)\| \geq \|Md_k\|$,

da $2\|Md_k\| - 2\|\nabla F_M(x_k)\| \leq 2\|\nabla F_M(x_k) + Md_k\| < \|Md_k\|$ und

- $2\|Md_k\| \geq \|\nabla F_M(x_k)\|$,

da $\|\nabla F_M(x_k)\| - \|Md_k\| \leq \|\nabla F_M(x_k) + Md_k\| < \|Md_k\|$.
Nun kommen wir zu der Abschätzung

$$\hat{F}_M(x_j) - \check{F}_M(x_j + \tau_j s_j) \geq \bar{\eta}\|\nabla F_M(x_j)\|^2.$$

Dafür betrachten wir Lemma 8.2 und unterscheiden zwei Fälle.
Für $m = 0$ bekommen wir die Abschätzung für $j \in I$ aus

$$\begin{aligned}
\hat{F}_M(x_j) - \check{F}_M(x_j + \tau_j s_j) &\geq \eta_2 s_j^T M d_j \\
&= \eta_2 s_j^T B_j s_j \\
&= \eta_2 \frac{s_j^T B_j s_j}{(B_j s_j)^T (B_j s_j)} \|Md_j\|^2 \\
&= \eta_2 \frac{(\Delta x_j)^T B_j \Delta x_j}{(B_j \Delta x_j)^T (B_j \Delta x_j)} \|Md_j\|^2 \\
&\geq \eta_2 \frac{\|\Delta x_j\|}{\|B_j \Delta x_j\|} \frac{(\Delta x_j)^T B_j \Delta x_j}{\|\Delta x_j\| \|B_j \Delta x_j\|} \|Md_j\|^2 \\
&\geq \frac{\eta_2 \beta_1}{\beta_2} \|Md_j\|^2,
\end{aligned}$$

da $\Delta x_j = x_{j+1} - x_j = \tau_j s_j$ und $s_j := B_j^{-1} M d_j$ ist.
Für $m > 0$ bekommen wir die Abschätzung für alle $j \in I, j \geq \bar{k}$ und einer

beliebigen Konstanten c aus

$$\begin{aligned}
& \hat{F}_M(x_j) - \check{F}_M(x_j + \tau_j s_j) \\
\geq\ & \eta_1 \frac{(s_j^T M d_j)^2}{\|s_j\|^2} - \frac{\eta_1}{(1-\sigma)} \frac{s_j^T(\nabla F_M(x_j) + M d_j)(s_j^T M d_j)}{\|s_j\|^2} \\
\geq\ & \eta_1 \frac{(s_j^T M d_j)^2}{\|s_j\|^2} - \frac{\eta_1}{(1-\sigma)} \frac{\|s_j\|\|\nabla F_M(x_j) + M d_j\|\|s_j\|\|M d_j\|}{\|s_j\|^2} \\
=\ & \eta_1 \frac{(s_j^T M d_j)^2}{\|s_j\|^2} - \frac{\eta_1}{(1-\sigma)} \frac{\|\nabla F_M(x_j) + M d_j\|\|M d_j\|^2}{\|M d_j\|} \\
\geq\ & \eta_1 \frac{(s_j^T M d_j)^2}{\|s_j\|^2} - \frac{c\eta_1}{(1-\sigma)} \|M d_j\|^2 \\
:=\ & \eta_1 \frac{(s_j^T B_j s_j)^2}{\|s_j\|^2} - \frac{\eta_1}{2} \beta_1^2 \|M d_j\|^2 \\
=\ & \eta_1 \left(\frac{\Delta x_j^T B_j \Delta x_j}{\|\Delta x_j\|\|B_j \Delta x_j\|} \right)^2 \|M d_j\|^2 - \frac{\eta_1}{2} \beta_1^2 \|M d_j\|^2 \\
\geq\ & \frac{\eta_1}{2} \beta_1^2 \|M d_j\|^2.
\end{aligned}$$

Um eine gemeinsame Konstante η zu finden, definieren wir

$$\eta := \min \left\{ \frac{\eta_2 \beta_1}{\beta_2}, \frac{\eta_1}{2} \beta_1^2 \right\}.$$

Daher bekommen wir mit dem vorher Erwähnten für $j \in I$ und $j \geq \bar{k}$

$$\hat{F}_M(x_j) - \check{F}_M(x_j + \tau_j s_j) \geq \frac{\eta}{4} \|\nabla F_M(x_j)\|^2.$$

Wir erinnern daran, dass unser Anfangsziel war, eine Abschätzung über $F_M(x_k) - F_M(x^*)$ zu finden. Daher nutzen wir die gerade hergeleiteten Ergebnisse und betrachten diese Differenz nun genauer. Wegen der gleichmäßigen Konvexität gilt

$$\begin{aligned}
F_M(x_k) - F_M(x^*) &\leq \nabla F_M(x_k)^T (x_k - x^*) \\
&\leq \|\nabla F_M(x_k)\| \|x_k - x^*\| \\
&\leq \|\nabla F_M(x_k)\| \sqrt{\frac{2}{\alpha} \underbrace{(F_M(x_k) - F_M(x^*))}_{\geq 0}}
\end{aligned}$$

für alle $k \geq 0$, also auch

$$F_M(x_k) - F_M(x^*) \leq \frac{2}{\alpha} \|\nabla F_M(x_k)\|^2.$$

Es folgt zusammen mit dem ersten Teil für $j \in I, j \geq \bar{k}$ und Lemma 4.1

$$\begin{aligned}
& F_M(x_{j+1}) - F_M(x^*) - \epsilon_{j+1} \\
= \; & F_M(x_{j+1}) - F_M(x^*) - \hat{F}_M(x_{j+1}) + \check{F}_M(x_{j+1}) \\
\leq \; & F_M(x_{j+1}) - F_M(x^*) - \hat{F}_M(x_{j+1}) + \hat{F}_M(x_j) - \frac{\eta}{4}\|\nabla F_M(x_j)\|^2 \\
\leq \; & F_M(x_{j+1}) - F_M(x^*) - \hat{F}_M(x_{j+1}) + \hat{F}_M(x_j) \\
& -\frac{\eta\alpha}{8}(F_M(x_j) - F_M(x^*)) \\
\leq \; & \hat{F}_M(x_{j+1}) - F_M(x^*) - \hat{F}_M(x_{j+1}) + \hat{F}_M(x_j) + \check{F}_M(x_j) \\
& -\check{F}_M(x_j) - \frac{\eta\alpha}{8}(F_M(x_j) - F_M(x^*)) \\
= \; & \check{F}_M(x_j) - F_M(x^*) + \epsilon_j - \frac{\eta\alpha}{8}(F_M(x_j) - F_M(x^*)) \\
\leq \; & F_M(x_j) - F_M(x^*) + \epsilon_j - \frac{\eta\alpha}{8}(F_M(x_j) - F_M(x^*)) \\
= \; & (1 - \frac{\eta\alpha}{8})(F_M(x_j) - F_M(x^*)) + \epsilon_j.
\end{aligned}$$

An dieser Ungleichung wollen wir gleich weiterarbeiten. Zuerst wollen wir ϵ_k genauer betrachten. Dazu verwenden wir die Abbruchbedingung $(*)$ aus dem Unteralgorithmus, die Lipschitzstetigkeit von ∇F_M mit der Konstanten $\|M\|$ und die gleichmäßige Konvexität von F_M und erhalten für alle $k \geq \bar{k}$

$$\begin{aligned}
\epsilon_k \; \leq \; & \delta_k (d_k)^T M d_k \\
= \; & \delta_k (M d_k)^T M^{-1} M d_k \\
\leq \; & \delta_k \|M^{-1}\| M d_k\|^2 \\
\leq \; & 4\delta_k \|M^{-1}\| \|\nabla F_M(x_k)\|^2 \\
= \; & 4\delta_k \|M^{-1}\| \|\nabla F_M(x_k) - \nabla F_M(x^*)\|^2 \\
\leq \; & 4\delta_k \|M^{-1}\| \|M\|^2 \|x_k - x^*\|^2 \\
\leq \; & \frac{8\delta_k \|M^{-1}\| \|M\|^2}{\alpha}(F_M(x_k) - F_M(x^*)).
\end{aligned}$$

Da $\{\delta_k\}$ gegen null konvergiert, kann man $\bar{k}$ so wählen, dass für alle $k \geq \bar{k}$

$$\frac{8\delta_k \|M^{-1}\| \|M\|^2}{\alpha} \leq \min\left\{\frac{1}{2}, \frac{\eta\alpha}{16}\right\}$$

gilt. Jetzt wollen wir an der oberen Ungleichung für alle $j \in I$ und $j \geq \bar{k}$ weiterarbeiten. In der linken Seite schätzen wir das ϵ_{j+1} mit

$$\begin{aligned}
& F_M(x_{j+1}) - F_M(x^*) - \epsilon_{j+1} \\
\geq\ & F_M(x_{j+1}) - F_M(x^*) - \frac{8\delta_{j+1}\|M^{-1}\|\|M\|^2}{\alpha}(F_M(x_{j+1}) - F_M(x^*)) \\
=\ & \left(1 - \frac{8\delta_{j+1}\|M^{-1}\|\|M\|^2}{\alpha}\right)(F_M(x_{j+1}) - F_M(x^*)) \\
\geq\ & \left(1 - \frac{1}{2}\right)(F_M(x_{j+1}) - F_M(x^*)) \\
>\ & 0
\end{aligned}$$

ab. Da x^* die eindeutige Lösung ist, haben wir auch ein echt größer null. In der rechten Seite schätzen wir das ϵ_j mit

$$\begin{aligned}
& \left(1 - \frac{\eta\alpha}{8}\right)(F_M(x_j) - F_M(x^*)) + \epsilon_j \\
\leq\ & \left(1 - \frac{\eta\alpha}{8}\right)(F_M(x_j) - F_M(x^*)) + \frac{8\delta_j\|M^{-1}\|\|M\|^2}{\alpha}(F_M(x_j) - F_M(x^*)) \\
\leq\ & \left(1 - \frac{\eta\alpha}{8}\right)(F_M(x_j) - F_M(x^*)) + \frac{\eta\alpha}{16}(F_M(x_j) - F_M(x^*)) \\
=\ & \left(1 - \frac{\eta\alpha}{16}\right)(F_M(x_j) - F_M(x^*)) \\
=:\ & r^{1/\omega}(F_M(x_j) - F_M(x^*))
\end{aligned}$$

ab, für $\omega \in (0,1)$ und eine Konstante r. Diese Konstante ist positiv, da die linke Seite echt größer null und $F_M(x_j) > F_M(x^*)$ ist. Eine Ungleichung, die hieraus hervorgeht, ist

$$\left(1 - 8\frac{\delta_{j+1}\|M^{-1}\|\|M\|^2}{\alpha}\right)(F_M(x_{j+1}) - F_M(x^*))$$

$$\leq \left(1 - \frac{1}{16}\eta\alpha\right)(F_M(x_j) - F_M(x^*))$$

für alle $j \in I$ und $j \geq \bar{k}$. Da

$$0 < \underbrace{\left(1 - \frac{1}{16}\eta\alpha\right)^1}_{\in(0,1)} < \left(1 - \frac{1}{16}\eta\alpha\right)^{\omega} < \left(1 - \frac{1}{16}\eta\alpha\right)^{\omega/2}$$

mit $\omega \in (0,1)$ bekommen wir

$$r^{1/\omega} < r < r^{1/2}.$$

Diese Beziehung werden wir gleich benötigen. Andererseits gilt aber auch

$$\left(1 - 8\frac{\delta_{k+1}\|M^{-1}\|\|M\|^2}{\alpha}\right)(F_M(x_{k+1}) - F_M(x^*))$$
$$\leq \left(1 + 8\frac{\delta_k\|M^{-1}\|\|M\|^2}{\alpha}\right)(F_M(x_k) - F_M(x^*))$$

