Contents

1 A Basic Bundle Method

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 [16]. To provide an easier understanding of the proximal bundle method in [30] 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 [6]. We focus here on the primal approach.

notation, definitions

already done in previous preliminaries chapter?

This section gives a short summery of the derivations and results of chapter XV in [8] 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 [8].

The optimization problem considered in this section is

$$\min_x f(x) \quad \text{s.t.} \quad x \in X \tag{1.1}$$

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

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

$$l_j(x) = f(x^j) + (g^j)^\top (x - x^j) \tag{1.2}$$

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

$$m_k(x) = \max_{j \in J_k} l_j(x). \tag{1.3}$$

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

$$\min_x m_k(x) \quad \text{s.t.} \quad x \in X. \tag{1.4}$$

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:

$$\min \tilde{m}_k(x) = m_k(x) + \frac{1}{2t_k}\|x - x^k\|^2 \quad \text{s.t.} \quad x \in X,\ t_k > 0. \tag{1.5}$$

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. [8]. 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 $\hat{x}^k$.

$$\delta_k = f(\hat{x}^k) - m_k(x^{k+1}) \geq 0 \tag{1.6}$$

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 [8, 31].

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 $\hat{x}^k$.

$$\begin{aligned} l_j(x) &= f(x^j) + {g^j}^\top (x - x^j) && (1.7)\\ &= f(\hat{x}^k) + {g^j}^\top (x - \hat{x}^k) - (f(\hat{x}^k) - f(x^j) + {g^j}^\top (x^j - \hat{x}^k)) && (1.8)\\ &= f(\hat{x}^k) + {g^j}^\top (x - \hat{x}^k) - e_j^k && (1.9) \end{aligned}$$

where

$$e_j^k := f(\hat{x}^k) - f(x^j) + {g^j}^\top (x^j - \hat{x}^k) \geq 0 \quad \forall j \in J_k \tag{1.10}$$

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 (1.5) can now be written as

$$\min_{\hat{x}^k + d \in X} \tilde{m}_k(\hat{x}^k + d) = f(\hat{x}^k) + \max_{j \in J_k} \{{g^j}^\top d - e_j^k\} + \frac{1}{2t_k}\|d\|^2 \tag{1.11}$$

$$\Leftrightarrow \quad \min_{\substack{\hat{x}^k+d\in X,\\ \xi\in\mathbb{R}}} \xi + \frac{1}{2t_k}\|d\|^2 \quad \text{s.t.} \quad f(\hat{x}^k) + {g^j}^\top d - e_j^k - \xi \leq 0, \quad j \in J_k \tag{1.12}$$

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 (1.12) if and only if

$$d^k \text{ solves the original subproblem (1.11) and} \tag{1.13}$$

$$\xi_k = \max_{j\in J_k} {g^j}^\top d^k - e_j^k = m_k(\hat{x}^k + d^k) - f(\hat{x}^k). \tag{1.14}$$

The new iterate is then given by $x^{k+1} = \hat{x}^k + d^k$.

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

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

This function is convex if and only if the set X is convex [23].

Subproblem (1.5) then writes with respect to the serious point $\hat{x}^k$

$$\min_{x\in\mathbb{R}^n} m_k(x) + \mathtt{i}_X(x) + \frac{1}{2t_k}\|x - \hat{x}^k\|^2. \tag{1.15}$$

The subproblem is now written as the *Moreau-Yosida regularization* of $\check{f} := m_k(x) + \mathtt{i}_X(x)$. The emerging mapping is also known as *proximal point mapping* [6] or *prox-operator*

$$prox_{t,f}(x) = \arg\min_{y\in\mathbb{R}^n} \left\{ \check{f}(y) + \frac{1}{2t}\|x - y\|^2 \right\}, \quad t > 0. \tag{1.16}$$

This special form of the subproblems gives the primal bundle method its name, *proximal bundle method*. The mapping also plays a key role when the method is generalized to nonconvex objective functions and inexact information.

