Nichtglatte Optimierung

Michael Ulbrich

Technische Universität München

Juni 2016

Inhaltsverzeichnis

2.7 Das ε-Subdifferential

Ein großer Nachteil der Richtungsableitung und des Subdifferentials besteht darin, dass man anhand von $f'(x,\cdot)$ bzw. $\partial f(x)$ nicht erkennen kann, ob sich x in der Nähe eines Minimums von f befindet (betrachte etwa $f(x) = |x|$). Nur, wenn x bereits das Minimum ist, erkennen wir dies anhand von $f'(x,\cdot)$ bzw. $\partial f(x)$.

Das Problem besteht darin, dass $f'(x,\cdot)$ und $\partial f(x)$ keine Umgebungsinformation enthalten.

Wir führen daher nun ein Subdifferential ein, das diese Schwierigkeiten überwindet.

Definition 2.7.1 (ε-Subgradient, ε-Subdifferential)**.** Sei $f : \mathbb{R}^n \to \mathbb{R}$ konvex und $\varepsilon \geq 0$. Der Vektor $g \in \mathbb{R}^n$ heißt ε-*Subgradient* von f im Punkt $x \in \mathbb{R}^n$, wenn gilt:

$$f(y) - f(x) \geq g^T(y - x) - \varepsilon \quad \forall\, y \in \mathbb{R}^n. \tag{2.21}$$

Die Menge $\partial_\varepsilon f(x) \subset \mathbb{R}^n$,

$$\partial_\varepsilon f(x) = \{g \in \mathbb{R}^n\,;\ g \text{ ist } \varepsilon\text{-Subgradient von } f \text{ in } x\}$$

heißt ε-*Subdifferential* von f im Punkt $x \in \mathbb{R}^n$. Das ε-Subdifferential induziert eine mengenwertige Abbildung $\partial_\varepsilon f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$.

Anschauliche Interpretation:

Der Vektor g ist ε-Subgradient von f in $\bar{x}$, falls die durch den Punkt $(\bar{x}, f(\bar{x}) - \varepsilon)$ verlaufende lineare Funktion l mit Gradient g, also

$$l(x) = f(\bar{x}) + g^T(x - \bar{x}) - \varepsilon$$

auf oder unterhalb des Graphen von f verläuft.

Aufgrund der Definition ist folgendes klar:

$$\partial f(x) = \partial_0 f(x) \subset \partial_\varepsilon f(x) \quad \forall \varepsilon \geq 0,\ x \in \mathbb{R}^n.$$

Wir können auch eine entsprechende Richtungsableitung definieren:

Definition 2.7.2 (ε-Richtungsableitung)**.** Sei $f : \mathbb{R}^n \to \mathbb{R}$ konvex und $\varepsilon \geq 0$. Die ε-*Richtungsableitung* von f im Punkt x in Richtung $s \in \mathbb{R}^n$ ist definiert gemäß

$$f'_\varepsilon(x,s) = \inf_{t>0} \frac{f(x+ts) - f(x) + \varepsilon}{t}.$$

Anschauliche Interpretation:

Betrachten wir f ausgehend vom Punkt x entlang der Richtung s, so ergibt sich die Funktion $\phi : \mathbb{R}_+ \to \mathbb{R}$, $\phi(t) = f(x + ts)$, $t \geq 0$. Der Wert $a = f'_\varepsilon(x,s)$ ist nun die Steigung jener Geraden $g : \mathbb{R}_+ \to \mathbb{R}$, $g(t) = f(x) + at - \varepsilon$, die durch den Punkt $(0, f(x) - \varepsilon)$ verläuft und den Graphen der Funktion ϕ von unten berührt, wobei diese Interrpetation nur zutrifft, wenn das Infimum in der Definition von $f'_\varepsilon(x,s)$ in einem $t = t^*$ angenommen wird. Der Berührpunkt ist dann $((x + t^*s)^T, f(x + t^*s))^T$.

Die folgende Darstellung ist kanonisch und folgt im wesentlichen [GK02]. Wir stellen zunächst einige Zusammenhänge zwischen f' und f'_ε her:

Satz 2.7.3. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ konvex. Dann gilt für alle $x, s \in \mathbb{R}^n$:*

a) $f'(x,s) = f'_0(x,s)$.

b) $f'(x,s) \leq f'_\varepsilon(x,s)$.

Beweis:

zu a): Folgt sofort aus Satz 2.4.1 b).

zu b): Für $t > 0$ gilt

$$\frac{f(x+ts) - f(x)}{t} \leq \frac{f(x+ts) - f(x) + \varepsilon}{t}$$

Bilden von $\inf\limits_{t>0}$ liefert nun die Behauptung. □

Für $\partial_\varepsilon f$ und $f'_\varepsilon(\cdot,\cdot)$ gelten ganz ähnliche Aussagen wie für ∂f und $f'(\cdot,\cdot)$.

Wir formulieren nun eine ε-Entsprechung zu Satz 2.4.3:

Satz 2.7.4. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ konvex und $x \in \mathbb{R}^n$. Dann gilt:*

a) $\partial_\varepsilon f(x) = \left\{ g \in \mathbb{R}^n \,;\, g^T s \leq f'_\varepsilon(x,s) \;\; \forall\, s \in \mathbb{R}^n \right\}$.

b) *$\partial_\varepsilon f(x)$ ist nichtleer, konvex und kompakt.*

c) $f'_\varepsilon(x,s) = \max\limits_{g \in \partial_\varepsilon f(x)} g^T s \quad \forall\, s \in \mathbb{R}^n$.

Beweis: Erfolgt in ganz ähnlicher Weise wie der Nachweis der entsprechenden Aussagen in Satz 2.4.3.

Satz 2.7.5. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ eine konvexe Funktion, $x^* \in \mathbb{R}^n$ und $\varepsilon \geq 0$. Dann sind die folgenden Aussagen äquivalent:*

a) *Der Punkt x^* ist ε-optimal, d.h.*

$$f(x^*) \leq f(x) + \varepsilon \quad \forall\, x \in \mathbb{R}^n.$$

b) *Es gilt* $\quad f'_\varepsilon(x^*, s) \geq 0 \quad \forall\, s \in \mathbb{R}^n$.

c) *Es gilt* $\quad 0 \in \partial_\varepsilon f(x^*)$.

Beweis:

a) $\Longrightarrow$ b):

Für beliebiges $s \in \mathbb{R}^n$ folgt $f(x^* + ts) + \varepsilon \geq f(x^*)$ für $t > 0$ und somit

$$f'_\varepsilon(x^*, s) = \inf_{t>0} \frac{f(x^* + ts) - f(x^*) + \varepsilon}{t} \geq 0.$$

b) $\Longrightarrow$ c):

Wegen

$$0^T s = 0 \leq f'_\varepsilon(x^*, s) \quad \forall\, s \in \mathbb{R}^n$$

folgt $0 \in \partial_\varepsilon f(x^*)$ gemäß Satz 2.7.4 a).

c) $\Longrightarrow$ a):

Da 0 ein ε-Subgradient ist, folgt nach Definition

$$f(x) - f(x^*) \geq 0^T(x - x^*) - \varepsilon = -\varepsilon \quad \forall\, x \in \mathbb{R}^n.$$

Somit ist x^* ε-optimal. □

Zu Beginn hatten wir versprochen, dass für $\varepsilon > 0$ das ε-Subdifferential $\partial_\varepsilon f(x)$ Informationen aus der Umgebung von x enthält. Dies wird nun präzisiert:

Satz 2.7.6. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ konvex, $x \in \mathbb{R}^n$. Dann gilt:*

a) *Zu $\varepsilon > 0$ gibt es $\delta > 0$, so dass gilt:*

$$\bigcup_{y \in B_\delta(x)} \partial f(y) \subset \partial_\varepsilon f(x).$$

b) *Zu $\delta > 0$ gibt es $\varepsilon > 0$, so dass gilt:*

$$\bigcup_{y \in B_\delta(x)} \partial f(y) \subset \partial_\varepsilon f(x).$$

Beweis:

Fall a): Die Funktion f ist nach Satz 2.4.1 a) auf einer offenen Umgebung U von x Lipschitz-stetig mit Konstante $L > 0$. Wähle nun $\delta > 0$ mit $B_\delta(x) \subset U$ und $2L\delta \leq \varepsilon$. Der Rest des Beweises wird gemeinsam mit b) geführt.

