6 Numerische Optimierung

In den ersten fünf Kapiteln dieses Skriptes haben wir Grundaufgaben der biomedizinischen Bildgebung eingeführt, im Sinne von Variationsmethoden modelliert und ihre Analyse in geeigneten Funktionenräumen diskutiert. Letztendlich ist aber das Ziel, die Methoden auf konkrete Daten der Biomedizin anzuwenden. Da Variationsmethoden auf Minimierungsaufgaben für kontinuierliche Funktionen beruhen, führt dies einerseits zu der Frage nach geeigneten Verfahren zur Diskretisierung und anderseits zu der Frage nach passenden numerischen Optimierungsverfahren. In diesen Themengebieten existiert eine sehr große Anzahl an Originalarbeiten, die sich entweder mit relativ spezifischen Variationsmethoden zur Bildverarbeitung oder andererseits mit sehr abstrakten Konzepten der Optimierung beschäftigen.

In diesem Kapitel geht es darum, die Frage der Diskretisierung von unterschiedlichen Seiten zu beleuchten und Werkzeuge der numerischen Optimierung zur Verfügung zu stellen, die die Lösung von vorgestellten Variationsproblemen ermöglichen. Wir starten mit dem sehr grundlegenden Sachverhalt der Diskretisierung unendlich dimensionaler Optimierungsprobleme.

6.1 Diskretisierung

Die Lösung der Variationsprobleme aus den letzten Kapiteln ist ein Problem der Variationsrechnung, bei der man sich mit Optimierungsproblemen der Form

$$J(u) = \int_\Omega G(x, (Ku)(x), u(x), \nabla u(x))\, dx \;\rightarrow\; \min_u \tag{6.1}$$

beschäftigt.

Es gibt zwei unterschiedliche Vorgehensweisen zur Lösung von unendlich dimensionalen Optimierungsproblemen. Zum einen das Prinzip, "First discretize, then optimize" und zum anderen das Prinzip "First optimize, then discretize". Beide Strategien werden in der Forschung eingesetzt und es besteht weiterhin fortlaufend eine Diskussion über ihre jeweiligen Vor- und Nachteile.

(a) **First discretize, then optimize**: Die Idee dieses Ansatzes ist die unmittelbare Diskretisierung eines gegebenen Optimierungsproblems, d.h. die Ersetzung aller auftretenden Funktionenräume durch endlich dimensionale Räume, sowie der Ersetzung aller auftretenden Operatoren durch geeignete diskrete Pendanten. D.h. anstelle von (6.1) für einen Funktionenraum $\mathcal{U}$ betrachtet man

$$\min_{u_h \in U_h} J_h(u_h)$$

mit $J_h : U_h \to \mathbb{R}$, $U_h \subset \mathcal{U}$ und einem Diskretisierungsparameter h. Dies führt im Allgemeinen zu einem Problem der (diskreten) *nichtlinearen* Optimierung (engl.: nonlinear programming) in $\mathbb{R}^n$. Der Hauptvorteil besteht darin, dass man eine Vielzahl an existierenden, effizienten Methoden der *nichtlinearen* Optimierung (z.B. basierend auf *Innere Punkte Verfahren* oder Sequentielle Quadratische Programmierung (SQP) Verfahren) einsetzen kann. Ein Nachteil ist der Mangel an quantitativen Approximationsresultaten für nichtilineare Probleme.

(b) **First optimize, then discretize**: Die Idee dieses Ansatzes ist die Formulierung der Optimierungsmethode in unendliche dimensionalen Räumen und einer anschließenden Diskretisierung lediglich für die Lösung von (linearen oder quadratischen) Teilproblemen und für die Auswertung des Zielfunktionals. Anders ausgedrückt, man geht erst über zum Optimalitätssystem (Karush-Kuhn-Tucker System für beschränkte Optimierungsprobleme) und transferiert dann alle auftretenden Funktionenräume und Operatoren für eine diskrete algorithmische Umsetzung. Der Hauptvorteil dieser Strategie besteht darin, dass quantitative Abschätzungen für die Konvergenz von Optimierungsmethoden mit Fehlerabschätzungen für die Diskretisierung von Teilproblemen kombiniert werden können. Damit kann man Abschätzungen für den gesamten Fehler einer numerischen Optimierungsmethode ableiten.