1.3 Aggregate Objects

Look again at a slightly different formulation of the bundle subproblem

$$\min_{\substack{d\in\mathbb{R}^n,\\ \xi\in\mathbb{R}}} \quad \xi + \mathtt{i}_X + \frac{1}{2t_k}\|d\|^2 \tag{1.17}$$

$$\text{s.t.} \quad {g^j}^\top d - e_j^k - \xi \leq 0, \quad j \in J_k. \tag{1.18}$$

As the objective function is still convex (X is a convex set) the following Karush-Kuhn-Tucker (KKT) conditions have to be valid for the minimizer $\left(\xi_k, d^k\right)$ of the above subproblem [9] assuming a constraint qualification if the constraint set X makes it necessary [27].

There exist a subgradient $\nu^k \in \partial \mathtt{i}_X$ and Lagrangian multipliers $\alpha_j,\ j \in J^k$ such that

$$0 = \nu^k + \frac{1}{t_k} d^k + \sum_{j\in J^k} \alpha_j g^j \tag{1.19}$$

$$\sum_{j\in J_k} \alpha_j = 1, \tag{1.20}$$

$$\alpha_j \geq 0,\ j \in J^k, \tag{1.21}$$

$${g^j}^\top d^k - e_j^k - \xi_k \leq 0, \tag{1.22}$$

$$\sum_{j\in J^k} \alpha_j \left({g^j}^\top d^k - e_j^k - \xi_k\right) = 0. \tag{1.23}$$

From condition (1.19) follows then

$$d^k = t_k \left(G^k + \nu^k\right) \quad \text{with} \quad G^k := \sum_{j\in J^k} \alpha_j g^j \in \partial m_k(x^{k+1}) \tag{1.24}$$

with the *aggregate subgradient* G^k.

Rewriting condition (1.23) yields the *aggregate error*

$$E_k := \sum_{j\in J^k} \alpha_j e_j^k = (G^k)^\top d^k + f(\hat{x}^k) - m_k(x^{k+1}). \tag{1.25}$$

Here relation (1.14) was used to replace ξ_k.

The aggregate subgradient and error are used to formulate an implementable stopping condition for the bundle algorithm. The motivation behind that becomes clear with the following lemma.

Lemma 1.1. *[5, Theorem 6.68, p.387] Let $X = \mathbb{R}^n$. Let $\varepsilon > 0$, $\hat{x}^k \in \mathbb{R}^n$ and $g^j \in \partial f(x^j)$ for $j \in J^k$. Then the set*

$$\mathcal{G}_\varepsilon^k := \left\{ \sum_{j \in J^k} \alpha_j g^j \mid \sum_{j \in J^k} \alpha_j e_j \varepsilon, \sum_{j \in J^k} \alpha_j = 1, \alpha_j \geq 0, j \in J^k \right\}$$

is a subset of the ε-subdifferential of $f(\hat{x}^k)$

$$\mathcal{G}_\varepsilon^k \subseteq \partial_\varepsilon f(\hat{x}^k)$$

.

This means that at least in the unconstrained case $G^k \in \partial_{E_k} f(\hat{x}^k)$. So driving $\|G^k\|$ and E_k close to zero results in some approximate ε-optimality of the objective function. In the constraint case the stopping condition is written as

$$\delta_k = E^k + t_k \|G^k + \nu^k\|^2. \tag{1.26}$$

δ_k is the same measure that is also taken for the decrease test. The relation

$$\begin{aligned} \delta_k &= E^k + t_k \|G^k + \nu^k\|^2 && (1.27)\\ &= E^k - \langle G^k, d^k \rangle - \langle \nu^k, d^k \rangle && (1.28)\\ &= f(\hat{x}^k) - m_k(x^{k+1}) - \langle \nu^k, d^k \rangle && (1.29)\\ & && (1.30) \end{aligned}$$

where (1.24) and (1.25) were used, shows that the new δ_k is only a little variation of the model decrease. If the iterate x^{k+1} does not lie on the edge of the constraint set X, the vector $\nu^k = 0$ and the expression simplifies to the one stated in (1.6).

A totally different use of the aggregate objects was proposed by Kiwiel in [12]. The aggregate subgradient can be used to build the *aggregate linearization*

$$a_k(\hat{x}^k + d) := m_k(x^{k+1}) + \langle G^k, d - d^k \rangle. \tag{1.31}$$

This function can be used to avoid memory problems as it compresses the information of all bundle elements into one affine plane. Adding the function a_k to the cutting plane model preserves all assumptions put on the model and can therefore be used instead of or in combination with the usual cutting planes.

