Inhaltsverzeichnis

1 Preliminaries **2**

2 Bundle Methods **3**

2.1 A basic bundle method 3

2.2 Proximal bundle method for nonconvex functions with inexact information 9

2.3 Convergence analysis 13

3 How is inexact information dealt with? **17**

4 Extension with Limited Memory approach **17**

4.1 Convergence 19

Bibliography **19**

1 Preliminaries

Check if requirements on functions are stated and defined.

Throughout this thesis I consider the optimization Problem

$$\min_x f(x), \quad x \in X \subseteq \mathbb{R}^n \tag{1}$$

where f is a possibly nonsmooth function. Also write something about inexactness? specify X more precisely? Convex?
When it comes to nonsmooth objective functions the derivative based framework of nonlinear optimization methods does not work any more. Therefore the most important definitions and results needed when working with nonsmooth functions are stated in this section.
Just definition, lemma, theorem or a bit explanation around it?
See if requirements in definitions and theorems meet what is needed/provided later.

Definition 1.1. The function $f : \mathbb{R}^n \to \mathbb{R}$ is called *Lipschitz near* $x \in \mathbb{R}^n$ if there exist $C > 0$ and $\varepsilon > 0$ such that

$$|f(y_2) - f(y_1)| \leq C\|y_2 - y_1\| \quad \forall y_1, y_2 \in \mathbf{B}_\varepsilon(x).$$

Definition 1.2. The function $f : \mathbb{R}^n \to \mathbb{R}$ is called *locally Lipschitz* if it is Lipschitz on each bounded subset $B \subseteq \mathbb{R}^n$

$$|f(y) - f(x)| \leq C\|y - x\| \quad \forall x, y \in B, \quad C > 0.$$

Definition 1.3. The *directional derivative* of f at x in direction d is

$$f'(x, d) := \lim_{\lambda \downarrow 0} \frac{f(x + \lambda d) - f(x)}{\lambda}.$$

Definition 1.4. Let f convex. The *subdifferential* $\partial f(x)$ of f at x is the nonempty compact convex set

$$\partial f(x) = \{g \in \mathbb{R}^n | f'(x, d) \geq \langle g, d\rangle \forall d \in \mathbb{R}^n\}.$$

Definition 1.5. The *generalized directional derivative* of f at x in direction d is given by

$$f^\circ(x, d) := \limsup_{\substack{y \to x \\ \lambda \downarrow 0}} \frac{f(y + \lambda d) - f(y)}{\lambda}.$$

Definition 1.6. The *generalized gradient* of f at x is a nonempty convex compact set $\partial f(x)$ given by

$$\partial f(x) := \{g \in \mathbb{R}^n | f^\circ(x, d) \geq \langle g, d\rangle \forall d \in \mathbb{R}^n\}.$$

If f is a convex function the generalized gradient coincides with the subdifferential ∂f of f [?].

2 Bundle Methods

introduction

→ introduction sentences: main focus on algorithm in paper; therefore at first introduce simplest form of bundle methods. When bundle methods were first introduced in 1975 by Claude Lemaréchal and Philip Wolfe they were developed to minimize a convex (possibly nonsmooth) function f for which at least one subgradient at any point x can be computed [?].

To provide an easier understanding of the proximal bundle method in [?] and stress the most important ideas of how to deal with nonconvexity and inexactness first a basic bundle method is shown here.

Bundle methods can be interpreted in two different ways: From the dual point of view one tries to approximate the ε-subdifferential to finally ensure first order optimality conditions. The primal point of view interprets the bundle method as a stabilized form of the cutting plane method where the objective function is modeled by tangent hyperplanes [?]. I focus here on the primal approach.

In the next two sections the function f is assumed to be convex.

notation, definitions

already done in previous preliminaries chapter?

2.1 A basic bundle method

This section gives a short summery of the derivations and results of chapter XV in [?] where a primal bundle method is derived as a stabilized version of the cutting plane method. If not otherwise indicated the results in this section are therefore taken from [?].

The optimization problem considered in this section is

$$\min_x f(x) \quad \text{s.t.} \quad x \in X \tag{2}$$

with the convex function f and the closed and convex set $X \subseteq \mathbb{R}^n$.

