6.8 Newton Verfahren und Varianten

In den vorherigen Kapiteln haben wir grundlegende Gradienten-basierte Verfahren kennen gelernt, die man zur numerischen Optimierung von (unbeschränkten) Variationsmethoden einsetzen kann. Typischerweise zeigen diese Verfahren ein Konvergenzverhalten erster Ordnung. Möchte man die Konvergenzordnung deutlich verbessern, so macht es Sinn zu Newton-artigen Verfahren überzugehen. Das ist insbesondere bei Anwendungen sinnvoll, in denen eine relativ hohe Genauigkeit einer Lösung essentiell ist. In diesem Kapitel werden wir deshalb grundlegende Konzepte Newton-artiger Verfahren für unbeschränkte nichtlineare Variationsmethoden der Form

$$\min_u J(u) \tag{6.13}$$

mit J zweimal stetig Frechet differenzierbar behandeln, Varianten des Newton-Verfahrens für unterschiedliche Anwendungen kennen lernen und auf ihr Konvergenzverhalten eingehen.

6.8.1 (Exaktes) Newton Verfahren

In der numerischen Analysis ist das Newton-Verfahren (oder auch Newton-Raphson Verfahren) ein Verfahren zur Bestimmung von Nullstellen einer Gleichung in einer oder mehr Dimensionen. Die grundlegende Idee des Newton Verfahrens für eine nichtlineare Gleichung wie z.B.

$$J'(u) = 0 \ , \tag{6.14}$$

der notwendigen Optimalitätsbedingung unseres Ausgangsproblems (6.13), ist eine *lokale Linearisierung* an der Stelle u_k um

$$u_{k+1} = u_k + d_k \tag{6.15}$$

zu berechnen, d.h.

$$J'(u_k) + J''(u_k)\, d_k \ = \ 0 \ \Leftrightarrow \ J''(u_k)\, d_k \ = \ -J'(u_k) \tag{6.16}$$

wobei J' die Jacobi Matrix und J'' die Hesse Matrix von J darstellt, und die Suchrichtung

$$d_k = -J''(u_k)^{-1} J'(u_k)$$

als sogenannte Newton-Richtung bezeichnet wird. Für eine Visualisierung des (exakten) Newton Verfahrens für die Gleichung (6.14) betrachten wir Abbildung 6.6. Eine zweite Interpretation des Newton Verfahrens aus Sicht der numerischen Optimierung

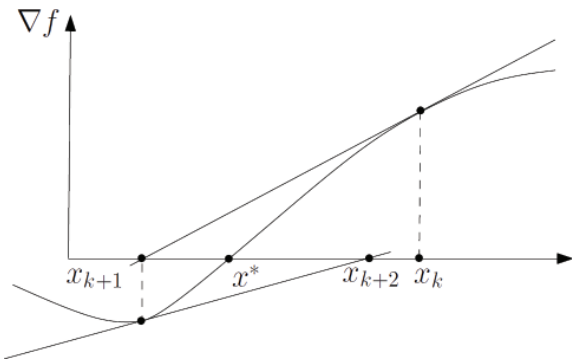

Figure 6.6: Erste Interpretation: Newton-Verfahren als lokale Linearisierung

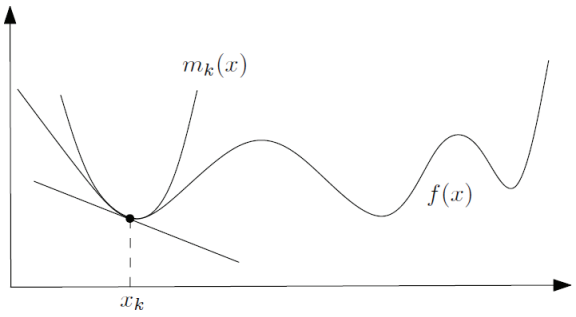

Figure 6.7: Zweite Interpretation: Newton-Verfahrens durch quadratisches Modell

erhält man durch ein quadratisches Modell des Zielfunktionals bzgl. der Suchrichtung d (ein quadratisches Modell kann verhältnismäßig leicht gelöst werden), d.h. einer Taylor Approximation zweiter Ordnung von J um u_k,

$$\begin{aligned} M_k(u_k + d) &:= J(u_k) + J'(u_k)d + \frac{1}{2}dJ''(u_k)d \\ &\approx J(u_k + d) \ . \end{aligned} \tag{6.17}$$

Eine Visualisierung dieser zweiten Interpretation ist in Abbildung 6.7 zu sehen. Falls die Hessematrix $J''(u_k)$ positiv definit ist, dann ist der Minimierer d_k von M_k eine eindeutige Lösung von $M_k'(u_k + d) = 0$, d.h.

