Contents

1 Introduction

There exists a sound and board theory of classical nonlinear optimization. However, this theory puts strong differentiability requirements on the given problem. Requirements that cannot always be fulfilled in practice. Examples for such practical application reach from problems in physics and mechanical engineering [2] over optimal control problems up to data analysis [1] and machine learning [12]. Other possible fields of applications are risk management and financial calculations [10, 13]. Additionally there exist so called stiff problems that are theoretically smooth but numerically nonsmooth due to rapid changes in the gradient [7].
There exists therefore a need for nonsmooth, that is not necessarily differentiable, optimization algorithms. A lot of the underlying theory and was developed in the 1970's, also driven by the "First World Conference on Nonsmooth Optimization" taking place in 1977 [9]. Now, there exists a well understood theoretical framework of nonsmooth analysis to create the basis for practical algorithms.
The most popular methods to tackle nonsmooth problems at the moment are bundle methods [14]. First developed only for convex functions [6] the method was soon extended to cope also with nonconvex objective functions [8].
Some time later these algorithms were again enhanced to deal with inexact information of the function value, the subgradient or both.
Some natural applications for these cases are derivative free optimization and stochastic simulations [14]. Some more examples from different sources? Bilevel Problems?

The basic idea of bundle methods is to model the original problem by a simpler function, often some sort of stabilized cutting plane model, that is minimized as a subproblem of the algorithm [4].

Adapt this part to what I finally really do:
In this thesis two different types of model functions will be examined that allow the use of inexact information in small to medium-scale problems as well as in large-scale problems. A limited memory approach is examined for the latter case.
what new? Combination of large-scale and inexact information - why needed
don't forget What - why - how

Adapt this part to what I finally really do:
This thesis is organized as follows:
introduction of the most important definitions and results of nonsmooth analysis. Then

the introduction of a very basic bundle algorithm which is then generalized for nonconvex functions with nonsmooth optimization.
Throughout study of this algorithm including comparison to other approaches to tackle inexact information.
Introduction of variable metric (bundle) algorithm to tackle large-scale applications. "discussion" how far this is compatible with inexactness.
Numerical testing
discussion

This thesis is written with the acedemic 'we'.

First a proximal bundle method Difference between different regularizations explained before... large-scale optimization: a metric bundle method instead of a proximal bundle method -> limited memory approach

from PhD-thesis

-

2 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 [9]. To provide an easier understanding of the proximal bundle method in [14] 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 [3]. We focus here on the primal approach.

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 [4] 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 [4].

The optimization problem considered in this section is

$$\min_x f(x) \quad \text{s.t.} \quad x \in X \tag{1}$$

where f is a convex but possibly nondifferentiable function and $X \subseteq \mathbb{R}^n$ is a closed and convex set.

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

$$l_j(x) = f(x^j) + (g^j)^\top (x - x^j) \tag{2}$$

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

$$m_k(x) = \max_{j \in J_k} l_j(x). \tag{3}$$

A new iterate x^{k+1} is found by solving the subproblem

$$\min_x m_k(x) \quad \text{s.t.} \quad x \in X. \tag{4}$$

Picture of function and cutting plane approximation of it

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 we 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:

$$\min \tilde{m}_k(x) = m_k(x) + \frac{1}{2t}\|x - x^k\|^2 \quad \text{s.t.} \quad x \in X \tag{5}$$

This new subproblem is strongly convex and has therefore always a unique solution.

This regularization term can be motivated and interpreted in many different ways, c.f. [4]. From different possible regularization terms the most popular in bundle methods is the penalty-like regularization used here.

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 *model 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 $\hat{x}^k$.

$$\delta_k = f(\hat{x}^k) - m_k(x^{k+1}) \geq 0 \tag{6}$$

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 serious iterate remains the same $\hat{x}^{k+1} = \hat{x}^k$.

Next to the model decrease other forms of decrease measures and variations of these are possible. Some are used in [4, 15].