Define Problem again?? Incorporate “set-constraint” by writing $h(x) := f(x) + \mathbb{I}_X$. → later???

explanation

2.1.1 Derivation of the bundle method

The geometric idea of the cutting plane method is to build a piecewise linear model of the objective function f that can be minimized more easily than the original objective function.
This model is built from a *bundle* of information that is gathered in the previous iterations. In the k'th iteration, the bundle consists of the previous iterates x^j, the respective function values $f(x^j)$ and a subgradient at each point $g^j \in \partial f(x^j)$ for all indices j in the index set J_k. From each of these triples, one can construct a linear function

$$l_j(x) = f(x^j) + (g^j)^\top (x - x^j) \tag{3}$$

with $f(x^j) = l_j(x^j)$ and due to convexity $f(x) \geq l_j(x), \ x \in X$.
One can now model the objective function f by the piecewise linear function

$$m_k(x) = \max_{j \in J_k} l_j(x) \tag{4}$$

and find a new iterate x^{k+1} by solving the subproblem

$$\min_x m_k(x) \quad \text{s.t.} \quad x \in X. \tag{5}$$

This subproblem should of course be easier to solve than the original task. A question that depends a lot on the structure of X. If $X = \mathbb{R}^n$ or a polyhedron, the problem can be solved easily. Still there are some major drawbacks to the idea. For example if $X = \mathbb{R}^n$ the solution of the subproblem in the first iteration is always $-\infty$.
In general one can say that the subproblem does not necessarily have to have a solution. To tackle this problem a penalty term is introduced to the subproblem:

$$\tilde{m}_k(x) = m_k(x) + \frac{1}{2t}\|x - x^k\|^2 \quad \text{s.t.} \quad x \in X \tag{6}$$

This new subproblem is strongly convex and has therefore always a unique solution.
how much explanation here? $\max_{j \in J_k} l_j(\hat{x}^k + d)$

Some nice sentences to explain the term a little bit more and to lead over to the next paragraph.
To understand the deeper motivation of this term see [?]. For this introduction it suffices to see that due to the regularization term the subproblem is now strongly convex and therefore always uniquely solvable.

The second major step towards the bundle algorithm is the introduction of a so called *stability center* or *serious point* $\hat{x}^k$. It is the iterate that yields the "best" approximation of the optimal point up to the k'th iteration (not necessarily the best function value though).

The updating technique for $\hat{x}^k$ is crucial for the convergence of the method: If the next iterate yields a decrease of f that is "big enough", namely bigger than a fraction of the decrease suggested by the model function for this iterate, the stability center is moved to that iterate. If this is not the case, the stability center remains unchanged.
In practice this looks the following:
Define first the *nominal decrease* δ_k which is the decrease of the model for the new iterate x^{k+1} compared to the function value at the current stability center $\hat{x}^k$.

$$\delta^k = f(\hat{x}^k) - \tilde{m}_k(x^{k+1}) + a_k \geq 0 \tag{7}$$

The nominal decrease is in fact stated a little differently for different versions of the bundle algorithm, this is why I added the constant $a_k \in \mathbb{R}$ here for generalization. In practice the difference between the decreases is not influencing the algorithm as δ_k is weighted by the constant $m \in (0,1)$ for the descent test which compensates a_k.

If the actual decrease of the objective function is bigger than a fraction of the nominal decrease

$$f(\hat{x}^k) - f(x^{k+1}) \geq m\delta_k, \quad m \in (0,1)$$

set the stability center to $\hat{x}^{k+1} = x^{k+1}$. This is called a *serious* or *descent step*.
If this is not the case a *null step* is executed and the serous iterate remains the same $\hat{x}^{k+1} = \hat{x}^k$.

The subproblem can be rewritten as a smooth optimization problem. For convenience rewrite the affine functions l_j with respect to the stability center $\hat{x}^k$.

citation for this???!!!

$$\begin{aligned} l_j(x) &= f(x^j) + {g^j}^\top (x - x^j) \qquad (8)\\ &= f(\hat{x}^k) + {g^j}^\top (x - \hat{x}^k) - (f(\hat{x}^k) - f(x^j) + {g^j}^\top (x^j - \hat{x}^k)) \qquad (9)\\ &= f(\hat{x}^k) + {g^j}^\top (x - \hat{x}^k) - e_j^k \qquad (10)\end{aligned}$$

where

$$e_j^k = f(\hat{x}^k) - f(x^j) + {g^j}^\top (x^j - \hat{x}^k) \geq 0 \quad \forall j \in J_k \tag{11}$$

is the *linearization error*. The nonnegativity property is essential for the convergence theory and will also be of interest when moving on to the case of nonconvex and inexact objective functions.

The subproblem can now be written as

$$\min_{\hat{x}^k+d\in X} \tilde{m}_k(d) = f(\hat{x}^k) + \max_{j\in J_k}\{{g^j}^\top d - e_j^k\} + \frac{1}{2t_k}\|d\|^2 \tag{12}$$

$$\Leftrightarrow \quad \min_{\hat{x}^k+d\in X} \xi + \frac{1}{2t_k}\|d\|^2 \quad \text{s.t.} \quad f(\hat{x}^k) + {g^j}^\top d - e_j^k - \xi \leq 0, \quad j \in J_k \tag{13}$$

where the constant term $f(\hat{x}^k)$ was discarded for the sake of simplicity.
If X is a polyhedron this is a quadratic optimization problem that can be solved using standard methods of nonlinear optimization. The pair (ξ_k, d^k) solves (13) if and only if d^k solves the original subproblem (12) and $\xi_k = f(\hat{x}^k) + \max_{j\in J_k} {g^j}^\top d^k - e_j^k$. The new iterate is then given by $x^{k+1} = \hat{x}^k + d^k$.

