1 Wichtig

- Vektoren oben indiziert, Zahlen unten, Matrizen?
- erklären, dass aufgrund einer Einheitlichen Notation ein wenig vom Paper abgewichen, zeigen wo?

2 Berechnungen und erste Gedanken zur Masterarbeit

2.1 Toughts about linesearch

In paper stated: Most *nonconvex* bundle methods need linesearch. Makes sense → in general often linesearch needed if nonconvex.
Here, because of convexification of the objective, linesearch not needed any more, because function "convex enough".

2.2 Gedanken zu Linearisierungsfehlern e_j

Problem: e_j häufig negativ, selbst in konvexem Fall, wo dies in der Theorie nicht auftritt.
Grund (wahrscheinlich): Rundungsfehler des Computers
Theorie: Das Bundle-Verfahren ist eine „Weiterentwicklung" des Schnittebenen-Verfahrens. Im Schnittebenen-verfahren, wird die zu minimierende Funktion durch die Tangenten in den Iterierten angenähert. Die „Lücke" zwischen den tatsächlichen Funktionswerten und denen der Approximation bei den Iterierten nennt man den Linearisierungsfehler e_j. Dieser ist in der Theorie bei konvexen Funktionen immer positiv. Durch Rundungsfehler im Rechner kann es dazu kommen, dass die Linearisierungsfehler nicht mehr positiv sind. Da Tangenten für die Annäherung der Funktion benutzt werden, führt jede Änderung in der Steigung dazu, dass aus der Tangente eine Sekante wird. Wenn diese Änderung „groß genug" ist, wird der Linearisierungsfehler in einer nahen Iterierten negativ. Bei den Tests ergab sich, dass der Linearisierungsfehler tnedenziell häufiger negativ ist, wenn die Dimension höher ist. Außerdem ist er im eindimensionalen Fall immer $geq0$.
Woran kann das liegen? Mehr Dimensionen = mehr „Kipprichtungen" für den Gradienten, daher e_j häufiger negativ? Abstieg nicht so schnell in höherer Dimension, daher Iterierte näher beieinander? Warum bei 1D gar kein Problem??? → Algorithmus in verschiedenen Dimensionen durchgehen. Sind die Iterierten unterschiedlich weit voneinander entfernt? Sonstige Unterschiede?
1D, Prabel: Iterierte liegen sehr weit auseinander, aber schwer allgemeingültige Aussagen zu treffen, weil Parabel ein sehr einfaches Problem Auch bei $f = x^4$ kein Problem. (Obwohl zB. das Minimum nicht gefunden wird, Funktion wahrscheinlich zu flach.) Hier fällt auf: Größen wie zB. α werden exakt berechnet. Vielleicht in 1D alles exakter (warum???) und deswegen e nie < 0? Edit: nach einer größeren Anzahl Iterationen ist α nicht mehr genau 1, aber noch sehr nah dran. Wahrscheinlich in 1D tatsächlich alles

genauer (zB. weil Subproblem genauer gelöst werden kann?) **Idee:** t könnte etwas damit zu tun haben, wie e_j sich ändert, da es die Zielfuntion konvexifiziert.
!!! t tritt nur im Subproblem auf, hat nichts mit der Annäherung der eigentlichen Funktion zu tun.

2.3 Nonconvex, exact

2.3.1 Berechnungen

Zusammenhang δ, ξ:

$$\delta_{k+1} = f(\hat{x}^k) + \frac{\eta_k}{2}|x^{k+1} - \hat{x}^k|^2 - m_k(x^{k+1})$$

$$\xi_k = m_k(x^{k+1}) - f(\hat{x}^k)$$
$$\Rightarrow \delta_{k+1} = -\xi_k + \frac{\eta_k}{2}|x^{k+1} - \hat{x}^k|^2 = -\xi_k + \frac{\eta_k}{2}|d^k|^2$$

Außerdem gilt (aus den KKT-Bedingungen):

$$\lambda_j \xi_k = \lambda_j (s_j^\top d_k - c_j^k) \quad \forall j \in J$$

$$\Rightarrow \xi_k = \sum_{j \in J} \lambda_j \xi_k = \sum_{j \in J} \lambda_j (s_j^\top d_k - c_j^k) = S^\top d_k - C$$