$$0 = J'(u_k) + J''(u_k)d$$

und damit erhält man wieder das (exakte) Newton-Verfahren

$$\begin{aligned} d_k &= -J''(u_k)^{-1}J'(u_k) \\ \Leftrightarrow u_{k+1} &= u_k - J''(u_k)^{-1}J'(u_k) \ . \end{aligned}$$

Selbstverständlich wird dabei nicht explizit die Inverse der Hessematrix berechnet, sondern stattdessen zur Berechnung der Suchrichtung d_k in (6.16) ein lineares Gleichungssystem gelöst und anschließend ein Update gemäß (6.15) durchgeführt. Verwendet man zusätzlich eine effiziente Schrittweite σ_k, so spricht man vom gedämpften

Newton-Verfahren. Startet man nicht sehr nahe an einem Minimum, dann ist eine effiziente Schrittweite in der Regel < 1, d.h., die Schrittweite des Newton-Verfahrens wird gedämpft.

Algorithm 6.8.1 (Gedämpftes Newton Verfahren)**.**
Wähle Startwert u_0 und setze $k := 0$.
Ist $J'(u_k) = 0$, % Stopp.
Berechne d_k aus

$$J''(u_k)d_k = -J'(u_k) \ ,$$

und mit effizienter Schrittweite σ_k setze

$$u_{k+1} = u_k + \sigma_k d_k.$$

Setze $k := k + 1$.

6.8.2 Varianten des Newton-Verfahrens

In diesem Abschnitt diskutieren wir unterschiedliche Varianten des Newton bzw. Variable-Metrik Verfahrens. Wir wollen dabei zwei Fragestellungen mit berücksichtigen: *Wie schnell ist die lokale Konvergenz der Verfahren? Kann man eine Konvergenz für jeden Startwert garantieren?* (globale Konvergenz)

Um das Konvergenzverhalten unterschiedlicher Newton-artiger Verfahren vergleichen zu können, wollen wir unterschiedliche Konvergenzraten einführen.

Definition 6.8.2 (Konvergenzraten)**.** *Angenommen $u_k \to u^*$. Dann konvergiert die Folge u_k:*

(i) Q-linear ("Quotient"-linear) $\Leftrightarrow$

$$\|u_{k+1} - u^*\| \leq C \|u_k - u^*\| \ \text{, mit } C < 1$$

für alle $k \geq k_0$.

$$\Leftrightarrow \ \limsup_{k\to\infty} \frac{\|u_{k+1} - u^*\|}{\|u_k - u^*\|} < 1$$

(ii) Q-superlinear $\Leftrightarrow$

$$\|u_{k+1} - u^*\| \leq C \|u_k - u^*\| \ \text{, mit } C \to 0$$

$$\Leftrightarrow \ \limsup_{k\to\infty} \frac{\|u_{k+1} - u^*\|}{\|u_k - u^*\|} = 0 \ .$$

(iii) Q-quadratisch $\Leftrightarrow$

$$\|u_{k+1} - u^*\| \leq C \|u_k - u^*\|^2 \text{ , mit } C < \infty$$

$$\Leftrightarrow \limsup_{k\to\infty} \frac{\|u_{k+1} - u^*\|}{\|u_k - u^*\|^2} < \infty \ .$$

Beispiele: ...

Jede Iteration der Form

$$u_{k+1} = u_k - A_k^{-1} J'(u_k) \tag{6.18}$$

mit A_k invertierbar, bezeichnet man als *Newton-artige Iteration* in der Optimierung. Für $A_k = J''(u_k)$ erhält man wieder das (exakte) Newton Verfahren, üblicherweise wählt man in der Praxis aber $A_k \approx J''(u_k)$. Man spricht bei Verfahren der Form (6.18) auch von Variable-Metrik Verfahren. Ist $A = J''(u)$ positiv definit, dann wird durch die Norm bzgl. A

$$\|u\|_A = \sqrt{\langle u, u\rangle_A} = \sqrt{\langle u, Au\rangle}$$

eine Metrik definiert. Zum Beispiel erhält man bei einer sehr einfachen Metrik bzgl. $A_k = I$, als Spezialfall wieder das bekannte Gradientenabstiegsverfahren. Der Vorteil des gedämpften Newton-Verfahrens gegenüber dem Gradientenverfahren ist, dass mit $J''(u_k)$ auch Informationen über die Krümmung von J in $u = u_k$ benutzt werden. Beim gedämpften Newton-Verfahren wird die Metrik zur Bestimmung des steilsten Abstiegs "variabel" an die Krümmung von J angepasst.