Fall b): Die Funktion f ist nach Satz 2.4.1 a) und Lemma 2.5.7 Lipschitz-stetig auf $\bar{B}_\delta(x)$ mit Konstante $L > 0$. Wähle nun $\varepsilon \geq 2L\delta$.

Fall a) und b) gemeinsam:

Für alle $y \in B_\delta(x)$ und alle $g \in \partial f(y)$ gilt $\|g\| \leq L$ nach Satz 2.4.3 b) und daher ergibt sich für alle $z \in \mathbb{R}^n$:

$$\begin{aligned} g^T(z - x) &= g^T(z - y) + g^T(y - x) \leq f(z) - f(y) + \|g\|\|y - x\| \\ &\leq f(z) - f(x) + |f(x) - f(y)| + \|g\|\|y - x\| \\ &\leq f(z) - f(x) + L\|x - y\| + L\|y - x\| \\ &\leq f(z) - f(x) + 2L\delta \leq f(z) - f(x) + \varepsilon. \end{aligned}$$

Dies zeigt $g \in \partial_\varepsilon f(x)$. □

Im Gegensatz zu f' können wir mit f'_ε robuste Abstiegsrichtungen berechnen. Genauer gilt:

Satz 2.7.7. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ konvex, $\varepsilon \geq 0$ und $x \in \mathbb{R}^n$ mit $0 \notin \partial_\varepsilon f(x)$. Weiter sei $s = -g$ mit $g = P_{\partial_\varepsilon f(x)}(0)$. Dann ist $\dfrac{s}{\|s\|}$ Lösung des Problems*

$$\min_{\|d\|=1} f'_\varepsilon(x, d) \tag{2.22}$$

und es gilt

$$f'_\varepsilon(x, s) = -\|s\|^2 = -\|g\|^2 < 0 \tag{2.23}$$

Ist weiter d eine Richtung mit $f'_\varepsilon(x, d) < 0$ (z.B. obige Richtung s), dann gilt

$$\inf_{t>0} f(x + ts) < f(x) - \varepsilon. \tag{2.24}$$

Beweis: Für $d \in \mathbb{R}^n$ mit $\|d\| = 1$ gilt $v^T d \geq -\|v\|$ für alle $v \in \mathbb{R}^n$ und somit

$$\begin{aligned} f'_\varepsilon(x,d) &= \max_{v \in \partial_\varepsilon f(x)} v^T d \geq \max_{v \in \partial_\varepsilon f(x)} -\|v\| = -\min_{v \in \partial_\varepsilon f(x)} \|v\| \\ &= -\|P_{\partial_\varepsilon f(x)}(0)\| = -\|g\| = -\|s\|. \end{aligned}$$

Weiter ergibt sich für alle $v \in \partial_\varepsilon f(x)$ wegen der Eigenschaft der Projektion

$$g^T(v-g) = (P_{\partial_\varepsilon f(x)}(0) - 0)^T (v - P_{\partial_\varepsilon f(x)}(0)) \geq 0 \quad \forall\, v \in \partial_\varepsilon f(x).$$

Daraus folgt

$$\min_{v \in \partial_\varepsilon f(x)} v^T g = \|g\|^2.$$

Dies zeigt

$$\begin{aligned} f'_\varepsilon\left(x, \frac{s}{\|s\|}\right) &= \frac{1}{\|g\|} f'_\varepsilon(x,-g) = \frac{1}{\|g\|} \max_{v \in \partial_\varepsilon f(x)} v^T(-g) = -\frac{1}{\|g\|} \min_{v \in \partial_\varepsilon f(x)} v^T g \\ &= -\frac{\|g\|^2}{\|g\|} = -\|g\| = -\|s\|. \end{aligned}$$

Damit ist $s/\|s\|$ Lösung von (2.22) und es gilt

$$f'_\varepsilon(x,s) = \|s\| f'_\varepsilon\left(x, \frac{s}{\|s\|}\right) = -\|s\|^2 < 0.$$

Ist nun d eine Richtung mit $f'_\varepsilon(x,d) < 0$, dann folgt

$$\inf_{t>0} \frac{f(x+td) - f(x) + \varepsilon}{t} = f'_\varepsilon(x,d) < 0.$$

Daher gibt es $t^* > 0$ mit

$$\frac{f(x+t^*d) - f(x) + \varepsilon}{t^*} < 0.$$

Daraus erhalten wir

$$\inf_{t>0} f(x+td) \leq f(x+t^*d) < f(x) - \varepsilon.$$

□

Wir können nun in Analogie zur Methode des steilsten Abstiegs das folgende Verfahren betrachten:

Algorithmus 2.7.8 (Modellalgorithmus)**.**

0. Wähle $x^0 \in \mathbb{R}^n$ und $\varepsilon > 0$.

Für $k = 0, 1, 2, \ldots$:

1. Bestimme $g^k = P_{\partial_\varepsilon f(x^k)}(0)$.
2. Falls $g^k = 0$, STOP.
3. Setze $s^k = -g^k$ und ermittle die optimale Schrittweite $\sigma_k \geq 0$ entlang s^k:

$$f(x^k + \sigma_k s^k) = \min_{\sigma \geq 0} f(x^k + \sigma s^k).$$

4. Setze $x^{k+1} = x^k + \sigma_k s^k$.

Bemerkung. Es würde genügen, wenn die Schrittweite σ_k so gewählt wird, dass sie einen Teil des maximal möglichen Abstiegs realisiert. Zudem könnte man auch mit anderen Richtungen, die $f'_\varepsilon(x^k, s_k) < 0$ erfüllen, arbeiten.