Umschreiben der beiden δ_{n+1}: Es gilt:

$$g^{-n} + \eta_n \Delta_{-n}^k = \mu_n (\hat{x}^k - x^{n+1})$$

und

$$\varphi_n(x^{n+1} = f(\hat{x}^k) - e_l^k - \eta_n d_l^k + \langle g^l + \eta_n \Delta_l^k, x^{n+1} - \hat{x}^k \rangle$$

Setzt man dies in δ_{n+1} ein, erhält man folgende Gleichungskette:

$$\begin{aligned}
\delta_{n+1} &= f(\hat{x}^k) + \frac{\eta_k}{2}|x^{n+1} - \hat{x}^k|^2 - m_k(x^{n+1}) \\
&= \frac{\eta_k}{2}|x^{n+1} - \hat{x}^k|^2 + e_l^k + \eta_n d_l^k - \langle g^l + \eta_n \Delta_l^k, x^{n+1} - \hat{x}^k \rangle \\
&= \frac{\eta_k}{2}|x^{n+1} - \hat{x}^k|^2 + e_l^k + \eta_n d_l^k - \langle \mu_n (\hat{x}^k - x^{n+1}), x^{n+1} - \hat{x}^k \rangle \\
&= \frac{\eta_k}{2}|x^{n+1} - \hat{x}^k|^2 + e_l^k + \eta_n d_l^k + \mu_n |x^{n+1} - \hat{x}^k|^2 \\
&= \frac{R + \mu_n}{2}|x^{n+1} - \hat{x}^k|^2 + e_l^k + \eta_n d_l^k
\end{aligned}$$

Mit $l = -n$ und der Benennung aus dem Paper.

Mit eigener Benennung:

$$c_j^k = e_k + \frac{\eta}{2}|x_j^{k+1} - \hat{x}^k|^2, \qquad C = \sum \alpha_j c_j^k = E + \frac{\eta}{2}\sum \alpha_j |x^j - \hat{x}^k|^2$$

$$\Rightarrow \delta_{k+1} = \frac{R+\mu_n}{2}|d|^2 + C = \frac{R+\mu_n}{2}|d|^2 + E + \frac{\eta}{2}\sum \alpha_j |x^j - \hat{x}^k|^2$$

Mit obiger Formulierung für ξ_k folgt:

$$\xi = -\mu|d^k|^2 - C$$

2.3.2 Unterschiedliche Formulierungen von Größen

Die Formulierungen sind auf jeden Fall numerisch nicht exakt gleich, wenn in einer ein Term mit α vorkommt und in der anderen nicht, da *alpha* nicht völlig exakt berechnet werden kann und durch das „Aggregieren“ keine Umformung von schon vorhandenen Größen stattfindet, sondern die besagte Größe tatsächlich anders berechnet wird.

δ**:**

$$\delta_{k+1} = -\xi_k + \frac{\eta_k}{2}|d^k|^2 = \frac{R+\mu_n}{2}|d|^2 + C$$

ξ**:**

$$\xi_k = m_k(x^{k+1}) - f(\hat{x}^k) = S^\top d_k - C$$

!hier müssen beide Gleichheitszeichen geprüft werden, da ξ aus der Optimierung von **quadProg** kommt! Erste Gleichheit jedoch blöd zu testen, da Modellfunktion nicht implementiert (blöd zu implementieren wegen max).
!Beachte, dass $S = -\frac{1}{t}d^k$ ersetzt. und somit die zwei Formulierungen von ξ nur gleich sein können, wenn auch diese Gleichung stimmt! Dies gilt jedoch nur, wenn man die Umformung auch verwendet und S durch $-\frac{1}{t}d^k$ ersetzt.

C**:**

$$C = \sum \alpha_i c_i = E + \frac{\eta}{2}\sum \alpha_j |x^j - \hat{x}^k|^2$$

2.4 Testfunktionen aus nonconv-exact-Paper

Im Paper werden 5 Testfunktionen $f_1, ...f_5$ verwendet. Diese sind aus den Funktionen h_i aufgebaut, welche folgendermaßen definiert sind:

$$h_i : \mathbb{R}^n \to \mathbb{R}; \qquad h_i(x) = (ix_i^2 - 2x_i - K) + \sum_{j=1}^{n} x_j$$

Wobei $K \in \mathbb{R}$ eine Konstante ist, die erst einmal $= 0$ gesetzt wird.