Remark: Setting $\check{f}(x) = f(x) + \mathbb{I}_X(x)$ the above optimization problem is ...
The *proximal point mapping* or *prox-operator*

$$prox_{t,f}(x) = \arg\min_y \left\{ \check{f}(y) + \frac{1}{2t}\|x-y\|^2 \right\}, \quad t > 0 \tag{14}$$

source??? This special form of the subproblems gives the proximal bundle method its name and will occur again later???

2.1.2 Aggregate objects

The constraint $\hat{x}^k + d \in X$ can also be incorporated directly in the objective function by using the indicator function

$$\mathbb{I}_X(x) = \begin{cases} 0, & \text{if } x \in X \\ +\infty, & \text{if } x \notin X \end{cases} .$$

The subproblem then writes as

$$\min_{\hat{x}^k+d\in R^n} \xi + \mathbb{I}_X + \frac{1}{2t_k}\|d\|^2 \quad \text{s.t.} \quad {g^j}^\top d - e_j^k - \xi \leq 0, \quad j \in J_k \tag{15}$$

Some introduction how this and the aggregate error expression relate to each other. Why it is in this case easier to write the model in the nonsmooth form...

Lemma XI 3.1.1 $\partial g = \partial f + \partial \mathbb{I}_X$ for $g = f + \mathbb{I}_X$.

One gets the following results about the step d^k of the subproblem:

Lemma 2.1. *The optimization problem (15) has for $t_k > 0$ a unique solution given by*

$$d^k = -t_k(G^k + \nu^k), \quad G^k \in \partial m_k(d^k), \quad \nu^k \in \partial \mathbb{I}_X. \tag{16}$$

Furthermore

$$m_k(\hat{x}^k + d) \geq f(\hat{x}^k) + {G^k}^\top d - E_k \quad \forall d \in \mathbb{R}^n \tag{17}$$

inequality because of aggregation technique. Is sharp when cutting plane model is used? source?

where

$$E_k := f(\hat{x}^k) - m_k(x^{k+1}) + {G^k}^\top d^k. \tag{18}$$

Comment on the inequality missing

The quantities G^k and E^k are the *aggregate subgradient* and the *aggregate error*.

Explain aggregation process in more detail

From the Karush-Kuhn-Tucker conditions (KKT-conditions) one can see that in the optimum there exist Lagrange or *simplicial multiplier* α_j^k, $j \in J_k$ such that

$$\alpha_j^k \geq 0, \quad \sum_{j \in J_k} \alpha_j^k = 1 \tag{19}$$

by rewriting and so on... one can see that the above expressions are in fact

From the dual problem one obtains that the aggregate subgradient and error can also be expressed as

$$E_k = \sum_{j \in J_k} \alpha_j^k e_j^k \qquad \text{and} \qquad G^k = \sum_{j \in J_k} \alpha_j^k g^j. \tag{20}$$

Finally use Lemma ??? in [?]

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

to reformulate the nominal decrease δ_k:

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

The nominal decrease in this case is defined as:
noch mal anschauen

$$\delta_k := E_k + t_k \|G^k + \nu^k\|^2 = f(\hat{x}^k) - m_k(x^{k+1}) - {\nu^k}^\top d^k \tag{21}$$

In practice the different definition of the decreases makes no difference because of the weighting with the descent parameter m.

The following basic bundle algorithm can now be stated:

Reformluate equations, model function
introduce aggregate expressions
say something to J-update, say something to t-update

see if all abbreviations ($f_j, g^j, \ldots$) are introduced
introduce prox-operator and proximal points

algorithm

Basic 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 $f(x^1)$ and g^1. Set the initial index set $J_1 := \{1\}$ and the initial stability center to $\hat{x}^1 := x^1$, $f(\hat{x}^1) = f(x^1)$ and select $t_1 > 0$.