Allgemein ist in jeder Iteration ein quadratisches Teilproblem zu lösen, um eine neue Suchrichtung d_k zu bestimmen:

$$d_k = \arg\min_d M_k(u_k + d) \ .$$

mit

$$M_k(u_k + d) = J(u_k) + J'(u_k)d + \frac{1}{2} dA_k(u_k)d \ ,$$

analog zu (6.17). Die Optimalitätsbedingung dieses Modells führt zum Update einer Suchrichtung in einer Newton-artigen Iteration:

$$\begin{aligned} 0 = M_k'(u_k + d_k) &= A_k d_k + J'(u_k) \\ \Longleftrightarrow\ d_k &= -A_k^{-1} J'(u_k) \ . \end{aligned}$$

Man beachte, dass d_k nur dann ein Minimierer von $M_k(u_k+d_k)$ ist, falls $A_k \succ 0$ (positiv definit). Für das (exakte) Newton Verfahren muss das nicht notwendigerweise der Fall sein, falls u_k weit weg ist von einer Lösung u^*.

Lemma 6.8.1 (Abstiegsrichtung)**.** *Falls $A_k \succ 0$, dann ist $d_k = -A_k^{-1} J'(u_k)$ eine Abstiegsrichtung.*

Proof.

$$J'(u_k)d_k = \underbrace{-J'(u_k)\underbrace{A_k^{-1}}_{\succ 0} J'(u_k)}_{>0} < 0 \ .$$

□

Definition 6.8.3 (Varianten des Newton Verfahrens)**.** *Varianten des Newton Verfahrens, die häufig in der Literatur verwendet werden, sind u.a. folgende:*

(a) *Quasi-Newton Verfahren*
Das Newton-Verfahren konvergiert zwar lokal quadratisch, jedoch liegt ein Problem bei der Anwendung des (exakten) Newton-Verfahrens darin, dass man die Hessematrix des Funktionals benötigt. Gerade bei hochdimensionalen Problemen in der Bildverarbeitung kann es in der Praxis schwierig sein die vollständige Hessematrix eines zu minimierenden Funktionals J zu berechnen (z.B. Speicherprobleme, Probleme mit der Rechenzeit).

Für solche Situationen wurden sogenannte *Quasi-Newton Verfahren* entwickelt, die eine Approximation der Hessematrix verwenden. Man approximiert dabei die Hessematrix A_{k+1} rekursiv mittels einer alten Approximation der Hessematrix A_k und Auswertungen der ersten Ableitung, $J'(u_{k+1})$ und $J'(u_k)$. Betrachtet man eine Taylor-Entwicklung von J'

$$J'(u_k) = J'(u_{k+1}) + J''(u_{k+1})(u_k - u_{k+1}) + o(\|u_k - u_{k+1}\|) \ ,$$

so erhält man mit $d_k = u_{k+1} - u_k$ die folgende wichtige Gleichung

$$A_{k+1}(u_{k+1} - u_k) = J'(u_{k+1}) - J'(u_k) \quad \text{(Sekanten-Bedingung)} \ . \tag{6.19}$$

Man bezeichnet diese Gleichung auch als Quasi-Newton-Gleichung. Der Preis für die Approximation der Hessematrix ist ein Verlust an Konvergenzgeschwindigkeit. Man kann beweisen, dass Quasi-Newton-Verfahren lokal superlinear konvergieren.

Das wichtigste Verfahren der Quasi-Newton-Klasse ist das sogenannte BFGS-Verfahren, das auf der folgenden Updateformel basiert,

$$A_{k+1} = A_k - \frac{A_k d d^T A_k}{d^T B_k d} + \frac{y y^T}{d^T y}$$

mit d und y definiert als

$$\begin{aligned} d &= u_{k+1} - u_k \ , \\ y &= J'(u_{k+1}) - J'(u_k) \ . \end{aligned}$$

Die Updateformel wurde von **B**royden, **F**letcher, **G**oldfarb und **S**hanno unabhängig voneinander gefunden. Man kann leicht nachrechnen, dass $A_{k+1}s = y$ gilt, also damit die Quasi-Newton-Gleichung in (6.19) erfüllt ist. Das BFGS-Verfahren ist eine sehr erfolgreiche Methode und man kann zeigen, dass die Folge der A_k gegen die Hessematrix J'' an der Stelle u^* konvergiert.