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

$$\begin{aligned}
l_j(x) &= f(x^j) + {g^j}^\top (x - x^j) && (7)\\
&= f(\hat{x}^k) + {g^j}^\top (x - \hat{x}^k) - (f(\hat{x}^k) - f(x^j) + {g^j}^\top (x^j - \hat{x}^k)) && (8)\\
&= f(\hat{x}^k) + {g^j}^\top (x - \hat{x}^k) - e_j^k && (9)
\end{aligned}$$

where

$$e_j^k := f(\hat{x}^k) - f(x^j) + {g^j}^\top (x^j - \hat{x}^k) \geq 0 \quad \forall j \in J_k \tag{10}$$

is the *linearization error*. Its 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.

Subproblem (5) can now be written as

$$\min_{\hat{x}^k+d\in X} \tilde{m}_k(d) = f(\hat{x}^k) + \max_{j\in J_k}\{{g^j}^\top d - e_j^k\} + \frac{1}{2t_k}\|d\|^2 \tag{11}$$

$$\Leftrightarrow \quad \min_{\substack{\hat{x}^k+d\in X,\\ \xi\in\mathbb{R}}} \xi + \frac{1}{2t_k}\|d\|^2 \quad \text{s.t.} \quad f(\hat{x}^k) + {g^j}^\top d - e_j^k - \xi \leq 0, \quad j\in J_k \tag{12}$$

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 (12) if and only if d^k solves the original subproblem (11) 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 $x^{k+1} = \hat{x}^k + d^k$.

2.1.2 The prox-operator

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

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

Subproblem (5) then writes as

$$\min_{\hat{x}^k+d\in R^n, \xi\in\mathbb{R}} \xi + \mathbb{I}_X + \frac{1}{2t_k}\|d\|^2 \quad \text{s.t.} \quad {g^j}^\top d - e_j^k - \xi \leq 0, \quad j\in J_k \tag{13}$$

check if f also not put into subproblem before

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???

2.1.3 Aggregate objects

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 (13) has for $t_k > 0$ a unique solution given by*

$$d^k = -t_k(G^k + \nu^k), \quad G^k \in \partial m_k(d^k), \quad \nu^k \in \partial \mathbb{I}_X. \tag{15}$$

Furthermore

$$m_k(\hat{x}^k + d) \geq f(\hat{x}^k) + {G^k}^\top d - E_k \quad \forall d \in \mathbb{R}^n \tag{16}$$

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

where

$$E_k := f(\hat{x}^k) - m_k(x^{k+1}) + {G^k}^\top d^k. \tag{17}$$

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

$$\alpha_j^k \geq 0, \quad \sum_{j \in J_k} \alpha_j^k = 1 \tag{18}$$

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

$$E_k = \sum_{j \in J_k} \alpha_j^k e_j^k \qquad \text{and} \qquad G^k = \sum_{j \in J_k} \alpha_j^k g^j. \tag{19}$$

Finally use Lemma ??? in [4]

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

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

$$\delta_k := E_k + t_k\|G^k + \nu^k\|^2 = f(\hat{x}^k) - m_k(x^{k+1}) - {\nu^k}^\top d^k \tag{20}$$

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 $(f_j, g^j, ...)$ are introduced
introduce prox-operator and proximal points

algorithm

Basic bundle method

Select descent parameter $m \in (0, 1)$ and a stopping tolerance `tol` ≥ 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

$$G^k = \sum_{j \in J_k} \alpha_j^k g_j^k, \quad E_k = \sum_{j \in J_k} \alpha_j^k e_j^k, \quad \text{and} \quad \delta_k = E_k + t_k \|G^k + \nu^k\|^2$$

If $\delta_k \leq \texttt{tol} \to$ STOP.