For $k = 1, 2, 3 \ldots$

1. Calculate
$$d^k = \arg\min_{d \in \mathbb{R}^n} m_k(\hat{x}^k + d) + \mathbb{I}_X + \frac{1}{2t_k}\|d\|^2$$
and the corresponding Lagrange multiplier α_j^k, $j \in J_k$. say how model m_k looks here. include $\mathbb{I}_X$
2. Set
$$G^k = \sum_{j \in J_k} \alpha_j^k g_j^k, \qquad E_k = \sum_{j \in J_k} \alpha_j^k e_j^k, \qquad \text{and} \qquad \delta_k = E_k + t_k \|G^k + \nu^k\|^2$$
If $\delta_k \leq \texttt{tol} \rightarrow$ STOP.
3. Set $x^{k+1} = \hat{x}^k + d^k$.
4. Compute $f(x^{k+1})$, g^{k+1}.
If
$$f^{k+1} \leq \hat{f}^k - m\delta_k \qquad \rightarrow \text{serious step.}$$
Set $\hat{x}^{k+1} = x^{k+1}$, $f(\hat{x}^{k+1}) = f(x^{k+1})$ and select suitable $t_{k+1} > 0$.
Otherwise → nullstep.
Set $\hat{x}^{k+1} = \hat{x}^k$, $f(\hat{x}^{k+1}) = f(x^{k+1})$ and choose t_{k+1} in a suitable way.
5. Select new bundle index set $J_{k+1} = \{j \in J_k | \alpha_j^{k+1} \neq 0\} \cap k+1$, calculate e_j for $j \in J_{k+1}$ and update the model m_k.

In steps 4 and 5 of the algorithm the updates of the steplength t_k and the index set J_k are are only given in a very general form.
The "suitable" choice of t_k will be discussed more closely in the convergence analysis of decide which method; say that $t_k > 0 \forall k$....
Comment on J_k update → depends on what is included in thesis.
For the choice of the new index set J_{k+1} different aggregation methods to keep the memory size controllable are available. The most easy and intuitive one is to just take those parts of the model function, that are actually active in the current iteration. This is done in this basic version of the method.
Refer to low memory bundling if later in thesis. Instead of keeping every index in the set

J_k different compression ideas exist. For now I therefor stick to this update.
refer to later "low memory" thing??

explanation to t_k update. → include at which point??? This simple idea has however some major drawbacks [?]:

- Minimization of the cutting plane model of the objective function is not trivial. Indeed unconstrained minimization of the model is never possible in the first step, where it is just a line, unless the starting point is already a minimum.
- The convergence speed is very slow.

If convergence speed named here, does it have to be shown (rates)? For all algorithms??? Leave out? Argue about instability?

To address those issues a regularization is added to the cutting plane model. This ensures unique solvability of the minimization of the subproblem. By introducing a stability center and

2.2 Proximal bundle method for nonconvex functions with inexact information

introduction

This section focuses on the proximal bundle method presented in [?].
The idea is to extend the basic bundle algorithm for nonconvex functions with both inexact function and subgradient information.
The key idea of the algorithm is the one already developed for [?]: When dealing with nonconvex functions a very critical difference to the convex case is that the linearization errors are not necessarily nonnegative any more. To tackle this problem the errors are manipulated to enforce nonnegativity. In this case this is done my modeling not the objective function directly but a convexified version of it.
citation for that??? sources of nonconv-exact? something about primal-dual view???

2.2.1 New subsubsection?

"assumptions and notations"

introduce exact optimization problem that is used in this section and its properties if not already introduce in "Preliminaries".

Throughout this section the optimization problem

$$\min_x f(x) \quad \text{s.t.} \quad x \in X \tag{22}$$

where f is locally Lipschitz is considered. $X \subseteq \mathbb{R}^n$ is assumed to be a convex compact set. Both the function value as well as the subgradient can be provided in an inexact form.

For the function value inexactness is defined straight forwardly: If

$$\|\tilde{f} - f(x)\| \leq \sigma \tag{23}$$

then $\tilde{f}$ approximates the value $f(x)$ within σ.
For the subgradients inexactness is interpreted in the following way: $\tilde{g} \in \mathbb{R}^n$ approximates a subgradient $g \in \partial f(x)$ within $\theta \geq 0$ if

$$\tilde{g} \in \partial f(x) + B_\theta(0). \tag{24}$$

In the paper it is assumed that the errors are bounded although the bound does not have to be known.

$$|\sigma_j| \leq \bar{\sigma} \quad \text{and} \quad 0 \leq \theta_j \leq \bar{\theta} \quad \forall j \in J_k. \tag{25}$$

In the context of inexact information it is important to make a distinction between the (unknown) exact function value and its approximation. Throughout this chapter I therefore write $f(x)$ for the exact function value whereas the approximation will be written as f_j or $\hat{f}_k$ for the approximation at the current stability center.

The objective function $f : \mathbb{R}^n \to \mathbb{R}$ is assumed to be proper, (subdifferentially) regular and locally Lipschitz continuous with full domain.