Bis heute gibt es kein allgemeines Rezept, welcher der beiden Diskretisierungsstrategien vorzuziehen ist. Vielmehr hängt es von der Anwendung und den Ressourcen zum wissenschaftlichen Rechnen ab. Wichtig ist allerdings, dass der gewählte numerische Ansatz die Struktur des unendlich dimensionalen Optimierungsproblems zu einem gewissen Grad widerspiegelt und erhält.

Darüber hinaus kann es für beschränkte Optimierungsprobleme auch Sinn machen, Diskretisierungskonzepte nicht direkt auf das Ausgangsproblem wie in (a) oder direkt auf das Optimalitätssystem wie in (b) anzuwenden, sondern zunächst einen SQP-Ansatz auf der kontinuierlichen Ebene anzusetzen, um dann wie oben beschrieben fortzufahren.

Im Folgenden starten wir mit numerischen Methoden der beschränkten Optimierung und konzentrieren uns auf Verfahren, die dem zweiten Ansatz genügen, d.h. wir formulieren Optimierungsmethoden im unendlich dimensionalen Fall. Man beachte aber, dass die Resultate auch im Fall $\mathcal{U} = \mathbb{R}^n$ angewandt werden können.

6.2 Gradientenverfahren

Im Folgenden nehmen wir an, dass $\mathcal{U}$ ein Hilbertraum sei, falls nicht anders festgelegt. Zur Herleitung eines sehr einfachen numerischen Optimierungsverfahrens für (unbeschränkte) Optimierungsprobleme der Form (6.1) betrachten wir das Beispiel eines Regularisierungsfunktionals bzgl. $\mathcal{U} := H^1(\Omega) = W^{1,2}(\Omega)$:

$$J(u) := \frac{1}{2} \int_\Omega |\nabla u|^2 \, dx \ .$$

Der sogenannte Gradientenfluss ist definiert als $\frac{\partial u}{\partial t} = -J'(u)$ und ist in diesem Fall durch die Wärmeleitungsgleichung

$$\frac{\partial u}{\partial t} = \Delta u$$

gegeben. Um ein allgemeines Optimierungsverfahren zu erhalten, können wir einen Gradientenfluss in einem Hilbertraum $\mathcal{U}$ als

$$\frac{\partial u}{\partial t} = -J'(u) \tag{6.2}$$

einführen, wobei $J'(u) \in \mathcal{U}$ eine Element des Hilbertraums darstellt, das mit dem Gradienten von J an der Stelle u identifiziert werden kann. Mit anderen Worten, wir definieren die Evolution dieser Gleichung durch

$$\left\langle \frac{\partial u}{\partial t}, v \right\rangle = -J'(u)v \qquad \forall v \in \mathcal{U} \ ,$$

wobei $\langle \cdot \, , \, \cdot \rangle$ das Skalarprodukt in $\mathcal{U}$ bezeichnet. Die Evolution des Gradientenflusses impliziert eine Evolution des Zielfunktionals, die gegeben ist durch:

$$\frac{\partial}{\partial t} \left(J(u) \right) = J'(u) \frac{\partial u}{\partial t} = - \left\| \frac{\partial u}{\partial t} \right\|^2 \leq 0 \ .$$

Dies bedeutet, dass das Zielfunktional monoton fallend ist, und $\frac{\partial}{\partial t}(J(u)) = 0$ gilt, genau dann wenn $\frac{\partial u}{\partial_t} = 0$. Darüber hinaus impliziert die Struktur des Gradientenflusses (6.2), dass $\frac{\partial u}{\partial_t} = 0$ gilt, genau dann wenn $J'(u) = 0$, d.h. u ist ein stationärer (man sagt auch: kritischer) Punkt, bzw. erfüllt die notwendige Optimalitätsbedingung erster Ordnung. Folglich können wir erwarten, dass das Zielfunktional mit dem Gradientenfluss abfällt bis schließlich ein stationärer Punkt erreicht wird.

Um ein iteratives Optimierungsverfahren abzuleiten, diskretisieren wir das Optimalitätssystem in (6.2) mit Hilfe einer *expliziten Zeitdiskretisierung* des Flusses, d.h.