Dieses Verfahren hat sehr schöne Konvergenzeigenschaften. Es gilt nämlich:

Satz 2.7.9. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ konvex und nach unten beschränkt. Weiter sei $\varepsilon > 0$. Dann terminiert Algorithmus* 2.7.8 *nach endlich vielen Iterationen mit einem ε-optimalen Punkt von f:*

$$f(x^k) \leq \inf_{x \in \mathbb{R}^n} f(x) + \varepsilon.$$

Beweis: Solange $g^k \neq 0$ ist, gilt $0 \notin \partial_\varepsilon f(x^k)$ und somit ergibt sich nach Satz 2.7.7:

$$f(x^{k+1}) < f(x^k) - \varepsilon.$$

Irgendwann gilt dann $f(x^k) \leq \inf_x f(x) + \varepsilon$ und dann muss spätestens $g^k = 0$ gelten, da eine f-Abnahme um mehr als ε nicht mehr möglich ist. Umgekehrt folgt aus $g^k = 0$, dass x^k ε-optimal ist, siehe Satz 2.7.5. □

Da $\partial_\varepsilon f(x)$ in der Praxis schwer oder gar nicht zu berechnen ist, hat Algorithmus 2.7.8 nur theoretischen Wert. Er legt aber nahe, $\partial_\varepsilon f(x^k)$ geeignet durch eine Menge G^k_ε zu approximieren und dann $s^k = -P_{G^\varepsilon_k}(0)$ als Suchrichtung zu verwenden. Genau dies macht das Bundle-Verfahren, dem wir uns aber zunächst aus dem Blickwinkel der Schnittebenen-Verfahren nähern wollen.

2.8 Bundle Methoden

Bundle (=Bündel) Methoden bilden eine der effizientesten Verfahrensklasse der nichtglatten Optimierung. Die grundlegende Idee besteht darin, wie beim Schnittebenen-Verfahren aus den bisher berechneten Funktionswerten und Subgradienten ein „Bündel“ aus Informationen zu schnüren und dieses zu verwenden, um Suchrichtungen zu berechnen.

Es gibt nun zwei Sichtweisen, die dual zueinander sind:

Die eine interpretiert das Bundle-Verfahren als ein regularisiertes Schnittebenen-Verfahren und verwendet somit das Bündel, um f durch ein stückweise lineares Schnittebenenmodell zu approximieren. Die Schrittberechnung erfolgt dann durch Minimierung dieses Modells, versehen mit einer geeigneten Penalisierung, welche zu langen Schritten entgegenwirkt und das Teilproblem eindeutig lösbar macht.

In der dualen Sichtweise wird das Bündel verwendet, um durch geeignete Konvexkombinationen der Subgradienten eine innere Approximation G^k_ε des ε-Subdifferentials $\partial_{\varepsilon_k} f(x^k)$ zu erzeugen und dieses mittels $s^k = -P_{G^k_\varepsilon}(0)$ im Sinne des Modellalgorithmus 2.7.8 zur Berechnung von Suchrichtungen zu verwenden.

2.8.1 Das Bundle-Verfahren aus Sicht der Schnittebenenmethode

In der k-ten Iteration stehen für ein Schnittebenenmodell aus früheren Iterationen in dem Bündel die folgenden Informationen zur Verfügung:

$$y^j, \quad f_j = f(y^j), \quad g^j \in \partial f(y^j), \quad j \in J_k \subset \{0, \ldots, k\}.$$

Der Punkt x^k ist stets im Bündel vertreten, d.h. es gibt $j \in J_k$ mit $y^j = x^k$.

Das vom Bündel induzierte Schnittebenenmodell lautet nun

$$f_k^{\text{se}}(x) = \max_{j \in J_k} l_j(x) = \max_{j \in J_k} (f_j + {g^j}^T (x - y^j)),$$

wobei l_j die lineare Stützfunktion zum Bündel-Eintrag (y^j, f_j, g^j) ist:

$$l_j(x) = f_j + {g^j}^T (x - y^j), \quad f_j = f(y^j), \quad g^j \in \partial f(y^j).$$

Zur Berechnung eines Schrittes s^k wird nun das folgende Teilproblem gelöst:

$$\min_{s \in \mathbb{R}^n} \; f_k^{\text{se}}(x^k + s) + \frac{1}{2\gamma_k} \|s\|^2, \tag{2.25}$$

wobei $\gamma_k > 0$ geeignet gewählt ist. Der Penalty-Term $\frac{1}{2\gamma_k}\|s\|^2$ sorgt dafür, dass der Schritt s^k nicht zu weit von x^k wegführt. Je kleiner γ_k, desto kürzer wird der Schritt ausfallen. Man kann sogar folgendes zeigen:

Ist s^k Lösung von (2.25), so gibt es $\Delta_k > 0$, so dass s^k das folgende Trust-Region-Schnittebenen-Problem löst:

$$\min_s \; f_k^{\text{se}}(x^k + s) \quad \text{u.d.N.} \quad \|s\| \le \Delta_k. \tag{2.26}$$

Ist umgekehrt s^k eine Lösung von (2.26), in der die Nebenbedingung $\|s\| \le \Delta_k$ stark aktiv ist (d.h. nach Vergrößern von Δ_k wäre s^k nicht mehr optimal), so gibt es $\gamma_k > 0$, so dass s^k das Problem (2.25) löst.

Man kann γ_k geeignet anpassen (siehe [SZ92]), aber für ein konvergentes Verfahren ist das nicht nötig. Wir setzen ab jetzt $\gamma_k = 1$ (wie z.B. auch in [GK02]).

Im folgenden ist es günstig, l_j folgendermaßen umzuschreiben:

$$\begin{aligned} l_j(x) &= f_j + {g^j}^T (x - y^j) = f(x^k) + {g^j}^T (x - x^k) - (f(x^k) - f_j + {g^j}^T (y^j - x^k)) \\ &= f(x^k) + {g^j}^T (x - x^k) - \alpha_j^k \end{aligned} \tag{2.27}$$

mit

$$\alpha_j^k = f(x^k) - f_j - {g^j}^T (x^k - y^j) = f(x^k) - l_j(x^k). \tag{2.28}$$

Der Wert α_j^k ist also die Differenz zwischen $f(x^k)$ und $l_j(x^k)$ und somit immer nichtnegativ. Damit lautet das Schnittebenenmodell

$$f_k^{\text{se}}(x) = \max_{j \in J_k} l_j(x) = f(x^k) + \max_{j \in J_k} ({g^j}^T (x - x^k) - \alpha_j^k) =: f(x^k) + \bar{f}_k^{\text{se}}(x). \tag{2.29}$$

In dem Problem (2.25) lassen wir den konstanten Offset $f(x^k)$ weg und erhalten

$$\min_s \; \bar{f}_k^{\text{se}}(x^k + s) + \frac{1}{2}\|s\|^2. \tag{2.30}$$

Hierbei haben wir wie angekündigt $\gamma_k = 1$ gesetzt.

Gemäß Lemma 2.6.1 können wir dieses Problem als QP schreiben:

$$\min_{s \in \mathbb{R}^n, \xi \in \mathbb{R}} \xi + \frac{1}{2}\|s\|^2 \quad \text{u.d.N.} \quad {g^j}^T s - \alpha_j^k - \xi \le 0, \;\; j \in J_k. \tag{2.31}$$

Das Paar (s^k, ξ_k) ist genau dann Lösung von (2.31), wenn s^k Lösung von (2.30) ist und $\xi_k = \bar{f}_k^{\text{se}}(x^k + s^k)$ gilt.