Definition 2.2. [?] A function $f : \mathbb{R}^n \to \bar{\mathbb{R}} = [-\infty, +\infty]$ is called *proper* if $f(x) < \infty$ for at least one $x \in \mathbb{R}^n$ and $f(x) > \infty$ $\forall x \in \mathbb{R}^n$.

Definition 2.3. [?] $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is called *subdifferentially regular* at $\bar{x}$ if $f(\bar{x})$ is finite and the epigraph

$$epi(f) := \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} | \alpha \geq f(x)\}$$

is Clarke regular at $\bar{x}, f(\bar{x})$.

Closed convex sets are Clarke regular, so in particular the epigraph of lower $\mathcal{C}^2$-functions?.

Definition semismooth for later:

Definition 2.4. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *semismooth* ar $x \in \mathbb{R}^n$ if f is Lipschitz near x and for each $d \in \mathbb{R}^n$ and for any sequences $\{t_k\} \subseteq \mathbb{R}_+, \{\theta^k\} \subseteq \mathbb{R}^n$ and $\{g^k\} \subseteq \mathbb{R}^n$ such that

$$\{t_k\} \downarrow 0, \quad \{\theta^k / t_k\} \to 0 \in \mathbb{R}^n \quad \text{and} \quad g^k \in \partial f(x + t_k d + \theta^k),$$

the sequence $\{\langle g^k, d\rangle\}$ has exactly one accumulation point.

Definition 2.5. A point $x \in \mathbb{R}^n$ that satisfies $0 \in \partial f(x)$ is called a *stationary point* of f.

explanation

A main issue both nonconvexity and inexactness entail is that the linearization errors e_j^k are not necessarily nonnegative any more.
So based on the results in [**?**] not the objective function but a convexified version of it is modeled as the objective function of the subproblem.
explain locally convexified more precisely? Is it because no global convexification? different than in [18]??

When looking at the subproblem formulated as in (12) one can see that the new iterate x^{k+1} is in fact a *proximal point* of the subproblem.
The *proximal point mapping* or *prox-operator* is defined as

$$prox_{t,f}(x) = \arg\min_y \left\{ \check{f}(y) + \frac{1}{2t}\|x-y\|^2 \right\}, \quad t > 0 \tag{26}$$

For $\check{f}(x) := m_(x) + \mathbb{I}_X(x)$ and $\mu := \frac{1}{t_k}$ this is just subproblem (12) with the constraint $x \in X$ incorporated in the objective function. Because of this special form of the subproblems primal bundle methods are also called proximal bundle methods.

explain in much more detail when read about calculation of proximal points for nonconvex functions. At the moment just main ideas.

The key idea is now to use the relation

$$prox_{R=\mu+\eta,f}(x) = prox_{\mu,f+\eta/2\cdot|\cdot-x|^2}(x). \tag{27}$$

This means, that the proximal point of the function f for parameter $R = \eta + \mu$ is the same as calculating the proximal point of the regularized function

$$\tilde{f}(y) = f(y) + \frac{\eta}{2}|y-x|^2 \tag{28}$$

with respect to the parameter μ. η is therefore called the *convexification parameter* and μ is the *prox-parameter*.

So the function that will be modeled by the cutting plane approximation is no longer the original objective function f but the convexified version $\tilde{f}$.

say why/how... this is related to current stability center Because new function to be approximated, subgradients “new”: The linear functions forming the model have therefore a tilted slope

$$s_j^k = g^j + \eta_k \left(x^j - \hat{x}^k \right). \tag{29}$$

linearization error defined just as to be nonnegative. → any further motivation???

be careful now slightly different definition of linearization error because of inexact information

η is defined to be such that the augmented linearization error is nonnegative:

$$\eta_k \geq \max \left\{ \max_{j \in J_k, x^j \neq \hat{x}^k} \frac{-2e_j^k}{\|x^j - \hat{x}^k\|^2}, 0 \right\} + \gamma \tag{30}$$

With the "saveguarding parameter" $\gamma \geq 0$

explain, why linearization errors are defined like that

$$0 \leq c_j^k := e_j^k + b_j^k, \quad \text{with} \quad \begin{cases} e_j^k := \hat{f}_k - f_j - \langle g^j, \hat{x}^k - x^j \rangle \\ b_j^k := \frac{\eta_k}{2} \|x^j - \hat{x}^k\|^2 \end{cases} \tag{31}$$

The new model function can therefore be written as

$$M_k(d) := \hat{f}_k + \max_{j \in J_k} \left\{ {s_j^k}^\top d - c_j^k \right\} \tag{32}$$

check that M_k and m_k from basic algorithm above have same form.

Explain how the D comes into the whole thing
here (also) because already said above (but why?)
already introduce in basic algorithm!
maybe later say something to how it is done in "depth"-Paper???