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

mit $\sigma_k > 0$ als Schrittweite bzgl. einer (künstlichen) Iterations-Zeit und einer sogenannten Suchrichtung $d_k(u_k) := -J'(u_k)$. Dieses Verfahren bezeichnet man als *Gradientenverfahren*. Aus numerischer Sicht ist es offensichtlich, dass nur eine hinreichend kleine Wahl der Zeitschritte σ_k sinvoll ist, da explizite Zeitdiskretisierungen mit zu großen Schritten nicht stabil sind.

Das Gradientenverfahren, auch *Verfahren des steilsten Abstiegs* genannt, wurde bereits 1847 von Cauchy untersucht. Wie wir gesehen haben, bestimmt man bei diesem Verfahren im Punkt u_k diejenige Suchrichtung d_k, in der J am stärksten abnimmt. Man spricht von einer streng gradientenbezogenen Suchrichtung.

Die lokal optimale Sichtweise beim (projizierten) Gradientenverfahren muss global nicht die beste Vorgehensweise sein. Eine ungünstige Wahl der Schrittweite kann dazu führen, dass man keine globale Konvergenz erhält. Wir betrachten dazu das folgende Beispiel.

Beispiel 6.2.1. *Es sei $J(u) = u^2$ und $u_0 = 1$. Weiter sei*

$$d_k = -1 \quad \textit{und} \quad \sigma_k = \left(\frac{1}{2}\right)^{k+2} \qquad \forall k \geq 0 \; . \tag{6.3}$$

Dann ist

$$u_{k+1} = u_k - \sigma_k = u_0 - \sum_{i=0}^{k} \left(\frac{1}{2}\right)^{i+1} = \frac{1}{2} + \left(\frac{1}{2}\right)^{k+1} \; .$$

Also gilt $u_{k+1} < u_k$ und daher $J(u_{k+1}) < J(u_k)$ für alle $k \geq 0$, aber $u_k \to \frac{1}{2}$, d.h. (u_k) konvergiert nicht gegen das Minimum $u = 0$ von J.

Ist das Minimum eines Funktionals gesucht, so ist es bei gegebenem u_k naheliegend, bei der Berechnung von u_{k+1} das Ziel

$$J(u_{k+1}) < J(u_k) \tag{6.4}$$

anzustreben. Verfahren, die eine solche Strategie realisieren, nennt man *Abstiegsverfahren*. Auch wenn das Verfahren unter Umständen nicht gegen ein (lokales) Minimum konvergiert, so wird doch in jeder Iteration das Zielfunktional verkleinert und damit ein besserer Punkt berechnet, was in der Praxis oft schon zufriedenstellend ist.

Ein Abstiegsverfahren benutzt zur Berechnung von u_{k+1} eine Abstiegsrichtung, d.h. eine Suchrichtung mit der Eigenschaft

$$J(u_k + \sigma d_k) < J(u_k), \qquad \forall \sigma \in]0, s_k[$$

mit einem $s_k > 0$. Zur Konstruktion von Abstiegsverfahren gibt es zwei prinzipielle Vorgehensweisen:

(a) Verfahren mit Schrittweitensteuerung: Hier bestimmt man zunächst eine Abstiegsrichtung d_k aufgrund lokaler Informationen über die Zielfunktion im aktuellen Iterationspunkt u_k. Dann berechnet man eine Schrittweite $\sigma_k \in]0, s_k[$, mit der man einen möglichst großen Abstieg erzielt, und setzt $u_{k+1} = u_k + \sigma_k d_k$.

(b) Trust-Region-Verfahren: Hier wird basierend auf einem lokalen Modell des Zielfunktionals (beispielsweise einer quadratischen Approximation des Zielfunktionals) eine Trust-Region (Vertrauensbereich) berechnet, auf der das lokale Modell des Zielfunktionals hinreichend gut approximiert wird. Das lokale Modell erlaubt dann die Berechnung einer Abstiegsrichtung d_k, und man setzt $u_{k+1} = u_k + d_k$.

Im Folgenden konzentrieren wir uns auf Schrittweitenverfahren, um z.B. das Konvergenzverhalten des Gradientenverfahrens zu verbessern. Auf Trust-Region-Verfahren werden wir später im Zusammenhang von Levenberg-Marquardt nochmal eingehen.

6.3 Schrittweitenverfahren

Die fundamentale Idee der Verfahren in diesem Kapitel ist folgende:

(i) An einer Stelle u bestimmt man eine Suchrichtung d, bei der die Funktionalwerte reduziert werden (Abstiegsverfahren).