Ist der Schritt s^k berechnet, so prüfen wir in gewohnter Trust-Region-Manier, ob die tatsächliche Zielfunktionsabnahme

$$\text{ared}_k(s^k) = f(x^k) - f(x^k + s^k) \tag{2.32}$$

hinreichend groß ist im Vergleich zur Modellabnahme (da es stets $j \in J_k$ mit $y^j = x^k$ gibt, gilt $f_k^{\text{se}}(x^k) = f(x^k)$), der wie folgt definiert ist (predicted reduction):

$$\text{pred}_k(s^k) = f_k^{\text{se}}(x^k) - f_k^{\text{se}}(x^k + s^k) = f(x^k) - f_k^{\text{se}}(x^k + s^k) = -\bar{f}_k^{\text{se}}(x^k + s^k) = -\xi_k. \tag{2.33}$$

Hierzu verwenden wir die Bedingung

$$\text{ared}_k(s^k) \geq \eta \text{pred}_k(s^k) = -\eta\xi_k. \tag{2.34}$$

Ist diese erfüllt, so führen wir einen *wesentlichen Schritt* (serious step, dies ist die gängige Terminologie) aus:

$$x^{k+1} = x^k + s^k.$$

Sonst führen wir einen *Nullschritt* aus:

$$x^{k+1} = x^k.$$

In beiden Fällen wird der Punkt $y^{k+1} = x^k + s^k$ in das neue Bündel aufgenommen.

Es ergibt sich das folgende Verfahren:

Algorithmus 2.8.1 (Bundle-Verfahren)**.**

0. Wähle $x^0 \in \mathbb{R}^n$, $\eta \in (0,1)$ und $\varepsilon \geq 0$. Bestimme $g^0 \in \partial f(x^0)$, setze $y^0 = x^0$, $\alpha_0^0 = 0$ und $J_0 = \{0\}$.

Für $k = 0, 1, 2, \ldots$:

1. Berechne ein KKT-Tupel (s^k, ξ_k, λ^k) des Problems (2.31).
2. Berechne $v^k = -s^k$ und $\varepsilon_k = \sum_{j \in J_k} \lambda_j^k \alpha_j^k$.
3. Prüfe auf Abbruch: Falls $\|v^k\| \leq \varepsilon$ und $\varepsilon_k \leq \varepsilon$, STOP.
4. Gilt
$$f(x^k + s^k) - f(x^k) \leq \eta\xi_k,$$
so führe einen *wesentlichen Schritt* durch:
$$y^{k+1} = x^k + s^k, \;\; x^{k+1} = y^{k+1}, \;\; J_{k+1} = \left\{ j \in J_k \,;\; \lambda_j^k > 0 \right\} \cup \{k+1\}.$$
5. Gilt
$$f(x^k + s^k) - f(x^k) > \eta\xi_k,$$
so führe einen *Nullschritt* durch:
$$y^{k+1} = x^k + s^k, \;\; x^{k+1} = x^k, \;\; J_{k+1} = \left\{ j \in J_k \,;\; \lambda_j^k > 0 \text{ oder } y^j = x^k \right\} \cup \{k+1\}.$$

6. Berechne $f_{k+1} = f(y^{k+1})$, $g^{k+1} \in \partial f(y^{k+1})$ und

$$\alpha_j^{k+1} = f(x^{k+1}) - f_j - {g^j}^T(x^{k+1} - y^j), \quad j \in J_{k+1}.$$

Einige Bemerkungen:

- Die Abbruchbedingung wird erst mit Lemma 2.8.5 klar werden. Sie sichert ein gewisse Form von ε-Optimalität zu.
- Die Indexmenge J_k wird so aktualisiert, dass neben den Schnittebenen zu $y^{k+1} = x^k + s^k$ und zu x^k nur die im Punkt $x^k + s^k$ stark aktiven Schnittebenen im Bündel bleiben. Andere Varianten sind möglich.

2.8.2 Eine duale Sichtweise des Bundle-Verfahrens

Wir leiten nun zu dem Teilproblem (2.31) ein duales Problem her.

Lemma 2.8.2. a) (s^k, ξ_k) *ist genau dann Lösung von* (2.31)*, wenn die KKT-Bedingungen gelten, d.h. wenn es* $\lambda_j^k \in \mathbb{R}$, $j \in J_k$, *gibt mit*

$$\begin{aligned} s^k + \sum_{j \in J_k} \lambda_j^k g^j &= 0, \\ \sum_{j \in J_k} \lambda_j^k &= 1, \\ {g^j}^T s^k - \alpha_j^k - \xi_k \le 0, \;\; \lambda_j^k \ge 0, \;\; \lambda_j^k({g^j}^T s^k - \alpha_j^k - \xi_k) &= 0, \;\; j \in J_k. \end{aligned} \tag{2.35}$$

b) *Der Vektor* λ^k *ist genau dann Lösung des Problems*

$$\begin{aligned} \min_{\lambda} \quad & \frac{1}{2} \left\| \sum_{j \in J_k} \lambda_j g^j \right\|^2 + \sum_{j \in J_k} \lambda_j \alpha_j^k \\ \text{u.d.N.} \quad & \sum_{j \in J_k} \lambda_j = 1, \;\; \lambda_j \ge 0, \;\; j \in J_k, \end{aligned} \tag{2.36}$$

wenn die KKT-Bedingungen gelten. Weiter ist $(\lambda^k, \mu^k, \xi_k) \in \mathbb{R}^{|J_k|} \times \mathbb{R}^{|J_k|} \times \mathbb{R}$ *genau dann ein KKT-Tupel von* (2.36)*, wenn* (2.35) *für*

$$s^k = -\sum_{j \in J_k} \lambda_j^k g^j, \tag{2.37}$$

erfüllt ist und zusätzlich gilt:

$$\mu_j^k = -{g^j}^T s^k + \alpha_j^k + \xi_k, \;\; j \in J_k. \tag{2.38}$$

gilt.

c) *Sei* λ^k *eine Lösung von* (2.36)*. Dann ist* λ^k *auch Lösung des folgenden Problems:*

$$
\begin{aligned}
&\min_{\lambda} \quad \frac{1}{2}\left\|\sum_{j\in J_k}\lambda_j g^j\right\|^2 \\
&\text{u.d.N.} \quad \sum_{j\in J_k}\lambda_j = 1,\ \ \lambda_j \geq 0,\ \ j\in J_k,\quad \sum_{j\in J_k}\lambda_j\alpha_j^k \leq \varepsilon_k
\end{aligned}
\tag{2.39}
$$

mit

$$
\varepsilon_k = \sum_{j\in J_k}\lambda_j^k\alpha_j^k. \tag{2.40}
$$

Beweis: zu a):

Für das konvexe QP (2.31) sind die KKT-Bedingungen notwendig und hinreichend. Diese KKT-Bedingungen sind in a) angegeben.

zu b):

Die KKT-Bedingungen sind für das konvexe QP (2.36) notwendig und hinreichend und lauten:

$$
\begin{aligned}
g^{j^T}\sum_{i\in J_k}\lambda_i^k g^i + \alpha_j^k + \xi_k - \mu_j^k = 0,\ \ j\in J_k \\
\sum_{j\in J_k}\lambda_j^k = 1, \\
\lambda_j^k \geq 0,\ \ \mu_j^k \geq 0,\ \ \mu_j^k\lambda_j^k = 0,\ \ j\in J_k.
\end{aligned}
\tag{2.41}
$$

Gelte nun (2.35). Die erste Zeile in (2.35) liefert dann (2.37). Definieren wir nun μ_j^k gemäß (2.38), und setzen (2.37) ein, so ergibt sich die erste Gleichung in (2.41). Umschreiben der dritten Gleichung von (2.35) auf μ_j^k ergibt die dritte Gleichung in (2.41). Die zweite Gleichung in (2.35) und in (2.41) sind identisch.