Der Gradient von der Funktion h_i sieht folgendermaßen aus:

$$\nabla h_i(x) = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 2ix_i - 2 \\ \vdots \\ 0 \end{pmatrix} \leftarrow i\text{'te Position}$$

2.4.1 Testfunktionen und ihre Ableitungen

Hier werden die Testfunktionen und je ein spezieller Subgradient von ihnen aufgeschrieben.

$$f_1(x) = \sum_{i=1}^{n} |h_i(x)|$$

$$\nabla f_1(x) = \sum_{i=1}^{n} \operatorname{sgn}(h_i(x)) \cdot \left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 2ix_i - 2 \\ \vdots \\ 0 \end{pmatrix} \right) = \sum_{i=1}^{n} \operatorname{sgn}(h_i(x)) \cdot \nabla h_i(x)$$

Wenn ein $h_i(x)$ gleich 0 ist, kann man als Subgradienten den Nullvektor verwenden. Dies steckt bereits implizit in $\operatorname{sgn}(h_i(x))$, da gilt: $\operatorname{sgn}(0) = 0$.

$$f_2(x) = \sum_{i=1}^{n} (h_i(x))^2$$

$$\nabla f_2(x) = \sum_{i=1}^{n} 2h_i(x) \cdot \left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 2ix_i - 2 \\ \vdots \\ 0 \end{pmatrix} \right) = \sum_{i=1}^{n} 2h_i(x) \cdot \nabla h_i(x)$$

$$f_3(x) = \max_{i \in \{1,\dots n\}} |h_i(x)|$$

$$\nabla f_3(x) = \operatorname{sgn}(h_I(x)) \cdot \nabla h_I(x) \qquad \text{mit } I \text{ sodass } h_I(x) = f_5(x)$$

$$f_4(x) = \sum_{i=1}^{n} |h_i(x)| + \frac{1}{2}||x||^2$$
$$\nabla f_4(x) = \nabla f_1 + x$$

$$f_5(x) = \sum_{i=1}^{n} |h_i(x)| + \frac{1}{2}||x||$$
$$\nabla f_5(x) = \nabla f_1 + \frac{1}{2} \cdot \frac{x}{||x||}$$

3 Assumptions (about functions) in papers

3.1 Nonconvex inexact

objective function: proper, regular, locally Lipschitz with full domain. (Citations given) something on lower $\mathcal{C}^1/\mathcal{C}^2$; didn't understand, what exactly.

3.2 Nonconvex, exact

objective function: lower $\mathcal{C}^2$.

4 Convergence proofs

Lemma 5 *Suppose the cardinality of the set* $\{j \in J^k | \alpha_j^k > 0\}$ *is uniformly bounded in* k.
If $E^k \to 0$ *as* $k \to \infty$, *then*
(i) $\sum_{j \in J^k} \alpha_j^k |x^j - \hat{x}^k| \to 0$ *as* $k \to \infty$
If, in addition, for some subset $K \subseteq 1, 2, \ldots,$

$$\hat{x}^k \to \bar{x}, G^k \to \bar{G} \text{ as } K \ni k \to \infty, \text{ with } \{\eta^k | k \in K\} \text{ bounded},$$

then we also have
(ii) $\bar{G} \in \partial f(\bar{x}) + B_{\bar{\theta}}(0)$.
If, in addition, $G^k + \nu^k \to 0$ *as* $K \ni k \to \infty$, *then*
(iii) $\bar{x}$ *satisfies the following approximate stationary condition:*

$$0 \in (\partial f(\bar{x}) + \partial \mathtt{i}_D(\bar{x})) + B_{\bar{\theta}}(0).$$

Finally, if in addition, f is lower-$\mathcal{C}^1$, then
(iv) for each $\varepsilon > 0$ there exists $\rho > 0$ such that

$$f(y) \geq f(\bar{x}) - (\bar{\theta} + \varepsilon)|y - \bar{x}| - 2\bar{\sigma}, \quad \textit{for all } y \in D \cap B_{rho}(\bar{x}).$$

!!! Is the assumption of the uniformly bounded set reasonable?