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

4. Compute $f(x^{k+1}),\ g^{k+1}$.
If

$$f^{k+1} \leq \hat{f}^k - m\delta_k \quad \to \text{serious step.}$$

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 $\to$ 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 $\to$ 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. $\to$ include at which point??? This simple idea has however some major drawbacks [5]:

- 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

3 Noll Part

3.1 Introduction

3.2 Keywords

important in Noll for me: optimize model + $d^\top(Q + \frac{1}{t_k}\mathbb{I})d$ -> some kind of second order information

important: $Q + \frac{1}{t_k}\mathbb{I}$ must have all eigenvalues ≥ 0.

idea to get Q: BFGS like in Fin-papers; theory

!!! check stopping criterion connection between d^k and G^k/S^k now: Optimality condition:

$$0 \in \partial M_k(x^{k+1}) + \partial \mathtt{i}_D(x^{k+1}) + \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k \tag{21}$$

$$\Rightarrow \quad S^k(+\nu^k) = -\left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k \tag{22}$$

From this derivation of $\delta_k \to$ nominal (model) decrease:

$$\delta_k = \hat{f}_k - M_k(x^{k+1}) - (\nu^k)^\top d^k \tag{23}$$

$$= \hat{f}_k - A_k(x^{k+1}) - (\nu^k)^\top d^k \tag{24}$$

$$= C_k - (S^k)^\top d^k - (\nu^k)^\top d^k \tag{25}$$

$$= C_k - (S^k + \nu^k)^\top d^k \tag{26}$$

$$= C_k + (d^k)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k \tag{27}$$

3.3 important assumptions

eigenvalues of Q are bounded $\to$ possible by manipulating BFGS update

$$\text{if} \quad norm\left(\frac{y^k {y^k}^\top}{{y^k}^\top d^k}\right) > 10^{??} \tag{28}$$

$$\text{set } \frac{y^k {y^k}^\top}{threshold} \tag{29}$$

$$threshold = norm\left(y^k {y^k}^\top\right)/10^{??} \tag{30}$$

$$\text{end} \tag{31}$$

same procedure for next term; all $< 1/3C$ for some overall threshold C

$Q + \frac{1}{t_k}\mathbb{I}$ such that $\succ \xi\mathbb{I}$ for some fixed $\xi > 0$.

$$\min_{\hat{x}+d\in D} M^k(\hat{x}^k + d^k) + d^\top \frac{1}{2}\left(Q + \frac{1}{t_k}\mathbb{I}\right) d \tag{32}$$

3.4 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 metric matrix $Q = \mathbb{I}$, 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 $k = 1, 2, 3, \ldots$

1. Calculate

$$d^k = \arg\min_{d\in\mathbb{R}^n}\left\{M_k(\hat{x}^k + d) + \mathbb{I}_X(\hat{x}^k + d) + \frac{1}{2}d^\top\left(Q + \frac{1}{t_k}\mathbb{I}\right)d\right\}.$$

2. Set $\rightarrow$ other stopping condition!!!

$$G^k = \sum_{j\in J_k}\alpha_j^k s_j^k, \quad \nu^k = -\frac{1}{t_k}d^k - G^k???????????$$
$$C_k = \sum_{j\in J_k}\alpha_j^k c_j^k$$
$$\delta_k = C_k + (d^k)^\top\left(Q + \frac{1}{t_k}\mathbb{I}\right)d^k$$

   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

$$f^{k+1} \leq \hat{f}^k - m\delta_k \quad \rightarrow \text{ serious step}$$

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

Lemma 5 in [14] stays the same; no Q involved

Theorem 3.1. *Theorem 6 in [14] $\rightarrow$ take only part with $\liminf_{k\to\infty} t_k > 0$ because other one not used in null steps and algorithm this way.*
Let the algorithm generate and infinite number of serious steps. Then $\delta_k \to 0$ as $k \to \infty$. Let the sequence $\{\eta_k\}$ be bounded. If $\liminf_{k\to\infty} t_k > 0$ then as $k \to \infty$ we have $C_k \to 0$, and for ever accumulation point $\bar{x}$ of $\{\hat{x}^k\}$ there exists $\bar{S}$ such that $S^k \to \bar{S}$ and $S^k + \nu^k \to 0$.
In particular if the cardinality of $j \in J^k | \alpha_j^k > 0$ is uniformly bounded in k then the conclusions of Lemma 5 in [14] hold.