Gelte nun umgekehrt (2.41). Definieren wir s^k gemäß (2.37), so folgt die erste Gleichung in (2.35). Einsetzen von (2.37) in die erste Gleichung von (2.41) ergibt (2.38) und Einsetzen von (2.38) in die dritte Zeile von (2.41) liefert die dritte Zeile von (2.35). Die zweiten Zeilen sind wiederum identisch.

zu c):

Für das konvexe QP in (2.39) sind die im folgenden aufgeschriebenen KKT-Bedingungen notwendig und hinreichend:

$$
\begin{aligned}
g^{j^T}\sum_{i\in J_k}\lambda_i^k g^i + \xi_k - \mu_j^k + \tau^k\alpha_j^k = 0,\ \ j\in J_k \\
\sum_{j\in J_k}\lambda_j^k = 1, \\
\lambda_j^k \geq 0,\ \ \mu_j^k \geq 0,\ \ \mu_j^k\lambda_j^k = 0,\ \ j\in J_k, \\
\sum_{j\in J_k}\lambda_j^k\alpha_j^k \leq \varepsilon_k,\ \ \tau^k \geq 0,\ \ \tau^k\left(\sum_{j\in J_k}\lambda_j^k\alpha_j^k - \varepsilon_k\right) = 0.
\end{aligned}
\tag{2.42}
$$

Ist nun λ^k Lösung von (2.36), so gelten die Bedingungen (2.41). Definieren wir ε_k gemäß (2.40) und setzen wir $\tau^k = 1$, so folgt unmittelbar (2.42). □

Wie soeben gezeigt, sind also die Probleme (2.31) und (2.36) äquivalent. Aus einer Lösung von (2.36) läßt sich s^k aus (2.37) zurückgewinnen. Der Wert ξ_k kann ebenfalls direkt berechnet werden. Hierzu benutzen wir die Komplementaritätsbedingung in (2.35), (2.37) und (2.40):

$$\begin{aligned}\xi_k &= \xi_k \sum_{j\in J_k} \lambda_j^k = \sum_{j\in J_k} \lambda_j^k \xi_k = \sum_{j\in J_k} \lambda_j^k ({g^j}^T s^k - \alpha_j^k) \\ &= \left(\sum_{j\in J_k} \lambda_j^k g^j\right)^T s^k - \sum_{j\in J_k} \lambda_j^k \alpha_j^k = -\|s^k\|^2 - \varepsilon_k.\end{aligned}$$

Damit ergeben sich folgende Formeln:

$$s^k = -v^k, \;\; \xi_k = -\|s^k\|^2 - \varepsilon_k = -\|v^k\|^2 - \varepsilon_k \tag{2.43}$$

$$\text{mit} \quad v^k = \sum_{j\in J_k} \lambda_j^k g^j, \quad \varepsilon_k = \sum_{j\in J_k} \lambda_j^k \alpha_j^k. \tag{2.44}$$

Wir kommen nun zu einem wichtigen Punkt: Gemäß Lemma 2.8.2 ist λ^k Lösung von (2.39) und daher gilt

$$v^k = P_{G_{\varepsilon_k}^k}(0),$$

wobei

$$G_\varepsilon^k = \left\{ \sum_{j\in J_k} \lambda_j g^j \,;\, \sum_{j\in J_k} \lambda_j \alpha_j^k \le \varepsilon, \;\; \sum_{j\in J_k} \lambda_j = 1, \;\; \lambda_j \ge 0, \;\; j \in J_k \right\} \tag{2.45}$$

Ein wesentlicher Schritt s^k des Bundle-Verfahrens, d.h.

$$x^{k+1} - x^k = s^k = -v^k = -P_{G_{\varepsilon_k}^k}(0)$$

erinnert also stark an den Schritt $x^{k+1} - x^k = -\sigma_k P_{\partial_\varepsilon f(x^k)}(0)$ im Modellalgorithmus 2.7.8. Dieser Eindruck wird nun erhärtet, indem wir zeigen, dass G_ε^k eine Teilmenge von $\partial_\varepsilon f(x^k)$ ist.

Lemma 2.8.3. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ konvex. Weiter seien $x^k \in \mathbb{R}^n$, $y^j \in \mathbb{R}^n$, $g^j \in \partial f(y^j)$, $j \in J_k$, $\varepsilon \ge 0$ und G_ε^k definiert gemäß* (2.45)*. Dann gilt*

$$G_\varepsilon^k \subset \partial_\varepsilon f(x^k).$$

Beweis:

Wir zeigen zunächst $g^j \in \partial_{\alpha_j^k} f(x^k)$:

Nach Definition von α_j^k und wegen $g^j \in \partial f(y^j)$ gilt für alle $x \in \mathbb{R}^n$

$$\begin{aligned}f(x) - f(x^k) &= f(x) - f(y^j) + f(y^j) - f(x^k) \ge {g^j}^T (x - y^j) + f(y^j) - f(x^k) \\ &= {g^j}^T (x - x^k) + {g^j}^T (x^k - y^j) + f(y^j) - f(x^k) = {g^j}^T (x - x^k) - \alpha_j^k,\end{aligned}$$

also $g^j \in \partial_{\alpha_j^k} f(x^k)$.

Sei nun $\lambda_j \ge 0$, $\sum_{j\in J_k} \lambda_j = 1$. Multiplizieren von

$$f(x) - f(x^k) \ge {g^j}^T (x - x^k) - \alpha_j^k$$

mit λ_j und Aufaddieren liefert

$$f(x) - f(x^k) \geq \left(\sum_{j \in J_k} \lambda_j g^j \right)^T (x - x^k) - \sum_{j \in J_k} \lambda_j \alpha_j^k \quad \forall\, x \in \mathbb{R}^n.$$

Damit ist

$$\sum_{j \in J_k} \lambda_j g^j \in \partial_{(\sum_{j \in J_k} \lambda_j \alpha_j^k)} f(x^k) \overset{\sum_{j \in J_k} \lambda_j \alpha_j^k \leq \varepsilon}{\subset} \partial_\varepsilon f(x^k)$$

gezeigt. □

Wir können daher Algorithmus 2.8.1 auch so formulieren:

Algorithmus 2.8.4 (Bundle-Verfahren 2.8.1 in dualer Formulierung)**.**

0. Wähle $x^0 \in \mathbb{R}^n$, $\eta \in (0,1)$ und $\varepsilon \geq 0$. Bestimme $g^0 \in \partial f(x^0)$, setze $y^0 = x^0$, $\alpha_0^0 = 0$ und $J_0 = \{0\}$.

Für $k = 0, 1, 2, \ldots$:

1. Berechne λ^k durch Lösen des Problems (2.36).
2. Berechne
$$v^k = \sum_{j \in J_k} \lambda_j^k g^j, \quad s^k = -v^k, \quad \varepsilon_k = \sum_{j \in J_k} \lambda_j^k \alpha_j^k, \quad \xi_k = -\|v^k\|^2 - \varepsilon_k.$$
3. Prüfe auf Abbruch: Falls $\|v^k\| \leq \varepsilon$ und $\varepsilon_k \leq \varepsilon$, STOP.
4. Gilt
$$f(x^k + s^k) - f(x^k) \leq \eta \xi_k,$$
so führe einen *wesentlichen Schritt* durch:
$$y^{k+1} = x^k + s^k, \;\; x^{k+1} = y^{k+1}, \;\; J_{k+1} = \left\{ j \in J_k \,;\; \lambda_j^k > 0 \right\} \cup \{k+1\}.$$
5. Gilt
$$f(x^k + s^k) - f(x^k) > \eta \xi_k,$$
so führe einen *Nullschritt* durch:
$$y^{k+1} = x^k + s^k, \;\; x^{k+1} = x^k, \;\; J_{k+1} = \left\{ j \in J_k \,;\; \lambda_j^k > 0 \text{ oder } y^j = x^k \right\} \cup \{k+1\}.$$
6. Berechne $f_{k+1} = f(y^{k+1})$, $g^{k+1} \in \partial f(y^{k+1})$ und
$$\alpha_j^{k+1} = f(x^{k+1}) - f_j - {g^j}^T (x^{k+1} - y^j), \quad j \in J_{k+1}.$$

Wir sehen uns nun die Abbruchbedingung näher an.