Teorem 3.2.2, Lemaréchal *Let algorithm generate an infinite sequence of serious steps ($\Leftrightarrow |K| = \infty$).*
(i) If

$$\sum_{k \in K} t_k = +\infty$$

Then $\{x^k\}$ is a minimizing sequence. (ii) If in addition $\{t_k\}$ has and upper bound on K and there is a nonempty set of solutions for the problem, then the whole sequence $\{x^k\}$ converges to one of the solutions.

Proof: (Lemma5) Recall that the first term in the right hand side of $\eta^k \geq \max\left\{\max_{j \in J^k, x^j \neq \hat{x}^k} \frac{-2e_j^k}{|x^j - \hat{x}^k|^2}, 0\right\}$ γ is the minimal value of $\eta \geq$ to imply that

$$e_j^k + \frac{\eta}{2}|x^j - \hat{x}^k|^2 \geq 0$$

for all $j \in J^k$. It is then easily seen that, for such η and for $\eta^k \geq \eta + \gamma$, we have that

$$c_j^k = e_j^k + \frac{\eta^k}{2}|x^j - \hat{x}^k|^2 \geq \frac{\gamma}{2}|x^j + \hat{x}^k|^2.$$

Taking into account that α_j^k and c_j^k are nonnegative, if $E^k \to 0$ then it follows from $E^k = \sum_{j \in J^k} \alpha_j^k c_j^k$ that $\alpha_j^k c_j^k \to 0$ for all $j \in J^k$. Hence,

$$\alpha_j^k c_j^k \geq (\alpha_j^k)^2 c_j^k \geq \frac{\gamma}{2}\left(\alpha_j^k |x^j - \hat{x}^k|\right)^2 \to 0.$$

Thus, $\alpha_j^k |x^j - \hat{x}^k|$ for all $j \in J^k$. As , by the assumption, the sum in the item (i) is over a finite set of indices and each element in the sum tends to zero, the assertion (i) follows. For each j, let p^j be the orthogonal projection of g^j onto the (convex, closed) set $\partial f(x^j)$. It holds that $|g^j - p^j| \leq \theta^j \leq \bar{\theta}$. By

$$d^k = -t^k(G^k + \nu^k), \text{ where } G^k := \sum_{j \in J^k} \alpha_j^k s_j^k, \quad \nu^k \in \partial \mathtt{i}_D(x_{k+1})$$

and

$$s_j^k := g^j + \eta^k\left(x^j - \hat{x}^k\right)$$

we have that

$$
\begin{aligned}
G^k &= \sum_{j\in J^k} \alpha_j^k g^j + \eta \sum_{j\in J^k} \alpha_j^k (x^j - \hat{x}^k) \\
&= \sum_{j\in J^k} \alpha_j^k p^j + \sum_{j\in J^k} \alpha_j^k (g^j - p^j) + \eta^k \sum_{j\in J^k} \alpha (x^j - \hat{x}^k).
\end{aligned}
$$

As the number of the active indices is uniformly bounded in k, by renumbering the indices and filling unused indices with $\alpha_j^k = 0$, we can consider that J^k is some fixed index set (say, $\{1, \ldots, N\}$). Let J be the set of all $j \in J^k$ such that $\liminf \alpha_j^k > 0$. Then item (i) implies that $|x^j - \hat{x}^k| \to 0$. Thus, $|x^j - \bar{x}| \leq |x^j - \hat{x}^k| + |\hat{x}^k - \bar{x}| \to 0$ for $j \notin J$, passing onto a further subsequence in the set K , if necessary, outer semicontinuity of the Clarke subdifferential implies that