Some Lemma why this is so

$$d^k = -t_k (G^k + \nu^k), \quad \text{where} \quad \nu^k \in \partial \mathbb{I}_D(x^{k+1}) \tag{33}$$

The definition of the aggregate objects follows straightforward:

$$S^k := \sum_{j \in J_k} \alpha_j^k s_j^k \tag{34}$$

$$C_k := \sum_{j \in J_k} \alpha_j^k c_j^k \tag{35}$$

explain how δ_k is derived.

$$\delta^k := C_k + t_k \|S^k + \nu^k\|^2 \tag{36}$$

algorithm

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 $k = 1, 2, 3, \dots$
  1. Calculate
       $d^k = \arg\min_{d \in \mathbb{R}^n} \left\{ M_k(\hat{x}^k + d) + \mathbb{I}_X(\hat{x}^k + d) + \frac{1}{2t_k}\|d\|^2 \right\}.$
  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 $\delta_k \leq \texttt{tol} \to$ STOP.
  3. Set $x^{k+1} = \hat{x}^k + d^k$.
  4. Compute $f^{k+1}, g^{k+1}$
     If
       $f^{k+1} \leq \hat{f}^k - m\delta_k \quad \to$ serious step
     Set $\hat{x}^{k+1} = x^{k+1}, \hat{f}^{k+1} = f^{k+1}$ and select $t_{k+1} > 0$.
     Otherwise $\to$ 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 $M^k$
```

2.3 Convergence analysis

2.3.1 Results for objectives with exact information

The main ideas of the algorithm are basicly the ones developed in [**?**] for the redistributed proximal bundle method for exact nonconvex problems.
Setting the error bounds $\bar{\sigma}$ and $\bar{\theta}$ to zero results therefore in the following convergence theorem.

Theorem 2.6. *Let the sequence* $\{\eta_k\}$ *be bounded,* $\liminf_{k\to\infty}$ *and the cardinality of the*

set $\{j \in J_k | \alpha_j^k > 0\}$ be uniformly bounded in k.
Then every accumulation point of sequence of serious iterates $\{\hat{x}^k\}$ is a stationary point of the problem.

think last condition only interesting in inexact case.

In the exact case boundedness of the sequence $\{\eta_k\}$ is proven for lower-$\mathcal{C}^2$ functions in [?]. This is not possible in the inexact case, even if the objective function f is convex.

A further simplification of the method for exact information is not necessary as the method is already almost as simple as the basic bundle method for nonconvex exact functions. Additionally no new concepts needed to be introduced when doing the step from nonconvex exact problems, for which the algorithm was originally designed, to problems with inexact information.

Remark: I want to add here, that the simplicity of the algorithm is rather special for methods suitable for nonconvex problems. Often a linesearch algorithm has to be inserted in the nonconvex case, which is not needed here.

2.3.2 Nonconvex bundle methods

There are different approaches for handling nonconvexity of the objective function in bundle methods. As the nonnegativity property of the linearization errors e_j^k is crucial for the convergence proof of convex bundle methods an early idea was forcing the errors to be so by different downshifting strategies. A very common one is using the *subgradient locality measure* [?, ?]. Here the linearization error is essentially replaced by the nonnegative number

$$\tilde{e}_j^k := \max_{j \in J_k} \{|e_j^k|, \gamma \|\hat{x}^k - x^j\|^2\}. \tag{37}$$

Remark on dual view? How subgradient locality measure measures how close subgradient is to subdifferential of f???
Methods using this kind of manipulation of the model function are endowed with a line search to provide sufficient decrease of the objective function. For the linesearch to terminate finitely, semismoothness of the objective function is usually needed.
It can be proven that every accumulation point of the sequence of serious points $\{\hat{x}^k\}$ is a stationary point of the objective function f under the additional assumptions that f is locally Lipschitz and the level set $\{x \in \mathbb{R}^n | f(x) \leq f(\hat{x}^1)\}$ is bounded [?].

A drawback to the method described a bove is that it is primarily supported from the dual point of view of the bundle algorithm. Newer concepts focus also on the primal point of view. This invokes for example having different model functions for the subproblem.

In [?, ?] the difference function

$$h(d) := f(x^j + d) - f(x^j) \quad j \in J_k \tag{38}$$

is approximated to find descent direction of f.