für alle $k \geq \bar{k}$. Denn wir wissen aus der Schrittweitenberechnung, dass

$$\begin{aligned}\check{F}_M(x_{k+1}) &\leq \hat{F}_M(x_k) - \underbrace{\sigma\tau_k}_{>0}\underbrace{(Md_k)^T B_k^{-1} Md_k}_{>0} \\ &< \hat{F}_M(x_k)\end{aligned}$$

wegen der Positivität von σ und τ_k und der positiven Definitheit von B_k für alle k gilt und daher gilt auch für alle $k \geq \bar{k}$

$$\begin{aligned}&\left(1 - 8\frac{\delta_{k+1}\|M^{-1}\|\|M\|^2}{\alpha}\right)(F_M(x_{k+1}) - F_M(x^*)) \\ \leq\ & F_M(x_{k+1}) - F_M(x^*) - \epsilon_{k+1} \\ =\ & F_M(x_{k+1}) - F_M(x^*) - \hat{F}_M(x_{k+1}) + \check{F}_M(x_{k+1}) \\ <\ & F_M(x_{k+1}) - F_M(x^*) - \hat{F}_M(x_{k+1}) + \hat{F}_M(x_k) + \check{F}_M(x_k) - \check{F}_M(x_k) \\ \leq\ & \hat{F}_M(x_{k+1}) - F_M(x^*) - \hat{F}_M(x_{k+1}) + \hat{F}_M(x_k) + F_M(x_k) - \check{F}_M(x_k) \\ =\ & F_M(x_k) - F_M(x^*) + \hat{F}_M(x_k) - \check{F}_M(x_k) \\ =\ & F_M(x_k) - F_M(x^*) + \epsilon_k \\ \leq\ & F_M(x_k) - F_M(x^*) + 8\frac{\delta_k\|M^{-1}\|\|M\|^2}{\alpha}(F_M(x_k) - F_M(x^*)) \\ =\ & \left(1 + 8\frac{\delta_k\|M^{-1}\|\|M\|^2}{\alpha}\right)(F_M(x_k) - F_M(x^*)).\end{aligned}$$

Für $k \geq \bar{k}$ definieren wir

$$\delta'_k := \frac{1 + 8\dfrac{\delta_k\|M^{-1}\|\|M\|^2}{\alpha}}{1 - 8\dfrac{\delta_{k+1}\|M^{-1}\|\|M\|^2}{\alpha}}$$

und daher gilt für alle $j \geq \bar{k}$

$$\frac{r^{1/\omega}}{1 - 8\dfrac{\delta_{j+1}\|M^{-1}\|\|M\|^2}{\alpha}} \leq \delta'_j r^{1/\omega}.$$

Jetzt ist es wichtig, genau auf die Indizes zu achten. Wir werden dafür einige Fallunterscheidungen vornehmen.

Für alle $k \geq \bar{k}$ existieren $k_{i-1}, k_i \notin K$ mit $k \in [k_{i-1}, k_i)$ und $i \geq 1$. Betrachten wir zuerst den Fall $k_i - k_{i-1} \leq 2$. Wir wollen hier noch genauer auf die Möglichkeiten eingehen. Es kann $k = k_{i-1}$ sein oder $k = k_i - 1$. Wir wissen, da $k_{i-1} \notin K$ aber $k_{i-1} \in I$ ist, dass

$$\begin{aligned} F_M(x_{k+1}) - F_M(x^*) &\leq \frac{r^{1/\omega}}{1 - 8\dfrac{\delta_{k+1}\|M^{-1}\|\|M\|^2}{\alpha}}(F_M(x_k) - F_M(x^*)) \\ &\leq \delta'_k r^{1/\omega}(F_M(x_k) - F_M(x^*)) \\ &< \delta'_k r(F_M(x_k) - F_M(x^*)) \\ &= \delta'_{k_{i-1}} r(F_M(x_{k_{i-1}}) - F_M(x^*)) \end{aligned}$$

für $k = k_{i-1}$ gilt. Für $k = k_i - 1 = k_{i-1} + 1$ erhalten wir

$$\begin{aligned} F_M(x_{k+1}) - F_M(x^*) &\leq \delta'_k(F_M(x_k) - F_M(x^*)) \\ &= \delta'_{k_i-1}(F_M(x_{k_i-1}) - F_M(x^*)) \\ &= \delta'_{k_{i-1}+1}(F_M(x_{k_{i-1}+1}) - F_M(x^*)) \\ &\leq \delta'_{k_{i-1}+1}\delta'_{k_{i-1}} r(F_M(x_{k_{i-1}}) - F_M(x^*)). \end{aligned}$$

Dies können wir auch zusammenfassen und erhalten somit

$$\begin{aligned} F_M(x_{k+1}) - F_M(x^*) &\leq \left(\prod_{j=k_{i-1}}^{k} \delta'_j\right) r(F_M(x_{k_{i-1}}) - F_M(x^*)) \\ &\leq \left(\prod_{j=k_{i-1}}^{k} \delta'_j\right) (r^{1/2})^{k-k_{i-1}+1}(F_M(x_{k_{i-1}}) - F_M(x^*)). \end{aligned}$$

Betrachten wir nun den Fall $k_i - k_{i-1} > 2$. Auch hier wollen wir noch einmal unterscheiden. Für den Fall $k \in [k_{i-1}, k_i - 1)$ gibt es mindestens $\left\lceil \frac{1}{2}(k - k_{i-1} + 1) \right\rceil$ Elemente in $I \cap [k_{i-1}, k]$. Dieser Durchschnitt hat also höchstens $k - k_{i-1} + 1$ Elemente. Da man bei weniger Elementen die folgende rechte Seite nur vergrößern kann, gilt für alle $k \in [k_{i-1}, k_i - 1)$

$$\begin{aligned} F_M(x_{k+1}) - F_M(x^*) &\leq \left(\prod_{j=k_{i-1}}^{k} \delta'_j\right) r^{k-k_{i-1}+1}(F_M(x_{k_{i-1}}) - F_M(x^*)) \\ &\leq \left(\prod_{j=k_{i-1}}^{k} \delta'_j\right) (r^{1/2})^{k-k_{i-1}+1}(F_M(x_{k_{i-1}}) - F_M(x^*)). \end{aligned}$$

Für den Fall $k = k_i - 1$ bekommen wir

$$\begin{aligned} F_M(x_{k+1}) - F_M(x^*) &= F_M(x_{k_i}) - F_M(x^*) \\ &\leq \delta'_{k_i-1}(F_M(x_{k_i-1}) - F_M(x^*)) \\ &\leq \left(\prod_{j=k_{i-1}}^{k_i-1} \delta'_j\right) r^{k_i-k_{i-1}-1}(F_M(x_{k_{i-1}}) - F_M(x^*)) \\ &\leq \left(\prod_{j=k_{i-1}}^{k_i-1} \delta'_j\right) (r^{1/2})^{k-k_{i-1}+1}(F_M(x_{k_{i-1}}) - F_M(x^*)), \end{aligned}$$

da für $k_i - k_{i-1} > 2$ gilt, dass $k_i - k_{i-1} - 1 > (k_i - k_{i-1})/2$. Fassen wir diese Fallunterscheidungen nun auch zusammen, so gilt für alle $k \in [k_{i-1}, k_i)$

$$F_M(x_{k+1}) - F_M(x^*) \leq \left(\prod_{j=k_{i-1}}^{k} \delta'_j\right) (r^{1/2})^{k-k_{i-1}+1}(F_M(x_{k_{i-1}}) - F_M(x^*)).$$

Ohne Beschränkung der Allgemeinheit sei $\bar{k} \notin K$. Dann haben wir insgesamt für alle $k \geq \bar{k}$

$$F_M(x_{k+1}) - F_M(x^*) \leq \left(\prod_{j=\bar{k}}^{k} \delta'_j\right) (r^{1/2})^{k-\bar{k}+1}(F_M(x_{\bar{k}}) - F_M(x^*)).$$

Da $\sum_{k=0}^{\infty} \delta_k < \infty$ und für alle $k \geq \bar{k}$ auch $0 < 8\delta_k \|M^{-1}\|\|M\|^2/\alpha \leq 1/2$ ist, gilt auch

$$\begin{aligned} \sum_{k=\bar{k}}^{\infty}(\delta'_k - 1) &= \sum_{k=\bar{k}}^{\infty} \left(\frac{1 + 8\frac{\delta_k\|M^{-1}\|\|M\|^2}{\alpha}}{1 - 8\frac{\delta_{k+1}\|M^{-1}\|\|M\|^2}{\alpha}} - \frac{1 - 8\frac{\delta_{k+1}\|M^{-1}\|\|M\|^2}{\alpha}}{1 - 8\frac{\delta_{k+1}\|M^{-1}\|\|M\|^2}{\alpha}} \right) \\ &= \sum_{k=\bar{k}}^{\infty} \frac{8\frac{\delta_k\|M^{-1}\|\|M\|^2}{\alpha} + 8\frac{\delta_{k+1}\|M^{-1}\|\|M\|^2}{\alpha}}{1 - 8\frac{\delta_{k+1}\|M^{-1}\|\|M\|^2}{\alpha}} \\ &\leq \sum_{k=\bar{k}}^{\infty} 8\frac{\|M^{-1}\|\|M\|^2}{\alpha}(\delta_k + \delta_{k+1}) \\ &< \infty \end{aligned}$$

und daher existiert nach Lemma 8.3 eine Konstante $\bar{C}$ mit

$$\prod_{k=\bar{k}}^{\infty} \delta'_k \leq \bar{C}.$$

Also gilt für alle $k \geq \bar{k}$

$$F_M(x_{k+1}) - F_M(x^*) \leq \bar{C}(r^{1/2})^{k-\bar{k}+1}(F_M(x_{\bar{k}}) - F_M(x^*)).$$

Jetzt haben wir die gewünschten Abschätzungen, so dass wir das Ergebnis angeben können. Da $r < 1$ ist, haben wir

$$\begin{aligned} \|x_k - x^*\| &\leq \left(\frac{2}{\alpha}\right)^{\frac{1}{2}} (F_M(x_k) - F_M(x^*))^{\frac{1}{2}} \\ &\leq \left(\frac{2\bar{C}(F_M(x_{\bar{k}}) - F_M(x^*))}{\alpha}\right)^{\frac{1}{2}} \left(r^{1/4}\right)^{k-\bar{k}} \\ &\leq \bar{c}q^k. \end{aligned}$$

Damit haben wir den Beweis beendet.

□

Bemerkungen: Es sollen noch zwei Sachen festgehalten werden.

- Insbesondere folgt aus dem Satz auch, dass

  $$\sum_{k=0}^{\infty} \|x_k - x^*\| < \infty$$

  gilt.

- Im Beweis kommen zwei bekannte Größen vor, die nur der Vollständigkeit wegen erwähnt werden sollen. Einmal ist der Winkel zwischen $B_k\Delta x_k$ und Δx_k definiert durch

  $$\cos\theta_k := \frac{\Delta x_k^T B_k \Delta x_k}{\|\Delta x_k\|\|B_k\Delta x_k\|}$$

  und andererseits der entsprechende Rayleigh-Quotient durch

  $$q_k := \frac{\Delta x_k^T B_k \Delta x_k}{\Delta x_k^T \Delta x_k}.$$

□

Im nächsten Unterabschnitt soll die superlineare Konvergenz bewiesen werden.