$$
\lim_{k\to\infty} \sum_{j\in J^k} \alpha_j^k p^j \in \partial f(\bar{x}).
$$

As $\sum_{j\in J^k} \alpha_j^k (g^j - p^j)$ is clearly in $B_{\bar{\theta}}(0)$, while $\eta^k \sum_{j\in J^k} \alpha (x^j - \hat{x}^k)$ tends to zero by item (i), this shows the assertion (ii).
Item (iii) follows from noting that $(G^k + \nu^k) \to 0$ as $K \ni k \to \infty$ implies that $\{\nu^k\} \to -\bar{G}$. As $\nu^k \in \partial \mathtt{i}_D(\bar{x})$. Adding the latter inclusion and result (ii) gives the desired result.
We finally consider item (iv). Fix any $\varepsilon > 0$. Let $\rho > 0$ be such that

$$
\forall \bar{x} \in \Omega, \forall \varepsilon > 0 \exists \rho > 0 : \forall x \in B_\rho(\bar{x}) \text{ and } g \in \partial f(x) \Rightarrow f(x+u) \geq f(x) + \langle g, u \rangle - \varepsilon |u|
$$

holds for $\bar{x}$. Let $y \in D \cap B_{rho}(\bar{x})$ be arbitrary but fixed. Again, we can consider that J^k is a fixed index set. Let J be the set of $j \in J^k$ for which $|x^j - \hat{x}^k| \to 0$. In particular, it then holds that $x^j \in B_\rho(\bar{x})$. By item (i) , we have that $\{\alpha_j^k \to 0\}$ for $j \notin J$.
Using (3) with the error bounds given in (4), for $j \in J$ we obtain that

$$
\begin{aligned}
f(y) &\geq f^j + \left\langle g^j, y - x^j \right\rangle + \sigma^j + \left\langle p^j - g^j, y - x^j \right\rangle - \varepsilon |y - x^j| \\
&\geq f^j + \left\langle g^j, y - x^j \right\rangle + \sigma^j - (\theta^j + \varepsilon) |y - x^j|.
\end{aligned}
$$

By $0 \leq c_j^k := e_j^k + b_j^k$ and the linearization error definition,

$$
f^j + \left\langle g^j, -x^j \right\rangle = \hat{f}^k - \left\langle g^j, \hat{x}^k \right\rangle + b_j^k - c_j^k.
$$

As a result, it holds that

$$
f(y) \geq \hat{f}^k - c_j^k + b_j^k + \left\langle g^j, y - \hat{x}^k \right\rangle + \sigma^j - (\theta^j + \varepsilon) |y - x^j|.
$$

Since $b_j^k \geq 0$ and $g^j = s_j^k - \eta^k(x^j - \hat{x}^k)$, we obtain that

$$f(y) \geq f(\hat{x}^k) - c_j^k + \left\langle s_j^k, y - \hat{x}^k \right\rangle - \eta^k \left\langle x^j - \hat{x}^k, y - \hat{x}^k \right\rangle + \sigma^j + \hat{\sigma}^k - (\theta^j + \varepsilon)|y - x^j|$$

Taking the convex combination in the latter relation using the simplicial multipliers α_j and using $E^k = \sum_{j \in J_k} \alpha_j^k c_j^k$, gives

$$f(y) \sum_{j \in J} \alpha_j^k \geq \sum_{j \in J} \alpha_j^k \left(f(\hat{x}^k) - c_j^k + \left\langle s : j^k, y - \hat{x}^k \right\rangle \right) - \eta_k \left\langle \sum_{j \in J} \alpha_j^k (x^j - \hat{x}^k) \right\rangle + \sum_{j \in J}$$

5 ...

5.1 Einleitung / Abstract

Dise Arbeit beschäftigt sich mit dem im... Paper vorgestellten bundle Verfahren für nichtkonvexe Probleme mit inexakter Infaormation.

5.2 Description of the method

The method described by ... in the paper generalizes the bundle method for optimizing nonsmooth functions to nonconvex objective functions. The objective function as well as all subgradients may be given in an inexact way.

5.2.1 Preliminaries

- explain subdifferential (also for nonconvex functions???)
- explain general assumptions for whole thesis?
- optimality conditions (Fermat)

5.2.2 General description of a bundle method

The general idea of bundle methods consists in approximating

- Verbesserung des Schnittebenenverfahrens
- generelle Idee: Aapprooximation der nichtglatten Zielfunction durch eine stückweise lineare Funktion
- gibt auch duale Sichtweise des Verfahrens → Approximation des ε-Subdifferentials und dann Verfahren des steilsten Abstieges

Bundle-verfahren wurden zunächst für konvexe Funktionen entwickelt. Hier soll zunächst der einfachste Fall eines Bundle.Verfahrens vorgestellt werden, um dann besser auf wichtige Unterschiede zum im Paper vorgestellten Algorithmus eingehen zu können.