The negative linearization errors are addressed by having two different bundles. One containing the indices with nonnegative linearization errors and one containing the other ones. From these two bundles two cutting plane approximations can be constructed which provide the bases for the calculation of the new iterate.
Convergence of the method to a stationary point is proven under the assumption of f being locally Lipschitz and semismooth.

In [**?**] Noll et al. follow an approach of approximating a local model of the objective function. The model can be seen as a nonsmooth generalization of the Taylor expansion and looks the following:

$$\Phi(y,x) = \phi(y,x) + \frac{1}{2}(y-x)^\top Q(x)(y-x) \tag{39}$$

The so called *first order model* $\phi(.,x)$ is convex but possibly nonsmooth and can be approximated by cutting planes. The *second order part* is a quadratic but not necessarily convex. The algorithm then proceeds a lot in the lines of a general bundle algorithm. For a locally Lipschitz objective function with a bounded levelset $\{x \in \mathbb{R}^n | f(x) \leq f(\hat{x}^1)\}$ convergence to a stationary point is established.

Other papers reach the same results only needing L-continuity and the boundedness of the level sets (in [**?**] established by introducing the set D). But they use a different concept. Explain "most common?" concepts for dealing with nonconvexity??? see [**?**].

- in paper other method for calculating $\mu = \frac{1}{t_k}$
  how is made sure that η big enough???
- convergence results for nonconvex functions:?
  - conditions on functions:
    exact: f lower-$\mathcal{C}^2$ near the minimizers of the problem
    inexact: proper, regular, locally Lipschitz with full domain; even better: lower-$\mathcal{C}^1$ (contains lower-$\mathcal{C}^2$) → conditions more general than in exact case
  - convergence results:
    exact: the limit of the sequence $\{x^k\}$ (which exists) or every accumulation point of the sequence $\{\hat{x}^k\}$ is a stationary point of f
    Does this mean: $0 \in \partial f(\bar{x})$, $\bar{x}$ being the respective limit??? incorporate set D?
    inexact: $0 \in (\partial f(\bar{x}) + \partial I_D(\bar{x})) + B_{\bar{\theta}}(0)$
    if f lower-$\mathcal{C}^1$:

    $$\forall \varepsilon > 0 \; \exists \rho > 0 : \quad f(y) \geq f(\bar{x}) - (\bar{\theta} + \varepsilon)\|y - \bar{x}\| - 2\bar{\sigma} \quad \forall y \in D \cap B_\rho(\bar{x})$$

  - obvious difference: no error terms in exact case

Results:

- better in comparison to other paper, because wider range of functions (lower-$\mathcal{C}^1$).

2.3.3 if convex function

Convergence for inexact convex functions:

- states in paper [**?**] (p. 14) that for convex functions error of $\bar{\sigma}$ instead of $2\bar{\sigma}$ possible (and for lower models; see depth paper?)

To Do:

- proof serious steps
- proof null steps
- limit of G^k
- proof in book; see if possible to leave out D; compare
  should be possible if bounded level sets assumed; check this!
  compare with "depth" $\rightarrow \phi$
- see if η_k can be bounded in exact case - yes for the class of functions mentioned in the paper
- find counterexample, that η can't be bounded in inexact case??? - main argument: have to assure, that convexified objective function is "convex on all bundle points x^j" from a certain η^k on
- find out about $\eta > \rho$-condition? Or unnecessary
- compare nonconvex exact $\leftrightarrow$ inexact convergence results
  only look at exact paper again if better results!
  - check if correct: inexact more general because choice of t_k more freely??? in exact μ_k only changed when restart
    update strategy not important for convergence; maybe for convergence speed?
  - check if update strategy important for convergence speed? - yes see napsu ... "Comparison ..."
- check if (ii) in Theorem 6 in paper can be assured by choice of t_k in algorithm (think yes)
- compare to convergence results of other papers
  check if better results can be carried over - results all the same; check for prerequisites on functions
- check if other papers have better prerequisites
  check if results can be carried over
  all need locally Lipschitz and either a compact? subset or bounded lower level sets
  Better in other papers?: neither $\{j \in J_k | \alpha_j^k > 0\}$ nor $\{\eta_k\}$ need to be bounded??
  any prerequisites on t_k in other papers???

- most other solutions for nonconvex functions: based on dual idea; remark on dual view on bundle method?