8.2 Superlineare Konvergenz

In diesem Kapitel möchten wir superlineare Konvergenz der aus dem Quasi-Newton-Bündel-Algorithmus entstandenen Folge $\{x_k\}$ gegen die Lösung x^* nachweisen. Das heißt, es gilt

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0.$$

Wie schon bei der R-linearen Konvergenz benötigen wir hier auch weitere Voraussetzungen. Zusätzlich zu der gleichmäßigen Konvexität der Zielfunktion soll hier unter anderem die Existenz sowie die Symmetrie und positive Definitheit von $\nabla^2 F_M$ an der Lösung x^* gefordert werden. Damit lässt sich die für superlineare Konvergenz bewährte Dennis-Moré-Bedingung (siehe z.B. S.207 in [20]) bezüglich unseres Falles beweisen.
Zuerst jedoch wird der BFGS1-Unteralgorithmus leicht abgeändert. Wir erhalten den BFGS2-Unteralgorithmus durch:

- Setze $B_0 := M$, wobei bekanntlich $M \in \mathbb{R}^{n\times n}$ eine symmetrische positiv definite Matrix ist.
- Für $k = 0, 1, \ldots$:
  - Falls für gegebene Konstanten $c_3 \in (0, \infty)$ und $c_4 \in (0, 1)$

    $$\|\Delta x_k\|_M(\sqrt{2\epsilon_k} + \sqrt{2\epsilon_{k+1}}) \le c_3 \Delta x_k^T \Delta y_k \qquad (\star)$$

    und

    $$2\|\Delta y_k\|_M(\sqrt{2\epsilon_k} + \sqrt{2\epsilon_{k+1}}) \le \min\{c_4, \delta_k^{1/3} + \delta_{k+1}^{1/3}\}\|\Delta y_k\|^2 \qquad (\star\star)'$$

    erfüllt sind, dann setze $B_{k+1} := BFGS(B_k, \Delta x_k, \Delta y_k)$,
    sonst setze $B_{k+1} := M$.

Dabei sind $\Delta x_k := x_{k+1} - x_k, \Delta y_k := -Md_{k+1} + Md_k$ und $\epsilon_k := \hat{F}_M(x_k) - \check{F}_M(x_k)$. Dieser Algorithmus ist eine Verallgemeinerung des BFGS1-Unteralgorithmus, bei dem nur Ungleichung $(\star\star)$ leicht erweitert wurde. Das $\delta(x)$ kennen wir aus dem Bündel-Algorithmus und hat die Genauigkeit der Approximationen mitbestimmt. Durch das nächste Lemma erhalten wir erste wichtige Abschätzungen.

Lemma 8.5 *Sei f gleichmäßig konvex auf einer kompakten Menge D und $\{B_k\}$ eine aus dem BFGS2-Unteralgorithmus gebildete Folge. Die Folgen $\{\|B_k\|\}$ und $\{\|B_k^{-1}\|\}$ sollen beschränkt sein. Bricht der Quasi-Newton-Bün-*

del-Algorithmus nicht vorzeitig mit einer Lösung ab, so liefert er eine Folge $\{x_k\}$ *mit*

$$\|x_k - x^*\| = O(\|\Delta x_k\|) \quad \textit{und} \quad \|x_{k+1} - x^*\| = O(\|\Delta x_k\|)$$

und für alle hinreichend großen k *gelten die beiden Bedingungen*

$$\|\Delta x_k\|_M(\sqrt{2\epsilon_k} + \sqrt{2\epsilon_{k+1}}) \leq c_3 \Delta x_k^T \Delta y_k \quad (\star)$$

und

$$2\|\Delta y_k\|_M(\sqrt{2\epsilon_k} + \sqrt{2\epsilon_{k+1}}) \leq \min\{c_4, \delta_k^{1/3} + \delta_{k+1}^{1/3}\}\|\Delta y_k\|^2 \quad (\star\star)'$$

aus unserem BFGS2-Update.

Beweis : Wir nehmen an, das Verfahren breche nicht ab. Sehen wir uns das Δx_k genauer an. Es ist definiert durch $\Delta x_k := x_{k+1} - x_k = \tau_k s_k = \tau_k B_k^{-1} M d_k$. Davon wissen wir, dass die Folgen $\{\|B_k\|\}$ und $\{\|B_k^{-1}\|\}$ beschränkt sind. Weiter wissen wir, dass τ_k von oben durch 1 beschränkt ist und zeigen nun, dass ein $\bar{k}$ existiert, so dass τ_k von unten für alle $k \geq \bar{k}$ durch eine positive Konstante beschränkt ist. Aus dem Beweis des Lemmas 8.2 kennen wir schon die erste Abschätzung mit $\rho, \sigma \in (0,1)$. Außerdem verwenden wir noch Lemma 4.3, die Abbruchbedingung des Bündel-Algorithmus

$$\epsilon(x_k) \leq \delta_k \min\{d_k^T M d_k, N\} \qquad (*)$$

und erhalten somit

$$\begin{aligned}
& \tau_k > -\rho \frac{s_k^T \nabla F_M(x_k) + \sigma s_k^T M d_k}{\|M\|\|s_k\|^2} \\
= \;& \frac{\rho}{\|M\|\|s_k\|^2}\Big(-(s_k^T \nabla F_M(x_k) + s_k^T M d_k) + (1-\sigma) s_k^T M d_k \Big) \\
= \;& \frac{\rho}{\|M\|}\Big(\frac{-s_k^T(\nabla F_M(x_k) + M d_k)}{\|s_k\|^2} + (1-\sigma)\frac{(Md_k)^T B_k^{-1} M d_k}{\|s_k\|^2} \Big) \\
\geq \;& \frac{\rho}{\|M\|}\Big(\frac{-\|s_k\|\|\nabla F_M(x_k) + M d_k\|}{\|s_k\|^2} + \lambda_{min}(B_k^{-1})(1-\sigma)\frac{\|Md_k\|^2}{\|s_k\|^2} \Big) \\
\geq \;& \frac{\rho}{\|M\|}\Big(\frac{-\sqrt{2\epsilon_k\|M\|}}{\|s_k\|} + \lambda_{min}(B_k^{-1})(1-\sigma)\frac{\|Md_k\|^2}{\|s_k\|^2} \Big) \\
\geq \;& \frac{\rho}{\|M\|}\Big(\frac{-\sqrt{2\delta_k d_k^T M d_k \|M\|}}{\|s_k\|} + \lambda_{min}(B_k^{-1})(1-\sigma)\frac{\|Md_k\|^2}{\|s_k\|^2} \Big) \\
= \;& \frac{\rho}{\|M\|}\Big(-\sqrt{2\|M\|}\frac{\sqrt{\delta_k}\|M^{1/2} d_k\|}{\|s_k\|} + \lambda_{min}(B_k^{-1})(1-\sigma)\frac{\|Md_k\|^2}{\|s_k\|^2} \Big) \\
\geq \;& \bar{\tau},
\end{aligned}$$

wobei $\lambda_{min}(B_k^{-1})$ der kleinste Eigenwert der Matrix B_k^{-1} ist und $\bar{\tau}$ eine positive Konstante. Bei der letzten Abschätzung haben wir berücksichtigt, dass $Md_k = B_k s_k$ ist, $\|B_k\|$ beschränkt ist und daher auch $\|B_k s_k\|/\|s_k\|$. Dies liefert die Beschränktheit des zweiten Summanden. Da δ_k gegen null konvergiert, konvergiert der erste Summand gegen null und insgesamt folgt die Beschränktheit von τ_k. Aus der Beschränktheit der Matrizen und der Beschränktheit von τ_k folgt

$$\|Md_k\| = O(\|\Delta x_k\|).$$

Nun erhalten wir aus der gleichmäßigen Konvexität mit einer Konstanten $\alpha > 0$ und der Tatsache $\nabla F_M(x^*) = 0$

$$\begin{aligned} \|Md_k\| &\geq \|\nabla F_M(x_k)\| - \|\nabla F_M(x_k) + Md_k\| \\ &= \|\nabla F_M(x_k) - \nabla F_M(x^*)\| - \|\nabla F_M(x_k) + Md_k\| \\ &\geq \alpha\|x_k - x^*\| - \|\nabla F_M(x_k) + Md_k\| \end{aligned}$$

und daher

$$\|x_k - x^*\| \leq \frac{\|Md_k\|}{\alpha}\left(1 + \frac{\|\nabla F_M(x_k) + Md_k\|}{\|Md_k\|}\right) \leq c\|\Delta x_k\|$$

mit einer Konstanten $c > 0$. Die letzte Abschätzung ist richtig, da der zweite Summand gegen null konvergiert, wie wir aus dem Beweis des Satzes 8.4 auf Seite 51 wissen und daher

$$c_5 \geq \frac{\|\nabla F_M(x_k) + Md_k\|}{\|Md_k\|} \geq \frac{\|\nabla F_M(x_k) + Md_k\|}{c_6\|\Delta x_k\|}$$

für alle hinreichend großen k und positiven Konstanten c_5, c_6 gilt. Daraus folgt die erste Gleichung. Die zweite folgt dann aus der ersten, da

$$\begin{aligned} \|x_{k+1} - x^*\| &= \|x_k + \tau_k s_k - x^*\| \\ &\leq \|x_k - x^*\| + \|\tau_k B_k^{-1} Md_k\| \\ &= O(\|\Delta x_k\|) + O(\|\Delta x_k\|) \\ &= O(\|\Delta x_k\|) \end{aligned}$$

gilt.
Wir wollen nun die beiden Update-Bedingungen zeigen. Wir beginnen mit der Bedingung ($\star$). Genauer zeigen wir, dass

$$\lim_{k\to\infty} \frac{\|\Delta x_k\|_M(\sqrt{2\epsilon_k} + \sqrt{2\epsilon_{k+1}})}{\Delta x_k^T \Delta y_k} = 0$$

ist. Den Nenner können wir nach unten durch $\alpha_2\|\Delta x_k\|^2$ wie auf Seite 48 abschätzen. Das ϵ_k schätzen wir durch $4\delta_k\|M^{-1}\|\|M\|^2\|x_k - x^*\|^2$ nach oben

wie auf Seite 54 ab. Danach wenden wir den ersten Teil dieses Lemmas an und bekommen

$$\begin{aligned}
& \frac{\|\Delta x_k\|_M(\sqrt{2\epsilon_k}+\sqrt{2\epsilon_{k+1}})}{\Delta x_k^T \Delta y_k} \\
\leq\ & \frac{\|M^{1/2}\|\|\Delta x_k\|(\sqrt{2\epsilon_k}+\sqrt{2\epsilon_{k+1}})}{\alpha_2\|\Delta x_k\|^2} \\
\leq\ & \frac{\|M^{1/2}\|}{\alpha_2}\Big(\frac{\sqrt{8\delta_k\|M^{-1}\|\|M\|^2\|x_k-x^*\|^2}}{\|\Delta x_k\|}+ \\
& \frac{\sqrt{8\delta_{k+1}\|M^{-1}\|\|M\|^2\|x_{k+1}-x^*\|^2}}{\|\Delta x_k\|}\Big) \\
\leq\ & \tilde{c}\frac{(\sqrt{\delta_k}+\sqrt{\delta_{k+1}})\|\Delta x_k\|}{\|\Delta x_k\|} \\
=\ & \tilde{c}\left(\sqrt{\delta_k}+\sqrt{\delta_{k+1}}\right)
\end{aligned}$$

mit einer positiven Konstanten $\tilde{c}$. Da δ_k gegen null konvergiert, konvergiert der Ausdruck insgesamt gegen null. So gilt für alle hinreichend großen k die Bedingung $(\star)$. Die Bedingung $(\star\star)'$ folgt in ähnlicher Weise. Wir zeigen einerseits, dass

$$\lim_{k\to\infty}\frac{2\|\Delta y_k\|_M(\sqrt{2\epsilon_k}+\sqrt{2\epsilon_{k+1}})}{\|\Delta y_k\|^2}=0$$

ist und andererseits, dass

$$\lim_{k\to\infty}\frac{2\|\Delta y_k\|_M(\sqrt{2\epsilon_k}+\sqrt{2\epsilon_{k+1}})}{(\sqrt[3]{\delta_k}+\sqrt[3]{\delta_{k+1}})\|\Delta y_k\|^2}=0$$