5.2.3 Primales Bundle Verfahren

- wie bei Schnittebenen-Verfahren: Bündel aus bisher berechneten Funktionswerten und Subgradienten → daraus können Suchrichtungen berechnet werden
- Schritte werden im Bundle-Verfahren durch Minimierung einer stückweise linearen Modellfunktion + ein Penaltyterm / Regularisierung berechnet
- Regularisierung verhindert zu lange Schrittweiten und sorgt für eindeutige Lösung des Modells

Etwas genauer

- in jeder Iteration wird der x-Wert, (Funktionswert) und der Subgradient gespeichert
- die Lineare Funktion mit Steigung des Subgradienten und "Aufpunkt" Funktionswert ist Tangente an objective function (wie ist das mit der dualen Sichtweise und den ε-Subgradienten → die berühren ja nicht, aber diese Gradienten schon?)
- bildet man eine stückweise lineare Funktion indem man das Punktweise Maximum betrachtet ist diese Funktion überall eine untere Schranke für die Zeilfunktion.
- je mehr Informationen im Bundle, desto besser ist die Approximation
- Idee: minimiere statt Zielfunktioen die (hoffentlich, sonst sinnlos) einfach-Strukturierte Modell-Funktion (VL: Struktur hängt von Menge X ab, aus der die x kommen → X kann eine beliebig beschränkte Menge sein → dann die "entsprechenden" Probleme der glatten Optimierung???)
- Problem: Lösung nicht immer eindeutig (oder überhaupt existent (wenn X nicht kompakt ist)
- daher einfhren eines Penalty/regularisierungsterms, der die Lösung eindeutig macht (und immer existent)
- Penalty-term kann auch wie in Trust-Region gesehen werden: Die Modellfunktion nähert nur in Umgebung von x^k die Zielfunktion gut an, daher möchte man sich nicht weiter als einen bestimmten Wert von dort entfernen
- neue itterierte wird berechnet → kann sein dass nicht genügend Abstieg erreicht, dann lieber Modell verbessern(sieh auch Trust Region) → serious and null.steps
- Bundle-Update? wichtig? Speziell für bundle verfahren???

Use the third relation in Lemma 2.4.1 in [1]

$$m_k(x^{k+1}) = f(\hat{x}^k) - E_k - t_k\|G^k\|^2$$

to reformulate the nominal decrease δ_k:

$$\delta_k = f(x_k) - m_k(x^{k+1}) - \frac{1}{2}t_k\|G^k\|^2 = E_k + \frac{1}{2}t_k\|G^k\|^2$$

Damit hat man ein Grundlegendes Bundle-Verfahren

5.2.4 Einfaches Primales Bundleverfahren nach Lemaréchal?

Constraints C drin lassen, weil auch in Paper? Aber in Paper vielleicht schöner ohne? → siehe "depth"

Algorithmus aus Buch von Lemaréchal:

Basic Proximal Bundle Method

Select descent parameter $m \in (0,1)$ and a stopping tolerance $\texttt{tol} \geq 0$
Choose a starting point $x^1 \in \mathbb{R}^n$ and compute f_1 and g^1. Set the initial index set $J_1 := \{1\}$ and the initial prox-center to $\hat{x}^1 := x^1$, $\hat{f}_1 = f_1$ and select $t_1 > 0$