The proof is very similar to the one stated in [14] but minor changes have to be made due to the different formulation of the nominal decrease δ_k.

Proof. At each serious step k holds

$$\hat{f}_{k+1} \leq \hat{f}_k - m\delta_k \tag{33}$$

where $m,\ \delta_k > 0$. From this follows that the sequence $\{\hat{f}_k\}$ is nonincreasing. Since $\{\hat{x}^k\} \subset D$ the sequence is by the fact that f is ??????? which assumption says f bounded below??? and $|\sigma_k| < \bar{\sigma}$ the sequence $\{f(\hat{x}^k) + \sigma_k\} = \{\hat{f}_k\}$ is bounded below. Together with the fact that $\{\hat{f}_k\}$ is nonincreasing one can conclude that it converges.
Using (33), one obtains

$$0 \le m \sum_{k=1}^{l} \delta_k \le \sum_{k=1}^{l} \left(\hat{f}_k - \hat{f}_{k+1} \right), \tag{34}$$

so letting $l \to \infty$,

$$0 \le m \sum_{k=1}^{\infty} \delta_k \le \hat{f}_1 - \underbrace{\lim_{k\to\infty} \hat{f}_k}_{\neq \pm\infty}. \tag{35}$$

As a result,

$$\sum_{k=1}^{\infty} \delta_k = \sum_{k=1}^{\infty} \left(C^k + (d^k)^\top \left(Q + \frac{1}{t_k} \mathbb{I} \right) d^k \right) < \infty \tag{36}$$

Hence, $\delta_k \to 0$ as $k \to \infty$. As all quantities above are nonnegative due to positive (semi-)definiteness of $Q + \frac{1}{t_k}\mathbb{I}$, it also holds that

$$C_k \to 0 \quad \text{and} \quad (d^k)^\top \left(Q + \frac{1}{t_k} \mathbb{I} \right) d^k \to 0. \tag{37}$$

For any accumulation point $\bar{x}$ of the sequence $\{\hat{x}^k\}$ the corresponding subsequence $d^k \to 0$ for $k \in K \subset \{1, 2, ...\}$. As $\liminf_{k\to\infty} t_k > 0$ and the eigenvalues of Q are bounded the whole expression

$$S^k + \nu^k = \left(Q + \frac{1}{t_k} I \right) d^k \to 0 \quad \text{for} \quad k \in K. \tag{38}$$

And from local Lipschitz continuity of f follows then that $S^k \to \bar{S}$ for $k \in K$.

$\square$

Remark: If one assumes that the set $\Omega = \{x \in \mathbb{R}^n | f(x) \le f(x^1) + 2\bar{\sigma}\}$ is bounded, it is not necessary to use the constraint set D.
Because all $\{\hat{x}^k\} \subset \Omega$ one can deduce the boundedness of the sequence.

For the case of infinitely many null steps one show:

Theorem 3.2. *[14] Let a finite number of serious iterates be followed by infinite null steps. Let the sequence $\{\eta_k\}$ be bounded and* $\liminf k \to \infty > 0$.
Then $\{x^k\} \to \hat{x}$, $\delta_k \to 0$, $C_k \to 0$, $S^k + \nu^k \to 0$ and there exist $K \subset \{1, 2, ...\}$ and $\bar{S}$ such that $S^k \to \bar{S}^k$ as $K \ni k \to \infty$.
In particular if the cardinality of $j \in J^k | \alpha_j^k > 0$ is uniformly bounded in k then the conclusions of Lemma 5 in [14] hold for $\bar{x} = \hat{x}$.