ist. Den Nenner schätzen wir – wie schon gehabt – durch

$$\begin{aligned}
\|\Delta y_k\| &\geq \|\nabla F_M(x_{k+1})-\nabla F_M(x_k)\|-\|\Delta y_k-(\nabla F_M(x_{k+1})-\nabla F_M(x_k))\| \\
&\geq \alpha\|\Delta x_k\|-(\|Md_k+\nabla F_M(x_k)\|+\|Md_{k+1}+\nabla F_M(x_{k+1})\|) \\
&\geq \alpha\|\Delta x_k\|-\tilde{\alpha}\|\Delta x_k\| \\
&=: \bar{\alpha}\|\Delta x_k\|
\end{aligned}$$

für alle hinreichend großen k und positiven Konstanten $\tilde{\alpha},\bar{\alpha}$ ab. Daher folgt einerseits mit einer positiven Konstanten $\bar{c}$

$$\begin{aligned}
\frac{2\|\Delta y_k\|_M(\sqrt{2\epsilon_k}+\sqrt{2\epsilon_{k+1}})}{\|\Delta y_k\|^2} &\leq \bar{c}\frac{(\sqrt{\delta_k}+\sqrt{\delta_{k+1}})\|\Delta x_k\|}{\|\Delta y_k\|} \\
&\leq \frac{\bar{c}}{\bar{\alpha}}\left(\sqrt{\delta_k}+\sqrt{\delta_{k+1}}\right)
\end{aligned}$$

und andererseits

$$\begin{aligned}\frac{2\|\Delta y_k\|_M(\sqrt{2\epsilon_k}+\sqrt{2\epsilon_{k+1}})}{(\sqrt[3]{\delta_k}+\sqrt[3]{\delta_{k+1}})\|\Delta y_k\|^2} &\leq \frac{\bar{c}}{\bar{\alpha}}\left(\frac{\sqrt{\delta_k}+\sqrt{\delta_{k+1}}}{\sqrt[3]{\delta_k}+\sqrt[3]{\delta_{k+1}}}\right)\\ &\leq \frac{\bar{c}}{\bar{\alpha}}\left(\frac{\sqrt{\delta_k}}{\sqrt[3]{\delta_k}}+\frac{\sqrt{\delta_{k+1}}}{\sqrt[3]{\delta_{k+1}}}\right).\end{aligned}$$

Da δ_k gegen null konvergiert und auch $\{\sqrt{\delta_k}/\sqrt[3]{\delta_k}\}$, folgt die Behauptung.

Somit sind alle vier Behauptungen bewiesen.

□

Bemerkungen: Wir wollen noch zwei Dinge feststellen.

- Die Bedingung $(\star\star)'$ kann – wie der Beweis des eben bewiesenen Lemmas gezeigt hat – verallgemeinert werden. In dem Ausdruck $\delta_k^{1/3}+\delta_{k+1}^{1/3}$ kann man $1/3$ durch eine beliebige Zahl $\gamma < 1/2$ ersetzen.
- Falls die Bedingungen aus Lemma 8.5 erfüllt sind, dann besitzt die Menge K unendlich viele Elemente. Also sind nur für eine endliche Anzahl von Iterationen die Bedingungen $(\star)$ und $(\star\star)'$ nicht erfüllt und die neue Matrix wird daher nur für endlich viele Iterationen als $B_{k+1} := M$ gesetzt.

□

In Lemma 8.5 wurde die Beschränktheit der Folgen $\{\|B_k\|\}$ und $\{\|B_k^{-1}\|\}$ vorausgesetzt. Dies soll im nächsten Lemma nicht mehr vorausgesetzt, sondern hergeleitet werden. Außerdem wird die schon angekündigte Dennis-Moré-Bedingung bezüglich unseres Falles bewiesen werden. Diese hat sich erstmals bei J. E. Dennis und J. J. Moré als ein sehr nützliches Kriterium für superlineare Konvergenz herausgestellt (siehe [3]) und wurde seitdem häufig für Konvergenzbeweise verwendet. Das Resultat wollen wir wieder aus [2] übernehmen. Zu der hier gestellten Voraussetzung soll im Anschluss an das Lemma eine kurze Bemerkung folgen.
Wir wollen ab jetzt voraussetzen, dass nicht nur $\sum_{k=0}^{\infty}\delta_k < \infty$ ist, wie im Quasi-Newton-Bündel-Algorithmus gefordert, sondern $\sum_{k=0}^{\infty}\delta_k^{1/3} < \infty$ ist. Die Begründung folgt im nächsten Lemma.

Lemma 8.6 *Sei f gleichmäßig konvex auf einer kompakten Menge D und $\{B_k\}$ eine aus dem BFGS2-Unteralgorithmus gebildete Folge. Außerdem existiere $\nabla^2 F_M(x^*)$ und sei symmetrisch und positiv definit. Weiter existie-*

re eine Konstante $L > 0$, so dass für die aus dem Quasi-Newton-Bündel-Algorithmus entstandenen Folgen $\{\Delta x_k\}$ und $\{\Delta \bar{y}_k\}$ für alle $k \geq 0$

$$\frac{\|\Delta \bar{y}_k - \nabla^2 F_M(x^*)\Delta x_k\|}{\|\Delta x_k\|} \leq L \max\{\|x_{k+1} - x^*\|, \|x_k - x^*\|\}$$

gilt. Falls das Verfahren nicht vorzeitig mit einer Lösung abbricht, dann genügt die Folge $\{x_k\}$

$$\lim_{k\to\infty} \frac{\|(B_k - \nabla^2 F_M(x^*))\Delta x_k\|}{\|\Delta x_k\|} = 0$$

und die beiden Folgen $\{\|B_k\|\}$ und $\{\|B_k^{-1}\|\}$ sind beschränkt. Hierbei sind $\Delta x_k := x_{k+1} - x_k$ und $\Delta \bar{y}_k := \nabla F_M(x_{k+1}) - \nabla F_M(x_k)$.

Beweis: Wir nehmen an, das Verfahren breche nicht ab. Wir arbeiten wieder mit der Hilfsfunktion, die schon in dem Beweis von Satz 8.4 eingeführt wurde. Hierzu definieren wir

$$\Delta \tilde{x}_k := \nabla^2 F_M(x^*)^{1/2}\Delta x_k, \quad \Delta \tilde{y}_k := \nabla^2 F_M(x^*)^{-1/2}\Delta y_k,$$
$$\tilde{B}_k := \nabla^2 F_M(x^*)^{-1/2} B_k \nabla^2 F_M(x^*)^{-1/2}.$$

Aus diesen Definitionen erhalten wir für unsere BFGS-Update-Matrix

$$\begin{aligned}
\tilde{B}_{k+1} &= \nabla^2 F_M(x^*)^{-1/2} B_k \nabla^2 F_M(x^*)^{-1/2} \\
&\quad - \frac{\nabla^2 F_M(x^*)^{-1/2} B_k \Delta x_k \Delta x_k^T B_k \nabla^2 F_M(x^*)^{-1/2}}{\Delta x_k^T B_k \Delta x_k} \\
&\quad + \frac{\nabla^2 F_M(x^*)^{-1/2} \Delta y_k \Delta y_k^T \nabla^2 F_M(x^*)^{-1/2}}{\Delta x_k^T \Delta y_k} \\
&= \tilde{B}_k - \frac{\tilde{B}_k \Delta \tilde{x}_k \Delta \tilde{x}_k^T \tilde{B}_k}{\Delta \tilde{x}_k^T \tilde{B}_k \Delta \tilde{x}_k} + \frac{\Delta \tilde{y}_k \Delta \tilde{y}_k^T}{\Delta \tilde{y}_k^T \Delta \tilde{x}_k}
\end{aligned}$$

und für unsere Hilfsfunktion

$$\begin{aligned}
\psi(\tilde{B}_{k+1}) &= \psi(\tilde{B}_k) + \frac{\|\Delta \tilde{y}_k\|^2}{\Delta \tilde{y}_k^T \Delta \tilde{x}_k} - 1 - \ln \frac{\Delta \tilde{y}_k^T \Delta \tilde{x}_k}{\Delta \tilde{x}_k^T \Delta \tilde{x}_k} + \ln \left(\frac{\Delta \tilde{x}_k^T \tilde{B}_k \Delta \tilde{x}_k}{\|\tilde{B}_k \Delta \tilde{x}_k\| \|\Delta \tilde{x}_k\|} \right)^2 \\
&\quad + \left[1 - \frac{\|\tilde{B}_k \Delta \tilde{x}_k\|^2}{\Delta \tilde{x}_k^T \tilde{B}_k \Delta \tilde{x}_k} + \ln \frac{\|\tilde{B}_k \Delta \tilde{x}_k\|^2}{\Delta \tilde{x}_k^T \tilde{B}_k \Delta \tilde{x}_k} \right].
\end{aligned}$$

Auch dieses Mal wollen wir die Funktion genauer analysieren. Wir betrachten nun aber den vorderen Teil. Dafür leiten wir einige Abschätzungen her. Da K unendlich ist, exitiert ein $\bar{k}$, so dass für alle $k \geq \bar{k}$ die Ungleichungen

$$\|\Delta x_k\|_M (\sqrt{2\epsilon_k} + \sqrt{2\epsilon_{k+1}}) \leq c_3 \Delta x_k^T \Delta y_k \qquad (\star)$$

und

$$2\|\Delta y_k\|_M (\sqrt{2\epsilon_k} + \sqrt{2\epsilon_{k+1}}) \leq \min\{c_4, \delta_k^{1/3} + \delta_{k+1}^{1/3}\} \|\Delta y_k\|^2 \qquad (\star\star)'$$

für gegebene Konstanten $c_3 \in (0,\infty)$ und $c_4 \in (0,1)$ erfüllt sind. Aus der Voraussetzung dieses Lemmas, Lemma 4.3, der Bedingung $(\star\star)'$, der Tatsache $\|\Delta\bar{y}_k\|^2 \geq (1-c_4)\|\Delta y_k\|^2$ (siehe Seite 48 im Beweis von Satz 8.4) und der Lipschitzstetigkeit von $\nabla F_M(x)$ mit der Konstanten $\|M\|$ erhalten wir