For $k = 0, 1, 2, \ldots$

1. Calculate
$$d^k = \arg\min_{d \in \mathbb{R}^n} m_k(\hat{x}^k + d) + \frac{1}{2t_k}\|d^k\|$$
2. Set
$$G^k = \sum_{j \in ???} \alpha_j g_j^k$$
$$\delta_k = E_k + \frac{1}{2}t_k\|G^k\|^2$$
If $\delta_k \leq \texttt{tol} \rightarrow$ STOP
3. Set $x^{k+1} = \hat{x}^k + d^k$
4. compute f^{k+1}, g^{k+1}
If $f^{k+1} \leq \hat{f}^k - m\delta_k \quad \rightarrow$ serious step
Set $\hat{x}^{k+1} = x^{k+1}, \hat{f}^{k+1} = f^{k+1}$ and select $t_{k+1} > 0$??? noch mal nachschauen oder weglassen
Otherwise → nullstep
Set $\hat{x}^{k+1} = \hat{x}^k, \hat{f}^{k+1} = f^{k+1}$ and choose $0 < t_{k+1} \leq t_k$. ??? noch mal nachschauen oder weglassen
5. Select new bundle index set J_{k+1}.
Calculate e_j for $j \in J_{k+1}$ and update the model m^k.

Quelle für Update von J wie in Vorlesung finden.

5.2.5 Unterschiede zu conv, inex

Need to decide which algorithm to take for general Bundle method

- nun auch nichtkonvexe Funktionen zulässig
  Problem: Subgradienten nicht mehr Minoranten für die Ziellfunktion; Teile der Funktion, die unter den Subgradienten liegem, werden nicht beachtet → Minimum kann so übersehen weden
  Lösung: Kovexifiziere Funktion (lokal), sodass sie oberhalb der Modellfunktion durch die Subgradienten bleibt
  ⇒ neue Ausdrücke für die Subgradienten und Fehler, da diese nun für die konvexifizierte Funktkion berechnet werden
- D needed to keep $\hat{x}^k$ and therefore $\hat{f}^k$ bounded in convergence proof ⇒ $\delta_k \to 0$ (equations between (27) and (28) in paper)
  Why else ???????
  can the idea of "depth" be used? Is it nicer? why error in algorithm?
- everything else is the same, maybe look at differences to other papers for better understanding
- inexact information of the data covered by the "nonconvexity machanism" → does no directly appear in algorithm, but in results of convergence proof.

5.2.6 Algorithm

Algorithmus aus Paper:

Nonconvex Proximal Bundle Method with Inexact Information

Select parameters $m \in (0,1), \gamma > 0$ and a stopping tolerance $\texttt{tol} \geq 0$
Choose a starting point $x^1 \in \mathbb{R}^n$ and compute f_1 and g^1. Set the initial index set $J_1 := \{1\}$ and the initial prox-center to $\hat{x}^1 := x^1$, $\hat{f}_1 = f_1$ and select $t_1 > 0$

For $k = 0, 1, 2, \ldots$

1. Calculate
$$d^k = \arg\min_{d \in \mathbb{R}^n} \left\{ M_k(\hat{x}^k + d) + \mathtt{i}_D(\hat{x}^k + d) + \frac{1}{2t_k}\|d\|^2 \right\}$$

2. Set

$$G^k = \sum_{j \in J_k} \alpha_j^k s_j^k, \quad \nu^k = -\frac{1}{t_k} d^k - G^k$$
$$C_k = \sum_{j \in J_k} \alpha_j^k c_j^k$$
$$\delta_k = C_k + t_k \|G^k + \nu^k\|^2$$

   If $\delta_k \leq \texttt{tol} \rightarrow$ STOP

3. Set $x^{k+1} = \hat{x}^k + d^k$

4. compute f^{k+1}, g^{k+1}
   If $f^{k+1} \leq \hat{f}^k - m\delta_k \quad \rightarrow$ serious step
   Set $\hat{x}^{k+1} = x^{k+1}, \hat{f}^{k+1} = f^{k+1}$ and select $t_{k+1} > 0$
   Otherwise $\rightarrow$ nullstep
   Set $\hat{x}^{k+1} = \hat{x}^k, \hat{f}^{k+1} = f^{k+1}$ and choose $0 < t_{k+1} \leq t_k$

5. Select new bundle index set J_{k+1}, keeping all active elements. Calculate

$$\eta_k \geq \max \left\{ \max_{j \in J_{k+1}, x^j \neq \hat{x}^{k+1}} \frac{-2e_j^k}{|x^j - \hat{x}^{k+1}|^2}, 0 \right\} + \gamma$$

   and update the model M^k

5.2.7 Bennenung der Größen

Wie ordnen?????
kann Term “convexified” benutzt werden? welche Bennenung findet sich noch???
auch Formeln oder so dazuschreiben????