Proof. Let k be large enough such that $k \geq \bar{k}$ and $\hat{x}^k = \hat{x}$ and $\hat{f}_k = \hat{f}$ are fixed. Define the optimal value of the subproblem (32) by

$$\Psi_k := M_k(x^{k+1}) + \left(d^k\right)^\top \frac{1}{2}\left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k. \tag{39}$$

It is first shown that the sequence $\{\Psi_k\}$ is bounded above. Using the aggregate linearization

$$A_k(\hat{x}) = M_k(x^{k+1}) - \langle S^k, d^k \rangle. \tag{40}$$

Using $S^k + \nu^k = -\left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k$ and the subgradient inequality for $\nu^k \in \partial \mathtt{i}_D$ one obtains

$$\begin{aligned}
\Psi^k + \frac{1}{2}\left(d^k\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k &= A_k(\hat{x}) + \langle S^k, d^k \rangle + \left(d^k\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k \\
&= A_k(\hat{x}) - \langle \nu^k, k \rangle \\
&\leq A(\hat{x}) \\
&\leq M_k(\hat{x}) \\
&= \hat{f}
\end{aligned}$$

where the equations and inequalities follow from???

By boundedness of d^k and $Q + \frac{1}{t_k}\mathbb{I}$ this yields that $\Psi_k \leq \hat{f}$, so the sequence $\{\Psi_k\}$ is bounded above. In the next step is shown that $\{\Psi_k\}$ is increasing.

$$\begin{aligned}
\Psi_{k+1} &= M_k(x^{k+2}) + \frac{1}{2}\left(d^{k+1}\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^{k+1} && (41)\\
&\geq A_k(x^{k+2}) + \frac{1}{2}\left(d^{k+1}\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^{k+1} && (42)\\
&= M_k(x^{k+1}) + \langle S^k, x^{k+2} - x^{k+1} \rangle + \frac{1}{2}\left(d^{k+1}\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^{k+1} && (43)\\
&= \Psi_k - \frac{1}{2}\left(d^k\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k + \frac{1}{2}\left(d^{k+1}\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^{k+1} && (44)\\
&\qquad - \left(d^k\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right)\left(d^{k+1} - d^k\right) - \langle \nu^k, x^{k+2} - x^{k+1} \rangle && (45)\\
&\geq \Psi_k + \frac{1}{2}\left(d^{k+1} - d^k\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right)\left(d^{k+1} - d^k\right) && (46)
\end{aligned}$$

say where (in-)equalities come from

As Q is fixed in null steps and $\liminf_{k\to\infty} t_k > 0$ $\{\Psi_k\}$ is increasing. The sequence is therefore convergent. Consequently, taking into account that $1/t_k \geq 1/t_{\bar{k}}$, it follows

$$\|d^{k+1} - d^k\| \to 0, \quad k \to \infty. \tag{47}$$

By the definitions and characterizations that have to be specified one has

$$\hat{f} = \delta_k + M_k(\hat{x}) - C_k - \left(d^k\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right)\left(d^k\right) \tag{48}$$

$$= \delta_k + M_k(x^{k+1}) - \langle S^k, d^k \rangle - \left(d^k\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right)\left(d^k\right) \tag{49}$$

$$= \delta_k \qquad \geq \delta_k + M_k(\hat{x} + d^k) \tag{50}$$

Where the last inequality is given by $\nu^k \in \partial \mathtt{i}_D(x^{k+1})$. Therefore

$$\delta^{k+1} \leq \hat{f} - M_{k+1}(\hat{x} + d^{k+1}). \tag{51}$$

By the first inequality in assumption define assumption on the model, written for $d = d^{k+1}$,

$$-\hat{f}_{k+1} + c_{k+1}^{k+1} - \left\langle s_{k+1}^{k+1}, d^{k+1} \right\rangle \geq -M_{k+1}(\hat{x} + d^{k+1}). \tag{52}$$