For the model update only the following two conditions are assumed to be fulfilled in consecutive null steps:

$$\begin{aligned} m_{k+1}(\hat{x}^k + d) &\geq f(\hat{x}^{k+1}) - e_{k+1}^{k+1} + \left\langle g^{k+1}, d \right\rangle \\ m_{k+1}(\hat{x}^k + d) &\geq a_k(\hat{x}^k + d) \end{aligned} \tag{1.32}$$

The first condition means, that the newly computed information in always put into the bundle. The second one is important when updating the bundle index set J^k. It holds trivially if no or only inactive information j with $\alpha_j = 0$ is removed.

It is also always satisfied if the aggregate linearization a_k itself is added to the bundle. In this case active information can be removed without violating the condition. This is the key idea of Kiwiel's aggregation technique and ensures that the set $\{j \in J^k | \alpha_j > 0\}$ can be bounded. It can however impair the speed of convergence if the bundle is kept too small and provides hence less information about the objective function [2].

The following basic bundle algorithm can now be stated:

algorithm

Basic bundle method

Select a 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
$$d^k = \arg\min_{d \in \mathbb{R}^n} m_k(\hat{x}^k + d) + \texttt{i}_X(\hat{x}^k + d) + \frac{1}{2t_k}\|d\|^2$$
and the corresponding Lagrange multiplier $\alpha_j^k,\ j \in J_k$.

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

3. Set $x^{k+1} = \hat{x}^k + d^k$.
4. Compute $f(x^{k+1}),\ g^{k+1}$.
   If
   $$f(x^{k+1}) \leq f(\hat{x}^k) - m\delta_k \quad \rightarrow \text{serious step.}$$
   Set $\hat{x}^{k+1} = x^{k+1},\ f(\hat{x}^{k+1}) = f(x^{k+1})$ and select a suitable $t_{k+1} > 0$.
   Otherwise
   $$\rightarrow \text{ nullstep.}$$
   Set $\hat{x}^{k+1} = \hat{x}^k,\ f(\hat{x}^{k+1}) = f(x^{k+1})$ and choose $t_{k+1} > 0$ in a suitable way.
5. Select the new bundle index set J_{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 it is not specified how to update the parameter t_k, the index set J^k and the model m_k. For the convergence proof it is only necessary that $\liminf_{k\to\infty} t_k > 0$ and that conditions (1.32) are fulfilled.

In practice the choice of t_k can be realized by taking

$$t_{k+1} = \kappa_+ t_k, \quad \kappa_+ > 1 \tag{1.33}$$

at every serious step and

$$t_{k+1} = \max\{\kappa_- t_k, t_{min}\}, \quad \kappa_- < 1 \text{ and } t_{min} > 0 \tag{1.34}$$

at every null step. The idea behind this management of t_k is taken from the trust region method: If the computed iterate was good, the model is assumed to be reliable in a bigger area around this serious iterate so bigger step sizes are allowed. If a null step was taken, the model seems to be too inaccurate far from the current serious point. Then smaller step sizes are used. A more sophisticated version of this kind of step size management is also used by Noll et al. in [20] and [18]. The trust region idea was very much exploited by Schramm and Zowe in [24].

For a convergence proof of this basic bundle method c.f [8].

2 Variations of The Bundle Method

After their discovery in 1975 bundle methods soon became very successful. Only a few years later they were generalized to be used also with nonconvex objective functions. Early works, that contain fundamental ideas still used for these algorithms are [15] and [11]. It then took over 25 years that bundle methods were again generalized to the use of inexact information. The first works on this subject being [7, 13] and [26].

This section of the thesis shortly presents the key ideas of those two kinds of generalizations and different types of bundle methods that realize them. This is first done for the case of convex objective functions with inexact function value and/or subgradient information and then for nonconvex objective functions.

2.1 Convex Bundle Methods with Inexact Information

We focus here on *convex* bundle methods with inexact information. The reason for this is that there is a fundamental difference in treating inexactness between methods that assume convex and those that assume nonconvex objective functions. When dealing with nonconvex objective functions inexactness is treated as some additional nonconvexity. This is not possible if the convexity property is to be exploited for better convergence results. A throughout study on this subject including a synthetic convergence theory is done in [31]. Here the most important aspects of this paper are reviewed.

2.1.1 Different Types of Inexactness

At first it is important to notice that there are many different types of inexactness and further assumptions can be put on the noise to reach stronger convergence results.