(ii) Beginnend bei u, bewegt man sich in Richtung d so weit, wie die Funktionalwerte von J *hinreichend reduziert* werden. (Schrittweiten Steuerung)

Um (global) konvergente Verfahren zu erhalten, müssen wir sogenannte effiziente Schrittweiten bestimmen.

Definition 6.3.1. *Für gegebenes u und gegebene Suchrichtung d mit $J'(u)d < 0$ erfüllt eine Schrittweite das* ***Prinzip des hinreichenden Abstiegs****, falls*

$$J(u + \sigma d) \leq J(u) + c_1 \sigma \, J'(u)d \tag{6.5}$$

und

$$\sigma \geq -c_2 \frac{J'(u)d}{\|d\|^2} \tag{6.6}$$

mit einer von u und d unabhängigen Konstanten $c_1, c_2 > 0$ gilt. Eine Schrittweite σ heißt ***effizient****, falls*

$$J(u + \sigma d) \leq J(u) - c \left(\frac{J'(u)d}{\|d\|} \right)^2$$

mit einer von u und d unabhängigen Konstanten $c = c_1 \, c_2 > 0$ gilt.

Erfüllt eine Schrittweite das Prinzip des hinreichenden Abstiegs, dann ist sie effizient.

Wir betrachten die Situation von Beispiel (6.3). Die Wahl der Schrittweitenfolge in diesem Beispiel erfüllt nicht das Prinzip des hinreichenden Abstiegs, da nach Ungleichung (6.6) die Bedingung

$$\sigma_k \geq 2c_2 u_k$$

mit einer von u unabhängigen Konstanten c_2 gelten müsste.

Eine naheliegende Schrittweitenstrategie besteht darin, die Schrittweite σ durch Lösung eines eindimensionalen Optimierungsproblems

$$\min_{\sigma \geq 0} \phi(\sigma) \; = \; J(u + \sigma d)$$

zu berechnen. Man bezeichnet die Lösung dieses Problems als *exakte Schrittweite*. Allerdings ist zu beachten, dass nur unter zusätzlichen Voraussetzungen(z.B. Konvexität von J) sichergestellt werden kann, dass σ globale Lösung des Problems ist. In der Praxis geht man deshalb sinnvollerweise zu nicht exakten Schrittweitenverfahren, die aber dennoch effiziente Schrittweiten liefern.

Im Folgenden werden wir Verfahren zur Berechnung solcher Schrittweiten betrachten, man spricht von *line-search* Verfahren. Da wir nur an der Minimierung von Zielfunktionalen interessiert sind, aber nicht an der exakten Approximation der Lösung eines Gradientenflusses, basiert eine Schrittweitensteuerung lediglich auf dem Ziel einen hinreichenden Abstieg des Zielfunktionals zu finden.

Ein klassischer Ansatz dafür sind die sogenannten *Armijo-Goldstein Regeln*. Die Armijo-Regel ist ein wichtiges Element für alternierende Schrittweitenverfahren (wie Armijo-Goldstein), die das Prinzip des hinreichenden Abstiegs erfüllen.

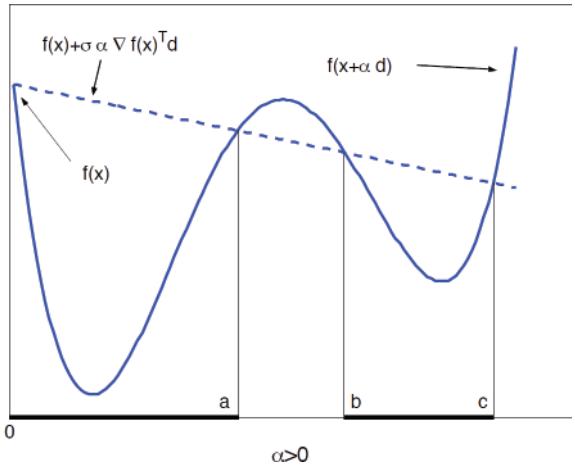

Figure 6.1: Illustration der Schrittweitensuche von Armijo

Definition 6.3.2 (Armijo Bedingung)**.** *Sei J ein Zielfunktional, u_k eine aktuelle Iterierte, d_k eine Suchrichtung und $c_1 \in (0,1)$ eine kleine Konstante. Dann führt eine geeignete Wahl der Schrittweite σ zu einem hinreichenden Abstieg des Zielfunktionals, d.h.*