Lemma 2.8.5. *In Algorithmus* 2.8.1 (*bzw. Algorithmus* 2.8.4) *sei für* $\varepsilon > 0$ *die Abbruchbedingung*

$$\|v^k\| \leq \varepsilon, \quad \varepsilon_k \leq \varepsilon.$$

erfüllt. Dann ist x_k ε*-optimal im folgenden Sinne:*

$$f(x^k) \leq f(x) + \varepsilon \|x - x^k\| + \varepsilon \quad \forall\, x \in \mathbb{R}^n.$$

Beweis: Für alle $j \in J_k$ und $x \in \mathbb{R}^n$ gilt nach (2.27)

$$g^{j^T}(x - x^k) = l_j(x) - f(x^k) + \alpha_j^k \leq f(x) - f(x^k) + \alpha_j^k.$$

Multiplizieren mit λ_j^k und Summieren ergibt:

$$\begin{aligned} v^{k^T}(x - x^k) &= \sum_{j \in J_k} \lambda_j^k g^{j^T}(x - x^k) \leq \sum_{j \in J_k} \lambda_j^k \big(f(x) - f(x^k) + \alpha_j^k\big) \\ &= f(x) - f(x^k) + \sum_{j \in J_k} \lambda_j^k \alpha_j^k = f(x) - f(x^k) + \varepsilon_k. \end{aligned}$$

Dies ergibt mit der Cauchy-Schwarz-Ungleichung

$$f(x_k) \leq f(x) - v^{k^T}(x - x^k) + \varepsilon_k \leq f(x) + \|v^k\| \|x - x^k\| + \varepsilon_k \leq f(x) + \varepsilon \|x - x_k\| + \varepsilon.$$

□

2.8.3 Globale Konvergenz

Wir weisen nun die globale Konvergenz des Verfahrens nach. Dies ist relativ aufwendig.

Die Iterationen, in denen wesentliche Schritte erfolgen, werden in der Menge $\mathcal{K}$ zusammengefasst:

$$\mathcal{K} = \{k \geq 0 \,;\, t_k = 1\},$$

wobei $t_k = 1$, falls $x^k \to x^{k+1}$ ein wesentlicher Schritt ist und $t_k = 0$, sonst.

Wir beginnen mit einem technischen Resultat:

Lemma 2.8.6. *Sei $f : \mathbb{R}^n \to \mathbb{R}$ konvex. Sei $\varepsilon = 0$ und die Folgen (x^k), (s^k) usw. seien von Algorithmus* 2.8.1 *erzeugt (insbesondere terminiere das Verfahren also nicht endlich). Weiter sei die Folge $(f(x^k))$ durch $f^* \in \mathbb{R}$ nach unten beschränkt. Dann gilt:*

a) $\lim_{k \to \infty} f(x^{k+1}) - f(x^k) = 0.$

b) $\displaystyle\sum_{k=0}^{\infty} t_k(\|v^k\|^2 + \varepsilon_k) \leq \frac{f(x^0) - f^*}{\eta}.$

c) *Erzeugt der Algorithmus unendlich viele wesentliche Schritte, d.h. gilt $|\mathcal{K}| = \infty$, so folgt*

$$\lim_{\mathcal{K} \ni k \to \infty} \|v^k\| = 0, \quad \lim_{\mathcal{K} \ni k \to \infty} \varepsilon_k = 0.$$

Beweis:

zu a):

Die Folge $(f(x^k))$ ist wegen der Bedingung für wesentliche Schritte in Schritt 4 monoton fallend und durch $f^* \in \mathbb{R}$ nach unten beschränkt. Daher folgt a) unmittelbar.

zu b):

Für alle wesentlichen Schritte gilt $t_k = 1$ und

$$f(x^{k+1}) - f(x^k) \leq \eta t_k \xi_k$$

Für alle Nullschritte ist dies wegen $x^{k+1} = x^k$ und $t_k = 0$ ebenfalls erfüllt. Damit haben wir

$$f(x^0) - f^* \geq f(x^0) - \lim_{k\to\infty} f(x^k) = \sum_{k=0}^{\infty}(f(x^k) - f(x^{k+1})) \geq -\eta \sum_{k=0}^{\infty} t_k \xi_k$$
$$= \eta \sum_{k=0}^{\infty} t_k(\|v^k\|^2 + \varepsilon_k) \geq \eta \sum_{k=0}^{\infty} t_k(\gamma^- \|v^k\|^2 + \varepsilon_k),$$

wobei wir (2.43) benutzt haben.

zu c):

Nach b) gilt

$$\sum_{k\in\mathcal{K}}^{\infty}(\gamma^- \|v^k\|^2 + \varepsilon_k) = \sum_{k=0}^{\infty} t_k(\gamma^- \|v^k\|^2 + \varepsilon_k) \leq \frac{f(x^0) - f^*}{\eta} < \infty.$$

daher ist $(\gamma^- \|v^k\|^2 + \varepsilon_k)_\mathcal{K}$ eine Nullfolge und somit (wegen $\gamma^- > 0$ und $\varepsilon_k \geq 0$) auch $(\|v^k\|)_\mathcal{K}$ und $(\varepsilon_k)_\mathcal{K}$.

□

Wir zeigen nun ein erstes Konvergenzresultat für den Fall, dass unendlich viele wesentliche Schritte durchgeführt werden:

Lemma 2.8.7. *Algorithmus* 2.8.1 *mit* $\varepsilon = 0$ *erzeuge unendlich viele wesentliche Schritte. Dann ist jeder Häufungspunkt von* (x^k) *ein (globales) Minimum von* f.

Beweis:

Sei x^* ein Häufungspunkt von (x^k). Die Folge $(f(x^k))$ ist monoton fallend und hat, da f stetig ist, $f(x^*)$ als Häufungspunkt. Daraus folgt $f(x^k) \downarrow f(x^*) =: f^* \in \mathbb{R}$. Wegen $|\mathcal{K}| = \infty$ gilt

$$\left\{x^k\,;\; k \geq 0\right\} = \{x^0\} \cup \left\{x^{k+1}\,;\; k \in \mathcal{K}\right\} = \left\{x^k\,;\; k \in \mathcal{K}\right\}$$

und daher ist dann x^* auch ein Häufungspunkt von $(x^k)_\mathcal{K}$. Es gibt somit eine Teilfolge $(x^k)_{\mathcal{K}'}$, $\mathcal{K}' \subset \mathcal{K}$, mit

$$(x^k)_{\mathcal{K}'} \to x^*.$$

Gemäß Lemma 2.8.3 gilt $v^k \in \partial_{\varepsilon_k} f(x^k)$ und aus Lemma 2.8.6 c) folgt

$$(\|v^k\|)_\mathcal{K} \to 0, \quad (\varepsilon_k)_\mathcal{K} \to 0.$$

Weiter haben wir für alle $x \in \mathbb{R}^n$

$$f(x) \geq f(x^k) + v^{k^T}(x - x^k) - \varepsilon_k \overset{\mathcal{K}'\ni k\to\infty}{\longrightarrow} f(x^*) + 0^T(x - x^*) + 0 = f(x^*).$$

Damit ist x^* globales Minimum von f.