- write that down nicely
- read depth paper
  - see what kind my algorithm is
  - check Lemma 5 (iv) stronger?
- remark on that inexactness makes objective nonconvex → relate to nonconvexity paper?
- check again uniformly bounded J_k
  aggregated objects (3.4), 5.2, 7.1 in "depth"
- generalized gradients may only be shifted not tilted?
  is this realistic assumption? → all ok, tilted and shifted!
- ask Simon to ε-subdifferentials
- Algorithm:
  - print something useful in every iteration
  - η should not get too big

If (newer) papers needed: look at citations of the ones I have.

(sub-)Level sets of continuous functions are closed (image of closed set closed under continuous function); should also hold for lower semicontinuous functions in metrizable(???) space

3 How is inexact information dealt with?

4 Extension with Limited Memory approach

Variable metric methods are also known as quasi-Newton methods.

PhD Thesis (p. 33) inheritance of positive definiteness of matrix update formulas
p. 34: number of correction pairs usually $3 \leq m \leq 30$

4.0.4 Thoughts about line search

- after what is written in "nonconv, inex"-paper: Line search not provable to be finite if inexact information
- is line search standard in all variable metric methods?
  - looks like it

  - there do exist versions without line search → other update???
  - does is work without line search??? -think yes
- does prox-parameter t_k have some relation to linesearch/stepsize?

Algorithm 1 in [?]

Variable metric limited memory bundle method with inexact information

Select parameters $m \in (0,1),\ \gamma > 0,\ K > 0,\ \rho \in (0,\ \frac{1}{2})$ and a stopping tolerance `tol` ≥ 0. Set the initial metric matrix $D_1 = \mathbb{I}_{n \times n}$. Choose a starting point $x^1 \in \mathbb{R}^n$ and compute f_1 and $s^1 = g^1 = S^1$. Set the initial serious iterate $\hat{x}^1 := x^1$, $\hat{f}_1 = f_1$, and $\hat{s}^1 = s^1$, and $c_1 = C_1 = 0$.
Set the correction indicator and the correction indicator for consecutive null steps $i_C = 0$.

For $k = 1, 2, 3, \ldots$

1. Compute
$$d^k = -D_k S^k$$
by using a limited memory BFGS update if the step before was a serious step and by using a limited memory SR1 update otherwise.
For $k = 1$: $d^1 = -S^1$.

2. If $-{S^k}^\top d^k < \rho {S^k}^\top S^k$ or $i_{CN} = 1$ then set
$$d^k = d^k - \rho S^k, \tag{40}$$
and $i_C = 1$. If the previous step was a null step set also $i_{CN} = 1$.
Otherwise set $i_C = 0$.

3. Set
$$\delta_k^w = -{S^k}^\top d^k + 2C_k \quad \text{and} \tag{41}$$
$$\delta_k^q = \frac{1}{2}{S^k}^\top S^k + C_k. \tag{42}$$
If $\delta_k^w <$ `tol` and $\delta_k^q <$ `tol` stop with $\hat{x}_k$ as the final solution.

4. Set
$$\theta_k = \min\{1, K/\|d^k\|\} \tag{43}$$
$$x^{k+1} = \hat{x}^k + \theta_k d^k. \tag{44}$$
Calculate the inexact function value f_{k+1} and an inexact subgradient g^{k+1}.

5. If
$$f^{k+1} \leq \hat{f}^k - m\delta_k^w \quad \rightarrow \text{ serious step}$$
Which δ???
Set $\hat{x}^{k+1} = x^{k+1}, \hat{f}_{k+1} = f_{k+1}$, $s^{k+1} = S^{k+1} = g^{k+1}$.

Otherwise → nullstep
Compute the convexification parameter

$$\eta_{k+1} = \max\{e_{k+1}, 0\} + \gamma$$

with $e_{k+1} = \hat{f}_{k+1} - f_{k+1} + \langle g^{k+1}, d^k \rangle$.
Set

$$\hat{x}^{k+1} = \hat{x}^k \tag{45}$$

$$\hat{f}^{k+1} = \hat{f}^k \tag{46}$$

$$s^{k+1} = g^{k+1} + \eta_{k+1} * (x^{k+1} - \hat{x}^{k+1}) \tag{47}$$

$$c_{k+1} = \max\{e_{k+1}, 0\} + \frac{\gamma}{2}. \tag{48}$$

Compute the new correction pair $u_1^k = \theta_k d^k$ and $u_2^k = s^{k+1} - \hat{s}^{k+1}$.

6. If this step was a serious step, go to 1.
In case of a null step determine multipliers $\alpha_i^k \geq 0,\ i = \{1, 2, 3\},\ \sum_i \alpha_i^k = 1$ that minimize the function

$$\begin{aligned} \phi(\alpha_1, \alpha_2, \alpha_3) &= (\alpha_1 \hat{s}^{k+1} + \alpha_2 s^{k+1} + \alpha_3 S^k) D_k (\alpha_1 \hat{s}^{k+1} + \alpha_2 s^{k+1} + \alpha_3 S^k) \\ &\quad + 2(\alpha_2 c_{k+1} + \alpha_3 C_k) \end{aligned}$$

where D_k is calculated by the same updating formula as in step 1 and $D_k = D_k + \rho\mathbb{I}$ if $i_C = 1$.
Compute the aggregate subgradient and error

$$S^{k+1} = \alpha_1 \hat{s}^{k+1} + \alpha_2 s^{k+1} + \alpha_3 S^k \tag{49}$$

$$C_{k+1} = (\alpha_2 c_{k+1} + \alpha_3 C_k) \tag{50}$$

and go back to step 1.

4.1 Convergence

Literatur