$$
\begin{aligned}
\frac{\|\Delta\tilde{y}_k - \Delta\tilde{x}_k\|}{\|\Delta\tilde{x}_k\|} &= \frac{\|\nabla^2 F_M(x^*)^{-1/2}\Delta y_k - \nabla^2 F_M(x^*)^{1/2}\Delta x_k\|}{\|\nabla^2 F_M(x^*)^{1/2}\Delta x_k\|} \\
&\leq \frac{\|\nabla^2 F_M(x^*)^{-1/2}\Delta y_k - \nabla^2 F_M(x^*)^{1/2}\Delta x_k\|}{\sqrt{\lambda_{min}(\nabla^2 F_M(x^*))}\|\Delta x_k\|} \\
&= \|\nabla^2 F_M(x^*)^{-1/2}\|\frac{\|\nabla^2 F_M(x^*)^{-1/2}\Delta y_k - \nabla^2 F_M(x^*)^{1/2}\Delta x_k\|}{\|\Delta x_k\|} \\
&\leq \|\nabla^2 F_M(x^*)^{-1/2}\|^2\frac{\|\Delta y_k - \nabla^2 F_M(x^*)\Delta x_k\|}{\|\Delta x_k\|} \\
&= \|\nabla^2 F_M(x^*)^{-1/2}\|^2\frac{\|\Delta y_k - \Delta\bar{y}_k + \Delta\bar{y}_k - \nabla^2 F_M(x^*)\Delta x_k\|}{\|\Delta x_k\|} \\
&\leq \|\nabla^2 F_M(x^*)^{-1/2}\|^2\left(L\max\{\|x_{k+1}-x^*\|,\|x_k-x^*\|\} + \frac{\|\Delta y_k - \Delta\bar{y}_k\|}{\|\Delta x_k\|}\right) \\
&=: \|\nabla^2 F_M(x^*)^{-1/2}\|^2\left(\epsilon_k' + \frac{\|\Delta y_k - \Delta\bar{y}_k\|}{\|\Delta x_k\|}\right) \\
&\leq \|\nabla^2 F_M(x^*)^{-1/2}\|^2\left(\epsilon_k' + \frac{\sqrt{2\epsilon_k\|M\|} + \sqrt{2\epsilon_{k+1}\|M\|}}{\|\Delta x_k\|}\right) \\
&\leq \|\nabla^2 F_M(x^*)^{-1/2}\|^2\left(\epsilon_k' + \frac{\sqrt{\|M\|}}{2}\left(\sqrt[3]{\delta_k} + \sqrt[3]{\delta_{k+1}}\right)\frac{1}{\|\Delta x_k\|}\frac{\|\Delta y_k\|^2}{\|\Delta y_k\|_M}\right) \\
&\leq \|\nabla^2 F_M(x^*)^{-1/2}\|^2\Bigg(\epsilon_k' + \\
&\quad \frac{\sqrt{\|M\|}}{2\sqrt{1-c_4}}\left(\sqrt[3]{\delta_k} + \sqrt[3]{\delta_{k+1}}\right)\frac{\|\Delta\bar{y}_k\|}{\|\Delta x_k\|}\frac{1}{\sqrt{\lambda_{min}(M)}}\Bigg) \\
&\leq \|\nabla^2 F_M(x^*)^{-1/2}\|^2\Bigg(\epsilon_k' + \\
&\quad \frac{\sqrt{\|M\|\|M^{-1}\|}}{2\sqrt{1-c_4}}\left(\sqrt[3]{\delta_k} + \sqrt[3]{\delta_{k+1}}\right)\frac{\|\nabla F_M(x_{k+1}) - \nabla F_M(x_k)\|}{\|\Delta x_k\|}\Bigg) \\
&\leq \|\nabla^2 F_M(x^*)^{-1/2}\|^2\left(\epsilon_k' + \frac{\sqrt{\|M\|^3\|M^{-1}\|}}{2\sqrt{1-c_4}}\left(\sqrt[3]{\delta_k} + \sqrt[3]{\delta_{k+1}}\right)\right) \\
&=: \bar{\epsilon}_k,
\end{aligned}
$$

wobei $\lambda_{min}(\nabla^2 F_M(x^*))$ der kleinste Eigenwert der Matrix $\nabla^2 F_M(x^*)$ und $\lambda_{min}(M)$ der kleinste Eigenwert von M ist. Da sowohl $\sum_{k=0}^{\infty} \epsilon'_k < \infty$ wegen $\sum_{k=0}^{\infty} \|x_k - x^*\| < \infty$ ist, als auch $\sum_{k=0}^{\infty} \sqrt[3]{\delta_k} < \infty$ ist, folgt insgesamt, dass auch

$$\sum_{k=0}^{\infty} \bar{\epsilon}_k < \infty$$

ist. Für alle hinreichend großen k können wir jetzt

$$\frac{\Delta\tilde{y}_k^T \Delta\tilde{x}_k}{\Delta\tilde{x}_k^T \Delta\tilde{x}_k} = 1 + \frac{(\Delta\tilde{y}_k - \Delta\tilde{x}_k)^T \Delta\tilde{x}_k}{\Delta\tilde{x}_k^T \Delta\tilde{x}_k} \geq 1 - \frac{\|\Delta\tilde{y}_k - \Delta\tilde{x}_k\|}{\|\Delta\tilde{x}_k\|} \geq 1 - \bar{\epsilon}_k$$

und

$$\begin{aligned} \frac{\|\Delta\tilde{y}_k\|^2}{\Delta\tilde{y}_k^T \Delta\tilde{x}_k} &\leq \left(\frac{\|\Delta\tilde{x}_k\| + \|\Delta\tilde{y}_k - \Delta\tilde{x}_k\|}{\|\Delta\tilde{x}_k\|} \right)^2 \frac{\|\Delta\tilde{x}_k\|^2}{\Delta\tilde{y}_k^T \Delta\tilde{x}_k} \\ &\leq \frac{(1+\bar{\epsilon}_k)^2}{1-\bar{\epsilon}_k} \leq 1 + c\bar{\epsilon}_k \end{aligned}$$

mit einer hinreichend großen Konstanten $c > 1$ herleiten. Für alle hinreichend großen k ist $\bar{\epsilon}_k < 1/2$ und daher

$$\ln \frac{\Delta\tilde{y}_k^T \Delta\tilde{x}_k}{\Delta\tilde{x}_k^T \Delta\tilde{x}_k} \geq \ln(1 - \bar{\epsilon}_k) \geq -2\bar{\epsilon}_k > -2c\bar{\epsilon}_k.$$

Daher lässt sich unsere Hilfsfunktion für alle hinreichend großen k durch

$$\begin{aligned} \psi(\tilde{B}_{k+1}) &\leq \psi(\tilde{B}_k) + 1 + c\bar{\epsilon}_k - 1 + 2c\bar{\epsilon}_k + \ln \left(\frac{\Delta\tilde{x}_k^T \tilde{B}_k \Delta\tilde{x}_k}{\|\tilde{B}_k \Delta\tilde{x}_k\| \|\Delta\tilde{x}_k\|} \right)^2 \\ &\quad + \left[1 - \frac{\|\tilde{B}_k \Delta\tilde{x}_k\|^2}{\Delta\tilde{x}_k^T \tilde{B}_k \Delta\tilde{x}_k} + \ln \frac{\|\tilde{B}_k \Delta\tilde{x}_k\|^2}{\Delta\tilde{x}_k^T \tilde{B}_k \Delta\tilde{x}_k} \right] \\ &< \psi(\tilde{B}_k) + 3c\bar{\epsilon}_k + \ln \left(\frac{\Delta\tilde{x}_k^T \tilde{B}_k \Delta\tilde{x}_k}{\|\tilde{B}_k \Delta\tilde{x}_k\| \|\Delta\tilde{x}_k\|} \right)^2 \\ &\quad + \left[1 - \frac{\|\tilde{B}_k \Delta\tilde{x}_k\|^2}{\Delta\tilde{x}_k^T \tilde{B}_k \Delta\tilde{x}_k} + \ln \frac{\|\tilde{B}_k \Delta\tilde{x}_k\|^2}{\Delta\tilde{x}_k^T \tilde{B}_k \Delta\tilde{x}_k} \right] \end{aligned}$$

abschätzen. Nach Aufsummieren gilt mit einer hinreichend großen Konstanten $\hat{c} > 0$ – welche auch für die nicht hinreichend großen k benötigt wird –

$$\psi(\tilde{B}_{k+1}) <$$

$$\hat{c} + \sum_{j=0}^{k} \left\{ \underbrace{\ln\left(\frac{\Delta\tilde{x}_j^T \tilde{B}_j \Delta\tilde{x}_j}{\|\tilde{B}_j\Delta\tilde{x}_j\|\|\Delta\tilde{x}_j\|}\right)^2}_{\leq 0} + \underbrace{\left[1 - \frac{\|\tilde{B}_j\Delta\tilde{x}_j\|^2}{\Delta\tilde{x}_j^T\tilde{B}_j\Delta\tilde{x}_j} + \ln\frac{\|\tilde{B}_j\Delta\tilde{x}_j\|^2}{\Delta\tilde{x}_j^T\tilde{B}_j\Delta\tilde{x}_j}\right]}_{\leq 0} \right\}.$$

Hieraus folgen unsere beiden Behauptungen. Denn einerseits folgt die Beschränktheit der Folgen $\{\|B_k\|\}$ und $\{\|B_k^{-1}\|\}$ aus der durch die Konstante $\hat{c}$ nach oben beschränkten Folge $\{\psi(\tilde{B}_k)\}$ und der ebenfalls im Beweis aus Satz 8.4 beschriebenen Darstellung

$$\psi(\tilde{B}_k) = \sum_{i=1}^{n} \underbrace{(\tilde{\lambda}_i^{(k)} - \ln\tilde{\lambda}_i^{(k)})}_{\geq 1} \leq \hat{c},$$

wobei $0 < \tilde{\lambda}_n^{(k)} \leq ... \leq \tilde{\lambda}_1^{(k)}$ die Eigenwerte der positiv definiten Matrix $\tilde{B}_k$ sind. Denn hieraus folgt $\hat{c} \geq \tilde{\lambda}_1^{(k)} - \ln\tilde{\lambda}_1^{(k)} \geq \ln\tilde{\lambda}_1^{(k)}$ und $\hat{c} \geq -\ln\tilde{\lambda}_n^{(k)}$, so dass

$$\|\tilde{B}_k\| = \tilde{\lambda}_1^{(k)} \leq \exp(\hat{c}), \quad \|\tilde{B}_k^{-1}\| = \frac{1}{\tilde{\lambda}_n^{(k)}} \leq \exp(\hat{c})$$

gilt. Aus der Beschränktheit von $\{\|\tilde{B}_k\|\}$ und $\{\|\tilde{B}_k^{-1}\|\}$ folgt dann auch die Beschränktheit von $\{\|B_k\|\}$ und $\{\|B_k^{-1}\|\}$ und somit die Behauptung. Andererseits folgt auch unter Berücksichtigung von $\psi(\tilde{B}_{k+1}) > 0$, dass die Folgenglieder gegen null gehen müssen. Also gilt auch

$$\ln\left(\frac{\Delta\tilde{x}_k^T\tilde{B}_k\Delta\tilde{x}_k}{\|\tilde{B}_k\Delta\tilde{x}_k\|\|\Delta\tilde{x}_k\|}\right)^2 \to 0$$

und

$$\left[1 - \frac{\|\tilde{B}_k\Delta\tilde{x}_k\|^2}{\Delta\tilde{x}_k^T\tilde{B}_k\Delta\tilde{x}_k} + \ln\frac{\|\tilde{B}_k\Delta\tilde{x}_k\|^2}{\Delta\tilde{x}_k^T\tilde{B}_k\Delta\tilde{x}_k}\right] \to 0.$$

Die letztere Funktion haben wir schon im Beweis von Satz 8.4 genauer betrachtet. Daher wissen wir auch aus Abbildung 8.1, dass

$$\lim_{k\to\infty} \frac{\|\tilde{B}_k\Delta\tilde{x}_k\|^2}{\Delta\tilde{x}_k^T\tilde{B}_k\Delta\tilde{x}_k} = 1$$

gilt und dass mit der eben hergeleiteten Tatsache

$$1 = \lim_{k\to\infty} \frac{\Delta\tilde{x}_k^T\tilde{B}_k\Delta\tilde{x}_k}{\|\tilde{B}_k\Delta\tilde{x}_k\|\|\Delta\tilde{x}_k\|} = \lim_{k\to\infty} \frac{\Delta\tilde{x}_k^T\tilde{B}_k\Delta\tilde{x}_k}{\Delta\tilde{x}_k^T\Delta\tilde{x}_k}$$

folgt. Nun erhalten wir die Dennis-Moré-Bedingung aus

$$\begin{aligned}
&\lim_{k\to\infty} \left(\frac{\|\nabla^2 F_M(x^*)^{-1/2}(B_k - \nabla^2 F_M(x^*))\Delta x_k\|^2}{\|\nabla^2 F_M(x^*)^{1/2}\Delta x_k\|^2} \right) \\
= &\lim_{k\to\infty} \left(\frac{\|(\tilde{B}_k - I)\Delta\tilde{x}_k\|^2}{\|\Delta\tilde{x}^k\|^2} \right) \\
= &\lim_{k\to\infty} \left(\frac{\|\tilde{B}_k\Delta\tilde{x}_k\|^2 - 2\Delta\tilde{x}_k^T\tilde{B}_k\Delta\tilde{x}_k + \Delta\tilde{x}_k^T\Delta\tilde{x}_k}{\|\Delta\tilde{x}_k\|^2} \right) \\
= &\lim_{k\to\infty} \left(\left(\frac{\|\tilde{B}_k\Delta\tilde{x}_k\|}{\|\Delta\tilde{x}_k\|} \right)^2 - 2\frac{\Delta\tilde{x}_k^T\tilde{B}_k\Delta\tilde{x}_k}{\Delta\tilde{x}_k^T\Delta\tilde{x}_k} + 1 \right) \\
= &\lim_{k\to\infty} \left(\left(\frac{\|\tilde{B}_k\Delta\tilde{x}_k\|\|\Delta\tilde{x}_k\|}{\Delta\tilde{x}_k^T\tilde{B}_k\Delta\tilde{x}_k} \frac{\Delta\tilde{x}_k^T\tilde{B}_k\Delta\tilde{x}_k}{\|\Delta\tilde{x}_k\|^2} \right)^2 - 2\frac{\Delta\tilde{x}_k^T\tilde{B}_k\Delta\tilde{x}_k}{\Delta\tilde{x}_k^T\Delta\tilde{x}_k} + 1 \right) \\
= &\ 1^2 - 2\cdot 1 + 1 = 0.
\end{aligned}$$