x^k	iterates
$\hat{x}^k$	current stability center
$f(x)$	exact evaluation of f at point x
f_k	value of f at point x^k from the oracle, may be inexact
$\hat{f}_k$	value of f at stability center $\hat{x}^k$ from the oracle, may be inexact
g^k	a subgradient of f at x^k (can be exact or inexact)
s^k	a subgradient of the convexified objective at x^k (can be exact or inexact)
m_k	cutting plane model of the objective function f
M_k	cutting plane model of the convexified objective function
e_j^k	linearization error
c_j^k	linearization error of the convexified objective function
η_k	convexification parameter
α_j^k	Lagrange multipliers of the subproblem
d^k	minimizer of the subproblem
ξ	variable from reformulation of the subproblem
J_k	index set at iteration k
t_k	prox-parameter
δ_k	nominal decrease
m	decrease parameter
G^k	aggregate subgradient
S^k	aggregate subgradient of convexified objective
E^k	aggregate error
C^k	aggregate error of convexified objective

5.2.8 Convergence proofs

Unterschiede in Results:
stationary points vs. minimum in nonconvex vs. convex case
error vs. no error in inexact vs. exact
only one accumulation point vs. all accumulation points it t_k not bounded from below $\rightarrow$ is t_k so important? Why on just leave it out?

Abbruchbedingung Schnittebenenverfahren mit anderen abbruchbedingungen vergleichen
Für all die Aussagen oben, die aus der VL kommen, eine quelle finden (VL-Quellen anschauen)
(Lemaréchal Convex Analysis and Minimization Algorithms II Advanced Theory and Bundle Methods).
It is immediate to verify that a function f is α-strongly convex if and only if $x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$ is convex → gleichmäßig konvex, Funktion muss quasi "konvexer" als die gestauchte (gestreckte) Parabel sein (https://blogs.princeton.edu/imabandit/2013/04/04/orf523-strong-convexity/) Quelle??? Aber reicht nicht schon strikte konvexität? ist in diesem Fall glm. nur einfacher zu erreichen???

Thoughts to differences between nonconv exact and inexact

Convergence proofs

- $f + \eta$ is "convex on bundle points"
  can't find out more, even in exact case → start from this as in convex case ???
- general idea of convergence proof: show that $E_k \to 0$ and $\|G^k + \nu^k\| \to 0$ follow stationary condition from that
- proof for serious steps almost same as in conv exact
  only D needed for boundedness of f (always given id f convex)
- Ulbrich shows that sequence of subproblem solutions (nonserious iterates x^k is bounded (above + below) other shows only bounded above and increasing
  $\Rightarrow$ both use it to show that $E_k \to 0$ and $\|G^k + \nu^k\| \to 0$

Update R, μ_k, η_k, t_k

- η_k has basicly same formula
  in older papers: nondecreasing sequence $\{\eta_k\}$
- R: in older papers this chosen
  update:
  in prox-points: if μ_k gets too small → break algorithm and demand greater R
  in conv-ex: if increase for next iterate is too big make μ_k and therefore R bigger
- μ_k prox-points: if too short steps: make μ smaller and η bigger; if not "normal" rule (calculate η_k and get μ_k from difference with R)
  in conv-ex: if increase for next iterate is too big make μ_k (and therefore R) bigger
- t_k: trust region like-update; R changes with every update
  t_k smaller (= μ_k bigger) if null step (= trust region smaller)
  t_k anyhow if serious step

Results for nonconv inex

- choose η_k such that function is “convex on bundle points”
- choose $t_k \Leftrightarrow \mu_k$ like for trust region algorithm (convex bundle)
- seems that after ”right” convexification of the objective function no more thinking about nonconvexity necessary

!!! has D anything to do with these parameters???

Literatur

[1] Jean-Baptiste Hiriart-Urruty and Claude LemarÃ©chal. *Convex analysis and minimization algorithms II*, volume 306 of *Grundlehren der mathematischen Wissenschaften*. Springer -Verlag Berlin Heidelberg, 1993.