$$J(u_k + \sigma d_k) \leq J(u_k) + c_1 \sigma \, J'(u_k) d_k \ . \tag{6.7}$$

Die praktische Umsetzung bei der Schrittweitensuche ist in der Regel folgende: Man startet mit einer Startschrittweite $\sigma_0 > 0$ und überprüft nach gewissen Bedingungen (hier Armijo), ob die Schrittweite zu einem hinreichenden Abstieg im Zielfunktional führt. Ist dies nicht der Fall, so wird die Schrittweite mit einem Parameter $\tau \in (0,1)$ solange verkleinert

$$\sigma_{k+1} = \tau \sigma_k \ ,$$

bis ein hinreichender Abstieg erreicht wird. Man spricht dabei von *backtracking line-search*. Allerdings allein die Armijo-Bedingung mittels *backtracking* zu verwenden, ist nicht ausreichend um einen hinreichenden Fortschritt der Minimierung zu garantieren, da die Bedingung unter Umständen schon für hinreichend kleine Werte der σ_k erfüllt sein kann.

Stattdessen testet man zusätzlich zur Armijo Bedingung noch auf eine darauf folgende Bedingung:

$$J(u_k + \sigma d_k) \geq J(u_k) + c_2 \sigma \, J'(u_k) d_k \ , \tag{6.8}$$

mit $0 < c_1 < c_2 < 1$. Die Schrittweitenstrategie (6.7) zusammen mit (6.8) bezeichnet man als *Armijo-Goldstein Regeln*. Die beiden Regeln sind gleichbedeutend mit dem Vergleich von *effektivem Abstieg*

$$D_{eff}(\sigma) := J(u_k + \sigma d_k) - J(u_k)$$

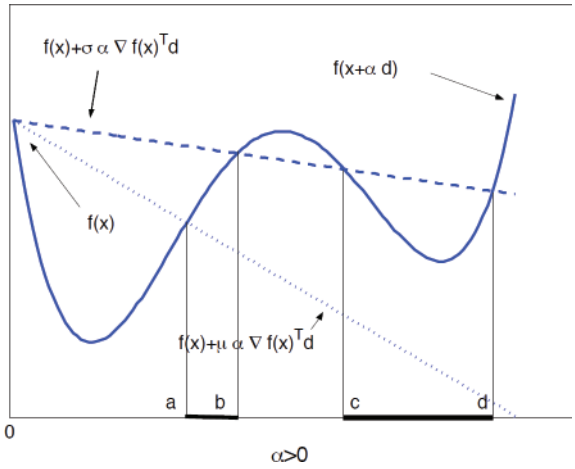

Figure 6.2: Illustration der Schrittweitensuche von Armijo-Goldstein

und *erwartetem Abstieg*

$$D_{exp}(\sigma) := \sigma J'(u_k)d_k \ .$$

Für hinreichend kleines σ stehen sie über die Taylorformel in Beziehung zueinander

$$D_{eff}(\sigma) = D_{exp}(\sigma) + o(\sigma) \ .$$

Man testet also, ob

$$c_2 D_{exp}(\sigma) \leq D_{eff}(\sigma) \leq c_1 D_{exp}(\sigma) \tag{6.9}$$

mit Konstanten $0 < c_1 < c_2 < 1$ erfüllt ist. Letztendlich akzeptieren wir eine Schrittweite falls (6.9) oder äquivalent (6.7) zusammen mit (6.8) erfüllt sind. Eine typische Wahl für die Konstanten sind

$$c_1 \approx 0.1, \quad c_2 \approx 0.9 \ .$$

Unter Verwendung der Armijo-Goldstein Regeln, kann man globale Konvergenz des Gradientenverfahrens beweisen, d.h. Konvergenz zu einem stationären Punkt für jeden beliebigen Startwert.

Theorem 6.3.3. *Sei J zweimal stetig Frechet-differenzierbar und schwach unterhalbstetig auf einem Hilbertraum $\mathcal{U}$. Weiterhin seien die Mengen*

$$\{u \in \mathcal{U} | J(u) \leq M\}$$

für jedes $M \in \mathbb{R}$ beschränkt in $\mathcal{U}$ und leer für M hinreichend klein. Dann hat die Folge (u_k) aus dem Gradientenverfahren mit Armijo-Goldstein Schrittweitensuche eine schwach konvergente Teilfolge, deren Grenzwert ein stationärer Punkt ist.