□

Bemerkung: Die Voraussetzung

$$\frac{\|\Delta\bar{y}_k - \nabla^2 F_M(x^*)\Delta x_k\|}{\|\Delta x_k\|} \leq L \max\{\|x_{k+1} - x^*\|, \|x_k - x^*\|\}$$

für alle $k \geq 0$ aus diesem Lemma lässt sich herleiten, wenn man für die lipschitzstetige Funktion $\nabla F_M : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ Differenzierbarkeit und Richtungsdifferenzierbarkeit vom Grad 2 in x^* voraussetzt. Dabei heißt Richtungsdifferenzierbarkeit vom Grad 2 in x^*, dass

$$\nabla F_M(x^* + h) - \nabla F_M(x^*) - \nabla F_M'(x^*; h) = O(\|h\|^2)$$

gilt, wobei $\nabla F_M'(x^*; h)$ die Richtungsableitung von ∇F_M in x^* mit Richtung h ist. Damit existiert nach Proposition 2.2 in [15] eine Konstante L_1 mit

$$\|\Delta\bar{y}_k - \nabla^2 F_M(x^*)\Delta x_k\| \leq L_1 \max\{\|x_{k+1} - x^*\|^2, \|x_k - x^*\|^2\}.$$

Außerdem folgt aus Lemma 8.5 mit einer Konstanten L_2

$$\max\{\|x_{k+1} - x^*\|, \|x_k - x^*\|\} \leq L_2\|\Delta x_k\|,$$

also insgesamt die Voraussetzung mit $L := L_1L_2$. Darauf soll aber hier nicht weiter eingegangen werden.

□

Jetzt sind alle Grundlagen gelegt, um superlineare Konvergenz beweisen zu können.

Satz 8.7 *Gegeben sei die unrestringierte Optimierungsaufgabe (P). Sei f auf einer kompakten Menge D gleichmäßig konvex, $\nabla^2 F_M$ existiere und sei symmetrisch und positiv definit in x^* und es existiere eine Konstante $L > 0$ mit*

$$\frac{\|\Delta\bar{y}_k - \nabla^2 F_M(x^*)\Delta x_k\|}{\|\Delta x_k\|} \leq L \max\{\|x_{k+1} - x^*\|, \|x_k - x^*\|\}$$

für alle $k \geq 0$. Der Quasi-Newton-Bündel-Algorithmus wird mit den Einschränkungen $\sigma < 1/2$ und $\sum_{k=0}^{\infty} \delta_k^{1/3} < \infty$ angewendet. Bricht das Verfahren nicht vorzeitig mit der Lösung x^* ab, dann konvergiert die aus diesem Algorithmus entstandene Folge $\{x_k\}$ gegen die eindeutige Lösung x^* superlinear. Es gilt also

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0.$$

Beweis: Wir nehmen an, das Verfahren breche nicht ab. Wir wollen zeigen, dass sich x_{k+1} als $x_k + s_k$ für alle hinreichend großen k darstellen lässt. Dies bedeutet, dass wir in der Schrittweitenbestimmung die Ungleichung für $m = 0$ erfüllt haben möchten, so dass

$$\check{F}_M(x_k + s_k) \leq \hat{F}_M(x_k) - \sigma s_k^T M d_k$$

erfüllt ist. Diese Ungleichung wollen wir im zweiten Schritt beweisen. Zuerst wollen wir die superlineare Konvergenz für den Fall $x_{k+1} = x_k + s_k$ zeigen, also

$$\lim_{k\to\infty} \frac{\|x_k + s_k - x^*\|}{\|x_k - x^*\|} = 0.$$

Hierbei ist x^* die eindeutige Lösung die existiert, da f auf einer kompakten Menge D gleichmäßig konvex ist. Da $\Delta x_k = x_{k+1} - x_k = s_k$ gilt, lässt sich das Ergebnis des Lemmas 8.6 umändern in

$$\lim_{k\to\infty} \frac{\|(B_k - \nabla^2 F_M(x^*))s_k\|}{\|s_k\|} = 0$$

beziehungsweise mit $s_k = B_k^{-1} M d_k$ in

$$\lim_{k\to\infty} \frac{\|Md_k - \nabla^2 F_M(x^*)s_k\|}{\|s_k\|} = 0.$$

Daraus und mit

$$\lim_{k\to\infty} \frac{\|\nabla F_M(x_k) + Md_k\|}{2\|\nabla F_M(x_k)\|} \leq \lim_{k\to\infty} \frac{\|\nabla F_M(x_k) + Md_k\|}{\|Md_k\|} = 0$$

– was uns aus dem Beweis von Satz 8.4 bekannt ist – ergibt sich für alle hinreichend großen k mit der Lipschitzstetigkeit des Gradienten ∇F_M

$$\begin{aligned}\nabla^2 F_M(x^*)s_k &= Md_k + o(\|s_k\|)\\ &= -\nabla F_M(x_k) + o(\|\nabla F_M(x_k)\|) + o(\|s_k\|)\\ &= O(\|x_k - x^*\|) + o(\|s_k\|).\end{aligned}$$

Da $\nabla^2 F_M(x^*)$ invertierbar ist, bedeutet dies

$$\begin{aligned}\|s_k\| &= O(\|x_k - x^*\|) + o(\|s_k\|)\\ &= O(\|x_k - x^*\|).\end{aligned}$$

Es existiert eine Annäherung von $\nabla F_M(x_k)$ durch

$$\nabla F_M(x_k) = \underbrace{\nabla F_M(x^*)}_{=0} + \nabla^2 F_M(x^*)(x_k - x^*) + o(\|x_k - x^*\|),$$

da die Differenz zwischen $\nabla F_M(x_k)$ und dem Polynom für $x_k \to x^*$ im Vergleich zu $(x_k - x^*)$ von höherer als erster Ordnung klein wird. Daher gilt insgesamt

$$\begin{aligned}&(B_k - \nabla^2 F_M(x^*))s_k\\ =\ & Md_k - \nabla^2 F_M(x^*)s_k\\ =\ & -\nabla F_M(x_k) + o(\|\nabla F_M(x_k)\|) - \nabla^2 F_M(x^*)s_k\\ =\ & -\nabla F_M(x^*) - \nabla^2 F_M(x^*)(x_k - x^*) + o(\|x_k - x^*\|)\\ & + o(\|\nabla F_M(x_k)\|) - \nabla^2 F_M(x^*)s_k\\ =\ & -\nabla^2 F_M(x^*)(x_k - x^*) - \nabla^2 F_M(x^*)s_k + o(\|x_k - x^*\|)\\ =\ & -\nabla^2 F_M(x^*)(x_k + s_k - x^*) + o(\|x_k - x^*\|).\end{aligned}$$

Aus den beiden Gleichungen erhalten wir für alle hinreichend großen k

$$\begin{aligned}\frac{\|\nabla^2 F_M(x^*)(x_k + s_k - x^*)\|}{\|x_k - x^*\|} &= o(1) + \underbrace{\frac{\|(B_k - \nabla^2 F_M(x^*))s_k\|}{\|s_k\|}}_{\to 0}\underbrace{\frac{\|s_k\|}{\|x_k - x^*\|}}_{\leq c}\\ &= o(1),\end{aligned}$$

woraus wir wieder mit der Invertierbarkeit von $\nabla^2 F_M(x^*)$ dieses Mal die Behauptung erhalten.
Wenn wir nun noch die oben beschriebene Ungleichung

$$\check{F}_M(x_k + s_k) \leq \hat{F}_M(x_k) - \sigma s_k^T M d_k$$

für alle hinreichend großen k für die Schrittweitenbestimmung zeigen könnten, wäre der Konvergenzsatz bewiesen. Wir wollen zeigen, dass

$$F_M(x_k + s_k) \leq F_M(x_k) - \sigma s_k^T M d_k$$

gilt, woraus nach Lemma 4.1 die oben beschriebene Ungleichung folgt. Wir können $F_M(x)$ durch

$$F_M(x) = F_M(x^*) + \frac{1}{2}(x - x^*)^T \nabla^2 F_M(x^*)(x - x^*) + o(\|x - x^*\|^2)$$

in derselben Weise wie oben annähern. Da wir gerade bewiesen haben, dass $\|x_k + s_k - x^*\| = o(\|x_k - x^*\|)$ gilt, gilt auch

$$\|x_k - x^*\| - \|s_k\| = o(\|x_k - x^*\|)$$

und auch

$$\|x_k - x^*\| = \|s_k\| + o(\|s_k\|).$$

Damit erhalten wir

$$\begin{aligned}
& F_M(x_k + s_k) - F_M(x_k) + \sigma s_k^T M d_k \\
= \; & -\frac{1}{2}(x_k - x^*)^T \nabla^2 F_M(x^*)(x_k - x^*) + o(\|x_k - x^*\|^2) \\
& + F_M(x_k + s_k) - F(x^*) + \sigma s_k^T B_k s_k \\
= \; & -\frac{1}{2}(x_k - x^*)^T \nabla^2 F_M(x^*)(x_k - x^*) + o(\|s_k\|^2) \\
& + \underbrace{\frac{1}{2}(x_k + s_k - x^*)^T \nabla^2 F_M(x^*)(x_k + s_k - x^*)}_{=o(\|x_k - x^*\|^2)} + o(\|x_k + s_k - x^*\|^2) \\
& + \sigma s_k^T B_k s_k \\
= \; & -\frac{1}{2} s_k^T \nabla^2 F_M(x^*) s_k + o(\|s_k\|^2) + \sigma s_k^T B_k s_k \\
= \; & -\sigma s_k^T (\nabla^2 F_M(x^*) - B_k) s_k + \left(\sigma - \frac{1}{2}\right) s_k^T \nabla^2 F_M(x^*) s_k + o(\|s_k\|^2) \\
\leq \; & \underbrace{\sigma \|s_k\| \|(\nabla^2 F_M(x^*) - B_k) s_k\|}_{=o(\|s_k\|^2)} + \left(\sigma - \frac{1}{2}\right) s_k^T \nabla^2 F_M(x^*) s_k + o(\|s_k\|^2) \\
= \; & \left(\sigma - \frac{1}{2}\right) s_k^T \nabla^2 F_M(x^*) s_k + o(\|s_k\|^2).
\end{aligned}$$

Da $\sigma < 1/2$ und $\nabla^2 F_M(x^*)$ positiv definit ist, gilt für alle hinreichend großen k

$$F_M(x_k + s_k) - F_M(x_k) + \sigma s_k^T M d_k < 0$$

und daher die Behauptung. Insgesamt haben wir nun auch den Konvergenzsatz bewiesen.