(b) *Gauss-Newton und Levenberg-Marquardt*
Einen schönen Zusammenhang zwischen Newton-Verfahren und inversen Problemen (z.B. in der Bildgebung) liefert die Klasse der Gauss-Newton und Levenberg-Marquardt Verfahren. Insbesondere für *nichtlineare* inverse Probleme der Form

$$F(u) - y = 0$$

mit einem *nichtlinearen* Operator F und gegebenen Daten y ist diese Art der Betrachtung interessant. Die Grundidee des Newton-Verfahrens für diese Gleichung ist eine lokale Linearisierung (1. Interpretation, Abbildung 6.6). Ein Schritt des Newton-Verfahrens würde die Lösung des linearen Systems

$$F'(u_k)\,(u_{k+1} - u_k) = -\,(F(u_k) - y) \tag{6.20}$$

beinhalten. Da im Fall eines inversen Problems $F'(u_k)$ keinen *regulären* linearen Operator darstellt, ist die Gleichung in (6.20) selbst ein lineares schlechtgestelltes Problem und folglich ist u_{k+1} nicht wohldefiniert. Eine übliche Strategie zur Konstruktion von Newton-Verfahren für nichtlineare schlechtgestellte Probleme ist (6.20) mit Hilfe einer Regularisierungstechnik für lineare schlecht-gestellte Probleme zu erweitern. Zum Beispiel erhält man durch Anwendung einer linearen Tikhonov-Regularisierung (betrachte $u_{k+1} - u_k$ als Unbekannte) das sogenannte *Levenberg-Marquardt Verfahren*

$$(F'(u_k)^* F'(u_k) + \alpha_k I)\,(u_{k+1} - u_k) = -F'(u_k)^*\,(F(u_k) - y)\ .$$

Im Sinne von Newton-Verfahren für Funktionale in Variationsmethoden kann man das folgendermaßen erklären. Für das zu minimierende L^2 Fitting-Funktional

$$J(u) := \frac{1}{2}\,\|F(u) - y\|_2^2$$

betrachten wir das folgende quadratische Modell

$$\begin{aligned} M_k(u_k + d) &= \frac{1}{2}\,\|F(u_k) - y + F'(u_k)d\|_2^2 \ + \ \frac{\alpha_k}{2}\,\|d\|_2^2 \\ &= \frac{1}{2}\,\|F(u_k)\|_2^2 + \langle F'(u_k)d, F(u_k) - y\rangle + \langle d, (F'(u_k)^* F'(u_k) + \alpha_k I)\,d\rangle \end{aligned}$$

Man beachte, dass α_k hier einen variablen Regularisierungsparameter darstellt und nicht mit einer Schrittweite σ_k verwechselt werden sollte. Im Vergleich zu Quasi-Newton-Verfahren haben wir hier

$$A_k = F'(u_k)^* F'(u_k) + \alpha_k I$$

und als Suchrichtung $d_k = -A_k^{-1} J'(u_k)$.

Nun stellt sich die Frage: Wann ist A_k nahe bei $J''(u_k)$? Berechnet man die Hessematrix $J''(u_k)$, so erhält man für den Unterschied

$$J''(u_k) - A_k = \langle (F(u) - y)'', F(u) - y \rangle \ ,$$

d.h. der Fehler wird klein, falls

a) die Komponenten von $(F(u) - y)''$ klein sind (F nahezu linear)

b) die Residuen $F(u) - y$ klein sind (gute Anpassung, wenig Restrauschen).

Hier ist wieder die Notwendigkeit einer Regularisierung sichtbar und zeigt, dass man (nur) im Fall einer Lösung mit perfektem Fitting eine *lokal quadratische* Konvergenz bei den letzten Iterierten erwarten kann. Im Allgemeinen kann man nur Q-lineare Konvergenz erwarten.

(c) *Inexakte Newton bzw. Newton-Krylow-Verfahren*
Bei inexakten Newton-Verfahren löst man das lineare System

$$J''(u_k)d = -J'(u_k)$$

in jedem Schritt des Newton-Verfahrens inexakt, z.B. durch iterative lineare Algebra. Dieser Ansatz ist gut geeignet für large-scale Probleme. Analog zur Minimierung von Variationsproblemen bietet sich auch die numerische Lösung nichtlinearer partieller Differentialgleichungen prinzipiell das Newton-Verfahren als Grundlöser an. Die entsprechende Jacobi-Matrix ist immer dünnbesetzt und so bieten sich Krylow-Unterraum-Verfahren zur Lösung der linearen Gleichungssysteme an. Man spricht dann von Newton-Krylow-Verfahren. Ein wichtiger Repräsentant dieser Klasse ist das Newton-CG Verfahren. Im Krylow-Verfahren selbst tritt die Jacobi-Matrix nur in Matrix-Vektorprodukten auf, welche als Richtungsableitungen interpretiert werden können. Approximiert man diese durch Finite Differenzen, so erhält man komplett matrixfreie Verfahren.

Übersicht Konvergenzraten:

a) *Exaktes Newton Verfahren* ist Q-*quadratisch* konvergent.

b) *Quasi-Newton-Verfahren* ist Q-*superlinear* konvergent.

c) *Gauss-Newton, Levenberg-Marquardt* und *Gradientenabstiegsverfahren* sind Q-*linear* konvergent.