As $\hat{f}_{k+1} = \hat{f}$, adding condition ??? to the inequality above, one obtains that

$$m\delta_k + \left\langle s_{k+1}^{k+1}, d^k - d^{k+1} \right\rangle \geq \hat{f} - M_{k+1}(\hat{x} + d^{k+1}). \tag{53}$$

Combining this relation with ??? yields

$$0 \leq \delta_{k+1} \leq m\delta_k + \left\langle s_{k+1}^{k+1}, d^k - d^{k+1} \right\rangle. \tag{54}$$

Since $m \in (0,1)$ and $\left\langle s_{k+1}^{k+1}, d^k - d^{k+1} \right\rangle \to 0$ as $k \to \infty$ due to (47) and the boundedness of $\{\eta_k\}$ using [11, Lemma 3, p.45] it follows from (54) that

$$\lim_{k\to\infty} \delta_k = 0. \tag{55}$$

From the formulation $\delta_k = C_k + \left(d^k\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k$ follows that $C_k \to 0$ as $k \to \infty$. As $Q + \frac{1}{t_k}\mathbb{I} \succ \xi\mathbb{I}$ it follows that

$$\xi \left(d^k\right)^\top d^k \leq \left(d^k\right)^{top} \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k \to 0 \tag{56}$$

□

References

[1] Adil Bagirov, Napsu Karmitsa, and Marko M. Mäkelä. *Introduction to Nonsmooth Optimization: Theory, Practice and Software.* Springer International Publishing Switzerland, 2014.

[2] Frank H. Clarke. *Optimization and nonsmooth analysis.* Classics in Applied Mathematics. Society for Industrial and Applied Mathematics Philadelphia, 1990.

[3] Warren Hare and Claudia Sagastizàbal. A redistributed proximal bundle method for nonconvex optimization. *SIAM Journal on Optimization*, 20(5):2442–2473, 2010.

[4] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms II*, volume 306 of *Grundlehren der mathematischen Wissenschaften.* Springer Berlin Heidelberg, 1993.

[5] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms I*, volume 305 of *Grundlehren der mathematischen Wissenschaften.* Springer Berlin Heidelberg, 2 edition, 1996.

[6] Claude Lemaréchal. Nonsmooth optimization and descent methods. Iiasa research report, International Institute for Applied Systems Analysis, 1978.

[7] Marko M. Mäkelä and Pekka Neittaanmäki. *Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control.* World Scientific Pub Co Pte Lt, 1992.

[8] Robert Mifflin. A modification and an extension of lemaréchal's algorithm for nonsmooth minimization. In *Mathematical Programming Studies*, volume 17, pages 77–90. Springer Nature, 1982.

[9] Robert Mifflin and Claudia Sagastizàbal. A science fiction story in nonsmooth optimization originating at iiasa. *Documenta Mathematica*, Extra Volume ISMP:291–300, 2012.

[10] Yurii Nesterov and Vladimir Shikhman. Algorithmic principle of least revenue for finding market equilibria. In Boris Goldengorin, editor, *Optimization and Its Applications in Control and Data Sciences*, volume 115 of *Springer Optimization and Its Applications*, pages 381–435. Springer Nature, 2016.

[11] Boris T. Polyak. *Introduction to Optimization.* Optimization Software , Inc., Publications Division, New York, 1987.

[12] A. J. Smola, S.v.n. Vishwanathan, and V. Le Quoc. Bundle methods for machine learning. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems*, number 20, 2007.

[13] Choon Hui Teo, A. J. Smola, S.V.N. Vishwanathan, and Quoc V. Le. Bundle methods for regulized risk minimization. 2010.

[14] Mikhail Solodov Warren Hare, Claudia Sagastizàbal. A proximal bundle method for nonsmooth nonconvex functions with inexact information. *Computational Optimization and Applications*, 63:1–28, 2016.

[15] Claude Lemaréchal Welington de Oliveira, Claudia Sagastizàbal. Convex proximal bundle methods in depth: a unified analysis for inexact oracles. *Mathematical Programming*, 148:241–277, 2014.