□

Wir konnten also in diesem Kapitel globale und superlineare Konvergenz nachweisen. Das Ergebnis wollen wir im nächsten Kapitel an zwei Beispielen testen.

Kapitel 9

Numerische Ergebnisse

Um die Konvergenzgeschwindigkeit zu testen, wollen wir den von uns in MATLAB (Version 6.1) programmierten Algorithmus anhand von zwei Beispielen testen. Die Programmierung und Auswertung fand auf den Rechnern im Institut für Numerische und Angewandte Mathematik der Universität Göttingen statt.
Die Arbeit soll mit einer kurzen Zusammenfassung enden.

9.1 Auswertung

Wir wollen den Quasi-Newton-Bündel-Algorithmus an zwei bekannten Beispielen testen. Um einen Vergleich herstellen zu können, wählen wir – wie M. Fukushima in seinem Paper „A descent algorithm for nonsmooth convex optimitzation“ [4] – die beiden Beispiele *MAXQUAD* und *TR48*. Diese beiden Beispiele sollen kurz vorgestellt werden. Wir haben sie dem Kapitel „A set of nonsmooth optimization test problems“ aus „Nonsmooth optimization“ [10] (ab S.151) von C. Lemaréchal und R. Mifflin entnommen.

1. Bei MAXQUAD wird eine Zielfunktion $f : \mathbb{R}^{10} \longrightarrow \mathbb{R}$ minimiert. Diese ist definiert durch

   $$f(x) = \max_{k=1,...,5} (x^T A_k x - b_k^T x),$$

   wobei $A_k \in \mathbb{R}^{10\times 10}$ eine symmetrische Matrix und b_k ein Vektor im $\mathbb{R}^{10}$ für alle $k = 1, ..., 5$ ist. Die symmetrische Matrix ist durch

   $$A_k(i,j) = e^{\frac{i}{j}} \cos(i \cdot j) \sin(k) \qquad i < j,$$

   mit Diagonalelementen

   $$A_k(i,i) = \frac{i}{10} |\sin(k)| + \sum_{j=1, j\neq i}^{10} |A_k(i,j)|$$

gegeben und der Vektor durch

$$b_k(i) = e^{\frac{i}{k}} \sin(i \cdot k).$$

2. Bei TR48 wird eine Zielfunktion $f : \mathbb{R}^{48} \longrightarrow \mathbb{R}$ minimiert. Diese ist definiert durch

$$f(x) = -\left\{ \sum_{i=1}^{48} s_i x_i + \sum_{j=1}^{48} d_j \min_{i=1,\ldots,48} (a_{ij} - x_i) \right\},$$

wobei $A \in \mathbb{R}^{48\times 48}$ eine symmetrische Matrix und s, d Vektoren im $\mathbb{R}^{48}$ sind. Es handelt sich um das duale Transportproblem mit 48 Quellen und 48 Zielen. Die Matrix mit den 2304 Kosten als Einträge und die Einträge der beiden Vektoren s (sources) und d (destinations) kann man in [10] (S.162/163) nachlesen.

Mit diesen beiden Beispielen wollen wir nun den Quasi-Newton-Bündel-Algorithmus testen. Bei der Implementation des Algorithmus halten wir uns an den in Kapitel 4, 5 und 8 beschriebenen Aufbau. Dazu sollen noch einige Erklärungen anhand des Quelltextes folgen.
Der gesamte Algorithmus sieht wie folgt aus:

```
function [x,iter,f,FZeval]=QNBTM()

global func subg;
format long e;

% Quasi-Newton bundle-type methods for nondifferentiable convex
% optimization

%*************************************************************

% INPUT-Parameter:
%       func      function to be minimized
%       subg      subgradient of func
%       x_0       initial iterate
%       max_iter  maximal number of iterations

% OUTPUT-Parameter:
%        x       approximate solution
%        iter    number of iterations performed
%        f       value of func in x
%        FZeval  number of function-/subgradientevaluations
```

```
%*****************************************************************

% INPUT:

% function to be minimized:
func=@funcMAXQUAD;
% subgradient of func:
subg=@subgMAXQUAD;
% initial iterate:
x_0=ones(10,1);
% maximal number of iterations:
max_iter=60;

%*****************************************************************

% parameter

sigma=0.0001;
% 0<sigma<0.5
rho=0.5;
% 0<rho<1;
tol=1.e-4;
% tol very small
M=0.5*eye(length(x_0));
% M positive definite and symmetric matrix

%*****************************************************************

% further definitions

xk=x_0;
k=0;
invM=inv(M);
invBk=invM;
deltak=1;
FZeval=[];

%*****************************************************************

% bundle algorithm for the beginning
```

```
[lowerFk, dk, epsilonk, j_sum_max] = Bundle(M, xk, deltak);

%************************************************************

% QNBT-algorithm

while (norm(M*dk)>=tol) & (k<max_iter)

 % computation of the search direction

 sk=invBk*M*dk;

 % line search

 m=0;
 epsilonkold=epsilonk;
 deltakold=deltak;
 deltak=(1/2)^(k+1);
 xkold=xk;
 dkold=dk;

 [fk]=feval(func, xk)
 upperFk=fk+0.5*dk'*M*dk;

 while (1)

  xk=xkold+rho^m*sk;

  FZeval=[FZeval;k,j_sum_max];
  [lowerFk, dk, epsilonk, auxil_var] = Bundle(M, xk, deltak);
  j_sum_max=auxil_var+j_sum_max;

  if (lowerFk<=upperFk-sigma*rho^m*sk'*M*dk)

   break

  else

   m=m+1;

  end
```

```
 end

 tauk=rho^m;

 % computation of the matrix with BFGS-Update

 c3=1;
 % c3 in (0,infinity)
 c4=0.2;
 % c4 in (0,1)

 Deltaxk=xk-xkold;
 Deltayk=-M*(dk-dkold);

 if sqrt(Deltaxk'*M*Deltaxk)*(sqrt(2*epsilonkold)+...
    sqrt(2*epsilonk))<=c3*Deltaxk'*Deltayk  & ...
    2*sqrt(Deltayk'*M*Deltayk)*(sqrt(2*epsilonkold)+...
    sqrt(2*epsilonk))<=min(c4, deltakold^(1/3)+...
    deltak^(1/3))*Deltayk'*Deltayk

  denom=Deltaxk'*Deltayk;
  invBk=(eye(size(M))-((Deltaxk*Deltayk')/denom))*invBk*...
  (eye(size(M))-((Deltayk*Deltaxk')/denom))+...
  ((Deltaxk*Deltaxk')/denom);

 else

  invBk=invM;

 end

 k=k+1

end

[fk]=feval(func,xk);

%*****************************************************************

% OUTPUT:

% approximate solution:
```

```
x=xk
% number of iterations performed:
iter=k
% value of func in x:
f=fk
% number of function-/subgradientevaluations:
FZeval=[FZeval;k,j_sum_max]

%*****************************************************************
```

Der Benutzer muss also den Startwert, die maximale Anzahl an Iterationen, die Zielfunktion und auch eine Funktion zur Berechnung der Subgradienten eingeben. Letzteres ist der einzig nicht so schöne Punkt des Algorithmus. Der Benutzer muss eine geeignete Funktion finden, mit der der Subgradient an den gewünschten Stellen ausgerechnet werden kann. Die Parameter können ebenfalls abhängig von den Beispielen geändert werden. Am Anfang wird die Funktion `Bundle` aufgerufen. Dort befindet sich die in Kapitel 4 vorgestellte Bündel-Methode und hat den folgenden Quelltext:

```
function [lowerFj, dj, epsilonj, jmax] = Bundle(M, x, delta)

global func subg;

%*****************************************************************

% this program solves the bundle problem

%*****************************************************************

N=1;
j=1;
U=x;
[F]=feval(func, x);
[Z]=feval(subg, x);

while (1)

 [lowerFj, dj]=funclower(x, M, U, F, Z);
 uj=x+dj;
 [f]=feval(func, uj);
 [z]=feval(subg, uj);
 F = cat(1,F,f);
 Z = cat(2,Z,z);
```

```
 U = cat(2,U,uj);
 upperFj=f+0.5*dj'*M*dj;
 epsilonj=upperFj-lowerFj;

 if (epsilonj>delta*min(dj'*M*dj,N))

  j=j+1;

 else

  break

 end

end

jmax=j+1;

%*****************************************************************
```

Dieser Bündel-Algorithmus beginnt mit der Berechnung der unteren Schranke $\check{F}_M^j(x)$ in der Funktion `funclower`. Dies erfolgt durch das Lösen der MinMax-Aufgabe

$$\text{Min.} \qquad \max_{i=1,\ldots,j}\{f(u^i) + z_{u^i}^T(x+d-u^i)\} + \frac{1}{2}\|d\|_M^2, \quad d \in \mathbb{R}^n.$$

Dieses Problem lässt sich in

$$\text{Min.} \qquad v + \frac{1}{2}d^T M d, \quad v \in \mathbb{R}, d \in \mathbb{R}^n$$

unter den Nebenbedingungen

$$f(u^i) + z_{u^i}^T(x+d-u^i) \leq v$$

umwandeln. Damit erhalten wir den folgenden Quelltext:

```
function [lowerf, d]=funclower(x, M, U, F, Z)

%*****************************************************************

% this program solves the minmax problem

%*****************************************************************
```

```
[n,j]=size(U);

% we use the Optimization toolbox in MATLAB
% X=quadprog(H,f,A,b) solves the following quadratic problem:
% min 0.5*x'*H*x + f'*x subject to A*x <=b, x

H=zeros(n+1,n+1);
H(1:n,1:n)=M;
f=zeros(n+1,1);
f(n+1)=1;
A=cat(2,Z',-ones(j,1));
b=zeros(j,1);

for i=1:j

 b(i)=-F(i)-Z(:,i)'*(x-U(:,i));

end

L=quadprog(H,f,A,b,[],[],[],[],[],optimset('Largescale','off'));
d=L(1:n);
v=L(n+1);
lowerf=v+1/2*d'*M*d;

%*****************************************************************
```

Hierbei haben wir aus der Optimization toolbox in MATLAB die Funktion `quadprog` verwendet. Um diese im Quelltext kurz beschriebene Funktion anwenden zu können, müssen wir unser Problem in Matrixschreibweise bringen. Das Ursprungsproblem lässt sich umschreiben in

$$\text{Min.} \quad \frac{1}{2}\begin{pmatrix} d \\ v \end{pmatrix}^T \begin{pmatrix} M & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} d \\ v \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} d \\ v \end{pmatrix}, \quad \begin{pmatrix} d \\ v \end{pmatrix} \in \mathbb{R}^{n+1}$$

unter den Nebenbedingungen

$$\begin{pmatrix} z_{u^1}^T & -1 \\ \vdots & \vdots \\ z_{u^j}^T & -1 \end{pmatrix} \begin{pmatrix} d \\ v \end{pmatrix} \leq \begin{pmatrix} -f(u^1) - z_{u^1}^T(x-u^1) \\ \vdots \\ -f(u^j) - z_{u^j}^T(x-u^j) \end{pmatrix}.$$

`Quadprog` liefert uns so unsere gewünschte Lösung $d^j(x)$, mit der der Bündel-Algorithmus fortfahren kann. Nach endlich vielen Schritten bricht der Bündel-Algorithmus ab und wir können den Quasi-Newton-Bündel-Algorithmus starten. Bei der Berechnung der Schrittweite setzen wir vereinfacht ρ konstant

auf 1/2. Beim anschließenden BFGS-Update berechnen wir gleich die inverse Matrix, da wir nur diese bei der Berechnung der Abstiegsrichtung benötigen. Dafür verwenden wir die Formel

$$B_{k+1}^{-1} = \left(I - \frac{\Delta x_k \Delta y_k^T}{\Delta x_k^T \Delta y_k}\right) B_k^{-1} \left(I - \frac{\Delta y_k \Delta x_k^T}{\Delta x_k^T \Delta y_k}\right) + \frac{\Delta x_k \Delta x_k^T}{\Delta x_k^T \Delta y_k}$$

(siehe S. 197 in [20]), wobei I die Identität ist. Der gesamte Algorithmus bricht entweder nach der maximalen Iterationszahl max_iter ab oder falls $\|Md_k\| < tol$ ist, wobei tol (tolerance) eine eingegebene Genauigkeitsgrenze ist. Danach liefert der Algorithmus den minimalen Funktionswert f, die Lösung x der Minimierungsaufgabe (P), die benötigte Anzahl an Iterationen $iter$ und die kumulierte Anzahl der Funkions-/Subgradientenauswertungen $FZeval$. An einigen Stellen lässt sich der Algorithmus auch effizienter programmieren (siehe [22]), worauf es uns aber nicht so sehr ankam.
Testen wir nun die beiden Beispiele. Nach den Tests sollen in einer kurzen Bemerkung zwei vorkommende Besonderheiten geklärt werden. Die Parameter bei unserem ersten Problem MAXQUAD kann man im Quelltext ablesen. Die optimale Lösung liegt hier bei

$$f = -0.8414$$

x = [-0.1263, -0.0343, -0.0068, 0.0264, 0.0673, -0.2784, 0.0742, 0.1385, 0.0840, 0.0386].