□

Wir untersuchen nun den Fall, dass nur endlich viele wesentliche Schritte auftreten. Hierzu treffen wir, um uns das Leben etwas zu erleichtern, die folgende Annahme:

(2.46)
$$\exists\, m > 0 :\ \forall\, k \geq m :$$
$$\{k-m, k-m+1, \ldots, k-1\} \cap \mathcal{K} = \emptyset \implies \gamma_k = \gamma^-.$$

In Worten: Nach einer Serie von m Nullschritten gilt stets $\gamma_k = \gamma^-$.

Lemma 2.8.8. *Algorithmus* 2.8.1 *mit* $\varepsilon = 0$ *erzeuge eine unendliche Folge* (x^k) *und es gelte* (2.46)*. Werden nur endlich viele wesentliche Schritte durchgeführt, d.h. gibt es* $l \geq 0$ *mit* $x^k = x^l$ *für alle* $k \geq l$*, so ist* x^l *globales Minimum von* f*.*

Beweis: Sei

$$J_k^+ = \left\{ j \in J_k \,;\; \lambda_j^k > 0 \right\}.$$

Wegen $x^k = x^l$ für alle $k \geq l$ gilt

$$\alpha_j^{k+1} = \alpha_j^k \quad \forall\, j \in J_k^+,\; k \geq l.$$

Ohne Einschränkung können wir wegen (2.46) annehmen, dass gilt:

$$\gamma_k = \gamma^- \quad \forall\, k \geq l.$$

Wir setzen nun

$$\theta_k := \sum_{j \in J_k^+} \lambda_j^k \alpha_j^{k+1} = \sum_{j \in J_k^+} \lambda_j^k \alpha_j^k = \varepsilon_k \quad \forall\, k \geq l. \tag{2.47}$$

Bezeichne weiter

$$Q_k(\lambda) = \frac{\gamma_k}{2} \left\| \sum_{j \in J_k} \lambda_j g^j \right\|^2 + \sum_{j \in J_k} \lambda_j \alpha_j^k$$

die Zielfunktion in (2.36).

Im folgenden sei $k > l$ beliebig.

Wir wählen nun zu beliebigem $\mu \in [0,1]$ den Vektor $\lambda^{k\mu}$ in folgender Weise:

$$\lambda_k^{k\mu} = \mu, \quad \lambda_j^{k\mu} = (1-\mu)\lambda_j^{k-1},\;\; j \in J_{k-1}^+, \quad \lambda_j^{k\mu} = 0,\;\; j \in J_k \setminus (J_{k-1}^+ \cup \{k\}).$$

Hierbei sei λ^{k-1} der durch Lösen des $(k-1)$-ten Teilproblems erhaltenene Vektor. Offensichtlich gilt dann

$$\lambda^{k\mu} \geq 0, \quad \sum_{j \in J_k} \lambda_j^{k\mu} = 1.$$

Somit ist $\lambda^{k\mu}$ zulässig für (2.36) und daraus folgt

$$Q_k(\lambda^k) \leq Q_k(\lambda^{k\mu}).$$

Zur Vereinfachung von $Q_k(\lambda^{k\mu})$ berechnen wir:

$$\sum_{j \in J_k} \lambda_j^{k\mu} g^j = \mu g^k + (1-\mu) \sum_{j \in J_{k-1}^+} \lambda_j^{k-1} g^j = \mu g^k + (1-\mu) \sum_{j \in J_{k-1}} \lambda_j^{k-1} g^j = \mu g^k + (1-\mu) v^{k-1}.$$

Ebenso ergibt sich

$$\sum_{j \in J_k} \lambda_j^{k\mu} \alpha_j^k = \mu \alpha_k^k + (1-\mu) \sum_{j \in J_{k-1}^+} \lambda_j^{k-1} \alpha_j^k = \mu \alpha_k^k + (1-\mu)\theta_{k-1}$$

mit θ_{k-1} wie in (2.47) definiert.

Damit erhalten wir

$$Q_k(\lambda^{k\mu}) = \frac{\gamma^-}{2}\|\mu g^k + (1-\mu)v^{k-1}\|^2 + \mu\alpha_k^k + (1-\mu)\theta_{k-1} = q_k(\mu)$$

mit

$$q_k(\mu) = \frac{\gamma^-}{2}\|\mu g^k + (1-\mu)v^{k-1}\|^2 + \mu\alpha_k^k + (1-\mu)\theta_{k-1}.$$

Bezeichne nun μ^k das Minimum von q_k auf $[0,1]$. Weiter sei $\nu_k = q_k(\mu_k)$. Dann ergibt sich

$$\begin{aligned}\nu_k = q_k(\mu_k) \le q_k(0) &= \frac{\gamma^-}{2}\|v^{k-1}\|^2 + \theta_{k-1} = \frac{\gamma^-}{2}\|v^{k-1}\|^2 + \varepsilon_{k-1} \\ &= Q_{k-1}(\lambda^{k-1}) \le Q_{k-1}(\lambda^{k-1,\mu_{k-1}}) = q_{k-1}(\mu_{k-1}) = \nu_{k-1}.\end{aligned} \tag{2.48}$$

Da q_k quadratisch ist, ergibt sich

$$q_k(\mu) = q_k(0) + \mu q_k'(0) + \frac{\mu^2}{2}q_k''(0) \le \nu_{k-1} + \mu q_k'(0) + \frac{\mu^2}{2}q_k''(0).$$

Wir schätzen nun $q_k'(0)$ ab:

$$q_k'(0) = \gamma^-(g^k - v^{k-1})^T v^{k-1} + \alpha_k^k - \theta_{k-1} = -\gamma^-\|v^{k-1}\|^2 + \gamma^- {g^k}^T v^{k-1} + \alpha_k^k - \theta_{k-1}.$$

Wir benutzen $x^k = x^{k-1}$, $y^k = x^{k-1} + s^k$, $\text{ared}_{k-1}(s^{k-1}) = f(x^{k-1}) - f^k < -\eta\xi_{k-1}$ (sonst wäre $k-1 \in \mathcal{K}$) sowie (2.43) und erhalten

$$\begin{aligned}\alpha_k^k &= f(x^k) - f_k - {g^k}^T(x^k - y^k) = f(x^k) - f_k - {g^k}^T(x^{k-1} - y^k) \\ &= f(x^{k-1}) - f_k + {g^k}^T s^{k-1} < -\eta\xi_{k-1} + {g^k}^T s^{k-1} = \eta\big(\varepsilon_{k-1} + \gamma^-\|v^{k-1}\|^2\big) + {g^k}^T s^{k-1} \\ &= \eta\big(\theta_{k-1} + \gamma^-\|v^{k-1}\|^2\big) - \gamma^- {g^k}^T v^{k-1}.\end{aligned}$$