Proof. Da das Gradientenverfahren ein Abstiegsverfahren ist, gilt

$$J(u_k) \leq J(u_0)$$

für alle $k \geq 0$. Das bedeutet, dass die Folge (u_k) beschränkt ist und daher eine schwach konvergente Teilfolge (u_{k_l}) mit Grenzwert $\bar{u}$ existiert. Mit der ersten Bedingung aus der Schrittweitensuche nach Armijo-Goldstein erhalten wir damit folgende Abschätzung

$$\begin{aligned}
\sum_{k=0}^{N} \|u_{k+1} - u_k\|^2 &= -\sum_{k=0}^{N} \sigma_k \, J'(u_k)(u_{k+1} - u_k) \\
&\leq \frac{1}{c_1} \sum_{k=0}^{N} (J(u_k) - J(u_{k+1})) \\
&= \frac{1}{c_1} (J(u_0) - J(u_{N+1})) \\
&\leq \frac{1}{c_1} \Big(J(u_0) - \inf_u J(u) \Big) =: p \, .
\end{aligned}$$

Da p unabhängig von N ist, erhalten wir für $N \to \infty$

$$\sum_{l=0}^{\infty} \|u_{k_l+1} - u_{k_l}\|^2 \leq \sum_{k=0}^{\infty} \|u_{k+1} - u_k\|^2 \leq p \, .$$

Daher existiert eine Teilfolge von (u_{k_l}), ohne Beschränkung der Allgemeinheit sei diese (u_{k_l}) selbst, mit

$$\|\sigma_{k_l} \, J'(u_{k_l})\| = \|u_{k_l+1} - u_{k_l}\| \to 0 \, .$$

Da J zweimal unterhalbstetig Frechet-differenzierbar ist, existiert eine Konstante $C < 0$ mit

$$J''(u_{k_l})(v, v) \leq C \, \|v\|^2 \, , \quad \forall v \in \mathcal{U} \, .$$

Damit impliziert die zweite Bedingung in Armijo-Goldstein

$$\begin{aligned}
c_2 \, \sigma_{k_l} J'(u_{k_l}) &\leq J(u_{k_l}) - J(u_{k_l}) \\
&\quad J'(u_{k_l})(u_{k_l+1} - u_{k_l}) + \frac{C}{2} \|u_{k_l+1} - u_{k_l}\|^2 \, .
\end{aligned}$$

Setzt man $u_{k_l+1} - u_{k_l} = -\sigma_{k_l} J'(u_{k_l})$ ein, so erhalten wir

$$(1 - c_2)\sigma_{k_l} \|J'(u_{k_l})\|^2 \leq \frac{C}{2} \sigma_{k_l}^2 \|J'(u_{k_l})\|^2 \, .$$

Damit gilt entweder $J'(u_{k_l}) = 0$ oder

$$\sigma_{k_l} \geq \frac{2(1 - \alpha)}{c} \, .$$

Falls $J'(u_k) = 0$ gilt, so hat der Algorithmus einen stationären Punkt erreicht und stoppt, d.h. $u_j = u_{k_l}$ für alle $j \geq k_l$, und die Konvergenz ist trivial. Im zweiten Fall ist σ_{k_l} gleichmäßig nach unten von Null weg beschränkt und deshalb gilt $\|J'(u_{k_l})\| \to 0$. Dies impliziert, dass $J'(\bar{u}) = 0$, d.h. der Grenzwert $\bar{u}$ ist ein stationärer Punkt. □

Betrachtet man für die Schrittweitensuche zusätzlich neben der Armijo-Schrittweite noch eine Krümmungseigenschaft, die garantiert, dass die Steigung von J hinreichend stark reduziert wird, so spricht man von den Wolfe-Bedingungen. Es wird auch Verfahren von Powell genannt, da er es bei der Untersuchung der globalen Konvergenz von Quasi-Newton-Verfahren benutzt hat, das wir in einem späteren Abschnitt noch kennenlernen werden.

Die vorgestellten Schrittweitenverfahren sind selbstverständlich nicht auf das Gradientenverfahren beschränkt, sondern sie können grundsätzlich bei Abstiegsverfahren eingesetzt werden.