Die folgende Tabelle zeigt die Konvergenz mit teuren kumulierten Funktions-/Subgradientenauswertungen:

iter	0	1	2	3	4
f	5337	-0.3601	-0.8394	-0.8414	-0.8414
FZeval	185	267	396	562	848

M. Fukushima dagegen benötigte 18 Iterationen, um das Ergebnis zu erhalten.
Auch bei dem Startwert $x_i = 0$ für $i = 1, ..., 10$, an dem die Funktion f einen Knick hat $(f_k(0) = 0, k = 1, ..., 5)$, erhalten wir nach 4 Iterationen das gewünschte Minimum. Wählen wir dagegen erneut den Startwert $x_i = 1$ für $i = 1, ..., 10$, und wählen wir dieses Mal $M = 10 \cdot I$, wobei I die Identität ist, so erhalten wir die folgende Tabelle:

iter	0	1	3	5	10	14
f	5337	9.4482	-0.4045	-0.8261	-0.8414	-0.8414
FZeval	26	41	68	106	257	466

Wir benötigen zwar mehr Iterationen, dafür sind die Funktions-/Subgra-

dientenauswertungen nicht so teuer wie vorher.
Die Parameter für unser zweites Problem TR48 sollen etwas anders gewählt werden. Wir setzen $M = 0.8 \cdot I$, wobei I die Identität ist und $\delta_k = (1/1.2)^{k+1}$ für die konvergente Reihe, welche die Genauigkeit der Approximationen mitbestimmt. Als Startwert wählen wir $x_i = 0$ für $i = 1, ..., 48$. Die optimale Lösung liegt hier bei

$$f = -638565$$

x = [12.1042, 125.1042, -131.8958, 351.1042, -42.8958, -296.8958, -203.8958, -383.8958, -219.8958, -309.8958, 179.1042, -5.8958, -124.8958, -266.8958, 26.1042, 77.1042, -30.8958, -223.8958, 97.1042, -51.8958, -36.8958, -60.8958, -375.8958, -29.8958, -143.8958, 0.1042, 205.1042, -70.8958, -27.8958, -8958.8958, 129.1042, -13.8958, -32.8958, -377.8958, 24.1042, -401.8958, 198.1042, -261.8958, 820.1042, -193.8958, 29.1042, 352.1042, -9.8958, 342.1042, 954.1042, 729.1042, -301.8958, 74.1042].

Die folgende Tabelle zeigt die Konvergenz:

iter	0	1	3	5	10
f	-464816	-533568.12	-576914.45	-605851.22	-621152.47
FZeval	34	69	153	245	534
iter	15	20	25	30	35
f	-634157.28	-637036.21	-638170.97	-637782.11	-638497.12
FZeval	824	1118	1642	1985	2339
iter	40	45	47	50	51
f	-638542.19	-638555.20	-638563.39	-638564.97	-638564.99
FZeval	3024	3940	4091	6024	7119

M. Fukushima dagegen benötigte 70 Iterationen, um in die Nähe des Ergebnises zu kommen (-638446). Seine Feststellung „... but it appears very difficult in practice to achieve this value closely...“ können wir nur bestätigen. Es war schwierig geeignete Parameter zu finden, um das Ergebnis in akzeptabler Zeit zu erhalten. Die letzten drei Iterationen benötigen nach wie vor sehr viele Funktions-/Subgradientenauswertungen. Auch C. Lemaréchal und R. Mifflin beschrieben das Problem als schwierig, gaben aber als Alternative dazu ein einfacheres Problem an, nämlich indem man die Einträge der Vektoren s und d auf Eins setzt. Das konnten wir bestätigen, denn bei geschickter Wahl der Parameter konnten wir in nur 5 Iterationsschritten die optimale Lösung

$$f = -9870$$

x = [-37.9204, 25.9313, -2.9204, 8.6185, -61.9204, 9.7242, -7.6708, 0, 15.6146, -74.3854, -7.9370, 11.0560, -53.9874, -70.5313, 7.0796, -7.8467, -14.6769,

19.4687, -18.2619, 0, -55.9204, -43.9482, -54.5052, -12.9204, -46.9204, -18.9204, -7.6694, -53.9204, -15.9204, 144.6613, 0, 3.0796, 169.0796, 0, -28.7973, 0, 79.0796, 44.4687, -10.9722, -31.3854, 46.0796, 17.0817, 40.0796, 69.0796, -71.9722, 73.0278, 4.5405, 24.0796]

finden. Hierbei konnte man deutlich feststellen, wie sehr die Matrix M Einfluss auf die Konvergenzgeschwindigkeit hat. Hier haben wir $M = 0.015 \cdot I$ gesetzt, wobei I die Identität ist. Setzt man zum Beispiel $M = 0.1 \cdot I$, so benötigt man schon 19 Iterationen beziehungsweise setzt man $M = I$, so sollte man nicht in akzeptabler Zeit ein Ergebnis erwarten.

Bemerkungen: Wir wollen noch zwei Dinge erwähnen.

- In $iter = 0$ wird normalerweise nur der Funktionswert in dem Startwert x_0 berechnet. Also benötigt man eine Funktions-/Subgradientenauswertung. Da man in unserem Algorithmus aber für die erste Iteration schon Startwerte aus dem Bündel-Algorithmus benötigt, haben wir die dafür benötigten Funktions-/Subgradientenauswertungen schon in $iter = 0$ berücksichtigt.

- Da man die Regularisierung nicht exakt berechnen kann, werden für die Berechnung der Schrittweite auch nur ihre Schranken verwendet. Daher kann man nicht erwarten, dass die Folge $\{F_M(x_k)\}$ monoton fällt. Im Gegensatz zur Berechnung der Armijo-Schrittweite im differenzierbaren Fall kann man hier nicht erwarten, dass $F_M(x_{k+1}) \leq F_M(x_k)$ gilt.

□

9.2 Zusammenfassung

Um noch einmal die wichtigsten Aussagen dieser Arbeit hervorzuheben, soll die Arbeit mit einer kurzen Zusammenfassung enden. Unser Ziel war es einen Algorithmus zu beschreiben, der den minimalen Punkt einer nichtdifferenzierbaren konvexen Funktion findet. Da es bei differenzierbaren Optimierungsproblemen schon sehr gute Verfahren gibt, wird unsere nichtdifferenzierbare Funktion durch eine differenzierbare Funktion – die Moreau-Yosida-Regularisierung – angenähert. Diese wiederum kann nicht exakt berechnet werden. Daher wird sie durch eine lineare Approximation beschrieben, was durch die Bündel-Methode geschieht. Hierauf nun kann man ein Quasi-Newton-Verfahren anwenden. Durch zusätzliche Voraussetzungen an die Regularisierung kann globale und superlineare Konvergenz bewiesen werden. Der Algorithmus lässt sich gut implementieren und liefert bei geeigneter Wahl der Parameter die optimale Lösung in akzeptabler Zeit.

Literaturverzeichnis

[1] J.Bonnans, J.Gilbert, C.Lemaréchal and C.Sagastizábal, *A family of variable metric proximal methods*, Mathematical Programming, 68 (1995), pp.15-47.

[2] R.H.Byrd and J.Nocedal, *A tool for the analysis of quasi-Newton methods with application to unconstrained minimization*, SIAM Journal on Numerical Analysis, 26 (1989), pp.727-739.

[3] J.E.Dennis and J.J.Moré, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Mathematics of Computation, 28 (1974), pp.549-560.

[4] M.Fukushima, *A descent algorithm for nonsmooth convex optimization*, Mathematical Programming, 30 (1984), pp. 163-175.

[5] M.Fukushima and L.Qi, *A globally and superlineary convergent algorithm for nonsmooth convex minimization*, SIAM Journal on Optimization, 6 (1996), pp.1106-1120.

[6] C.Geiger, C.Kanzow, *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*, Springer-Verlag, Berlin, Heidelberg (1999).

[7] J.B.Hiriart-Urruty and C.Lemaréchal, *Convex Analysis and Minimization Algorithms 1*, Springer-Verlag, Berlin, Heidelberg (1993).

[8] J.B.Hiriart-Urruty and C.Lemaréchal, *Convex Analysis and Minimization Algorithms 2*, Springer-Verlag, Berlin, Heidelberg (1993).

[9] K.C.Kiwiel, *Methods of descent for nondifferentiable optimization*, in Lecture Notes in Mathematics, Springer-Verlag, Berlin, Heidelberg (1985).

[10] C.Lemaréchal and R.Mifflin, *Nonsmooth Optimization*, Pergamon Press, Oxford (1978).

[11] C.Lemaréchal and C.Sagastizábal, *Practical Aspects of the Moreau-Yosida regularization 1: Theoretical preliminaries*, SIAM Journal on Optimization, 7 (1997), pp. 367-385.

[12] C.Lemaréchal and C.Sagastizábal, *Variable metric bundle methods: From conceptual to implementable forms*, Mathematical Programming, 76 (1997), pp.393-410.

[13] R.Mifflin, D.Sun and L.Qi, *Quasi-Newton bundle-type methods for non-differentiable convex optimization*, SIAM Journal on Optimization, 2 (1998), pp. 583-603.

[14] J.Nocedal, S.J.Wright, *Numerical optimization*, Springer-Verlag, New York (1999).

[15] L.Qi, *On superlinear convergence of quasi-Newton methods for non-smooth equations*, Operations Research Letters, 20 (1997), pp.223-228.

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

[17] R.T.Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Mathematics of Operations Research, 1 (1976), pp.97-116.

[18] N.Z.Shor, *Minimization Methods for Nondifferentiable Functions*, Springer-Verlag, Berlin (1985).

[19] J.Werner, *Numerische Mathematik 1*, Vieweg-Verlag, Braunschweig-Wiesbaden (1992).

[20] J.Werner, *Numerische Mathematik 2*, Vieweg-Verlag, Braunschweig-Wiesbaden (1992).

[21] J.Werner, Private Kommunikation, Göttingen (2002).

[22] J.Werner, *Unrestringierte Optimierung*, Skript zur Vorlesung, Göttingen (2002).

[23] C.Zhu, *Asymptotic convergence analysis of some inexact proximal point algorithms for minimization*, SIAM Journal on Optimization, 6 (1996), pp.626-637.

Ich möchte mich bei Herrn Professor Dr. Jochen Werner für die interessante Themenstellung sowie für seine Hilfe während der Erstellung dieser Arbeit bedanken.

Mein Dank gilt meinen Eltern, die mir dieses Studium ermöglicht haben und auch allen anderen, die mich immer unterstützt haben.