Daraus folgt

$$q_k'(0) = -\gamma^-\|v^{k-1}\|^2 + \gamma^- {g^k}^T v^{k-1} + \alpha_k^k - \theta_{k-1} \le -(1-\eta)\big(\theta_{k-1} + \gamma^-\|v^{k-1}\|^2\big).$$

Als nächstes schätzen wir $q_k''(0)$ ab. Zunächst gilt

$$\gamma^-\|v^{k-1}\|^2 \le \gamma^-\|v^{k-1}\|^2 + 2\theta_{k-1} \le 2\nu_{k-1} \le 2\nu_l.$$

Daraus wiederum folgt mit (2.48)

$$\begin{aligned}\|y^k\| = \|x^{k-1} + s^{k-1}\| &\le \|x^{k-1}\| + \|s^{k-1}\| = \|x^{k-1}\| + \gamma_{k-1}\|v^{k-1}\| \\ &= \|x^l\| + \gamma^-\|v^{k-1}\| \le \|x^l\| + \sqrt{2\gamma^-\nu_l}.\end{aligned}$$

Somit ist die Folge $(y^k)_{k>l}$ beschränkt und daher auch die Folge $(g^k)_{k>l}$, siehe Satz 2.4.3 und Lemma 2.5.7. Insgesamt gibt es daher eine Konstante $C > 0$ mit

$$\|v^{k-1}\| \le C, \quad \|g^k\| \le C \quad \forall\, k > l.$$

Nun folgt

$$q_k''(0) = \gamma^-\|g^k - v^{k-1}\|^2 \le \gamma^-(\|g^k\| + \|v^{k-1}\|)^2 \le 4\gamma^- C^2 \quad \forall\, k > l,$$

Für alle $\mu \geq 0$ ergibt sich

$$q_k(\mu) \leq \nu_{k-1} + \mu q_k'(0) + \frac{\mu^2}{2} q_k''(0) \leq \nu_{k-1} + \mu q_k'(0) + 2\mu^2 \gamma^- C^2 =: \bar{q}_k(\mu).$$

Bezeichne $\bar{\mu}_k$ das unrestringierte globale Minimum von $\bar{q}_k$. Dann gilt

$$\bar{\mu}_k = \frac{-q_k'(0)}{4\gamma^- C^2}$$

Im Fall $\bar{\mu}_k > 1$ erhalten wir

$$\nu_k = q_k(\mu_k) \leq q_k(1) \leq \bar{q}_k(1) = \nu_{k-1} + q_k'(0) + 2\gamma^- C^2 < \nu_{k-1} + q_k'(0) - \frac{q_k'(0)}{2} = \nu_{k-1} + \frac{q_k'(0)}{2}.$$

Im Fall $\bar{\mu}_k \leq 1$ ergibt sich

$$\nu_k = q_k(\mu_k) \leq q_k(\bar{\mu}_k) \leq \bar{q}_k(\bar{\mu}_k) = \nu_{k-1} + \bar{\mu}_k q_k'(0) + 2\bar{\mu}_k^2 \gamma^- C^2 = \nu_{k-1} - \frac{q_k'(0)^2}{8\gamma^- C^2}.$$

Die nichtnegative, monoton fallende Folge (ν_k) ist konvergent und daher eine Cauchy-Folge. Insbesondere ist $(\nu_{k-1} - \nu_k)$ eine Nullfolge. Wegen

$$\nu_{k-1} - \nu_k \geq \min\left\{ \frac{-q_k'(0)}{2}, \frac{q_k'(0)^2}{8\gamma^- C^2} \right\} > 0$$

folgt, dass $(q_k'(0))$ eine Nullfolge ist, und daraus wiederum

$$\lim_{k\to\infty} \left(\theta_{k-1} + \gamma^- \|v^{k-1}\|^2\right) = 0.$$

Daher haben wir wegen $\varepsilon_k = \theta_k$, $k \geq l$:

$$\varepsilon_k \to 0, \quad v^k \to 0, \quad k \to \infty.$$

Wir können nun fortfahren wie im zweiten Teil des Beweises von Lemma 2.8.7, um zu zeigen, dass jeder Häufungspunkt von (x^k) ein globales Minimum von f ist. Die stationäre Folge besitzt genau einen Grenzwert, nämlich x^l. □

Nehmen wir Lemma 2.8.7 und Lemma 2.8.8 zusammen, dann erhalten wir den folgenden Konvergenzsatz:

Satz 2.8.9. *Algorithmus* 2.8.1 *mit* $\varepsilon = 0$ *erzeuge die Folge* (x^k) *und es gelte* (2.46). *Dann ist jeder Häufungspunkt von* (x^k) *ein globales Minimum von* f.

Durch eine kleine Modifikation des Beweises von Lemma 6.79 und Satz 6.80 in [GK02] kann man außerdem noch zeigen:

Satz 2.8.10. *Die Funktion* f *besitze globale Minima und Algorithmus* 2.8.1 *mit* $\varepsilon = 0$ *erzeuge die Folge* (x^k) *und es gelte* (2.46). *Dann konvergiert* (x^k) *gegen ein globales Minimum* x^* *von* f.

Literaturverzeichnis

[BC11] H. H. Bauschke, P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2011.

[Ber99] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.

[Cla83] F. H. Clarke, *Optimization and nonsmooth analysis*, John Wiley & Sons, Inc., New York, 1983.

[Cla98] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern und P. R. Wolenski, *Nonsmooth analysis and control theory*, Graduate Texts in Mathematics, 178, Springer-Verlag, New York, 1998.

[GK99] C. Geiger, C. Kanzow, *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*, Springer-Verlag, 1999.

[GK02] C. Geiger, C. Kanzow, *Theorie und Numerik restringierter Optimierungsaufgaben*, Springer-Verlag, 2002.

[HU93a] J.-B. Hiriart–Urruty und C. Lemaréchal, *Convex analysis and minimization algorithms I*, Grundlehren der Mathematischen Wissenschaften, 305, Springer-Verlag, Berlin, 1993.

[HU93b] J.-B. Hiriart–Urruty und C. Lemaréchal, *Convex analysis and minimization algorithms II*, Grundlehren der Mathematischen Wissenschaften, 306, Springer-Verlag, Berlin, 1993.

[KS86] M. Kojima und S. Shindo, *Extension of Newton and quasi-Newton methods to systems of PC^1 equations*, J. Oper. Res. Soc. Japan 29 (1986), 352–375.

[Lem89] C. Lemaréchal, *Nondifferentiable optimization*, in: G.L. Nemhauser, A. H. G. Rinnooy Kan, M. J. Todd (eds.), Optimization, Handbooks in Operations Research and Management Science, 1, North-Holland Publishing Co., Amsterdam, 1989, 529–572.

[Mif77] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim. 15 (1977) 957–972.

[Ro70] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[RW09] R. T. Rockafellar, R. J.-B. Wets, Variational Analysis, Springer, 2009.

[Qi93] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res. 18 (1993), 227–244.

[QS93] L. Qi und J. Sun, *A nonsmooth version of Newton's method*, Math. Programming 58 (1993), 353–367.

[Sch94] S. Scholtes, *Introduction to piecewise differentiable equations*, Preprint No. 53/1994, Universität Karlsruhe, Institut f. Statistik u. Math. Wirtschaftstheorie, 1994.

[SZ92] H. Schramm und J. Zowe, *A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results*, SIAM J. Optim. 2 (1992) 121–152.