Generally inexact information for convex objective functions is defined in the following way:

$$\begin{aligned} f_x = f(x) - \sigma_x, \quad \sigma_x \leq \bar{\sigma} \\ g_x \in \mathbb{R}^n \text{ such that } f(\cdot) \geq f_x + \langle g_x, \cdot - x \rangle - \theta_x, \quad \theta_x \leq \bar{\theta}. \end{aligned} \tag{2.1}$$

From this follows with

$$f(\cdot) \geq f(x) + \langle g_x, \cdot - x \rangle - (\sigma_x + \theta_x) \tag{2.2}$$

that g_x is an ε-subgradient with $\varepsilon = \sigma_x + \theta_x \geq 0$ independently of the signs of the errors.

Two very interesting cases of more sophisticated errors are *asymptotically vanishing errors* and oracles that only underestimate the true function, so called *lower oracles*.

The case of asymptotically vanishing errors $\lim_{k\to\infty} \sigma_k = 0$ and $\lim_{k\to\infty} \theta_k = 0$ is important because it allows convergence to the exact minimum of the problem even if function values and subgradients are erroneous. In the case of $\bar{\theta} = 0$ it even suffices to show that the errors are asymptotically exact only for descent steps [10]. This observation was the motivation for the partly inexact bundle methods presented in [10] and [31]. The idea is to calculate a value of the objective function with a demanded accuracy (which is finally going to be exact) only if a certain target descent γ_x is reached. This approach can save a lot of (unnecessary) computational effort while still enabling convergence to the exact minimum c.f. [31].

Lower oracles provide f_x and g_x such that $f_x \leq f(x)$ and $f(\cdot) \geq f_x + \langle g_x, \cdot - x \rangle$. That means the cutting plane model is always minorizing the true function as it is the case in for exact information. In this case if the value to approximate the optimal function value is chosen properly, it is not necessary to include any other steps to cope with the inexactness, such as noise attenuation [31, Corollary 5.2].

2.1.2 Noise Attenuation

In the case of inexact information, especially if the inexact function value can overestimate the real one, it is possible that the aggregate linearization error E_k becomes very small (or even negative) even though the current iterate is far from the minimum of the objective function. To tackle this problem the authors propose a procedure called *noise attenuation* that was developed in [7] and [13]. The basic idea is to allow bigger step sizes t_k whenever the algorithm comes in the situation described above. This ensures that either some significant descent towards the real minimum can be done or shows that the point where the algorithm is stuck is actually such a minimum. Noise attenuation is triggered when E_k or respectively the descent δ_k that is used for the descent test is negative.

2.1.3 Convergence Results

Depending on the kind of error many slightly different convergence results can be proven for bundle methods that handle convex objective functions with inexact information. In case of the general error on the function value and subgradient it can be shown that for bounded sequences $\hat{x}^k$ every accumulation point $\bar{x}$ of an infinite series of serious steps or the last serious iterate before an infinite tail of null steps is a $\bar{sigma} + \bar{\theta}$-solution of the problem meaning that

$$f(\bar{x}) \leq f^* + \bar{\sigma} + \bar{\theta} \tag{2.3}$$

with f^* being the exact minimum of the objective function. Generally for asymptotically vanishing errors it is possible to show that the method converges to the exact minimum of the problem. For more detailed results refer to [31].

2.2 Nonconvex Bundle Methods with Exact Information

In the case of inexactness in convex bundle methods, where a lot of different assumptions can be put on the errors to reach different convergence results, the strategy to cope with these errors remains very much the same. In contrast to this in case of nonconvex objective functions the case to be studied is rather uniform still there exist very different approaches to tackle the problem. 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* [12, 15]. Here the linearization error is essentially replaced by the nonnegative number

$$\tilde{e}_j^k := \max_{j \in J_k} \{|e_j^k|, \gamma \|\hat{x}^k - x^j\|^2\} \tag{2.4}$$

or a variation of this expression.

The expression gradient locality measure comes from the dual point of view, where the aggregate linearization error provides a measure for the 'distance' of the calculated ε-subgradient to the objective function.

Methods that use downshifting for building the model function are often endowed with a line search to provide sufficient decrease of the objective function. For the linesearch to

terminate finitely, usually semismoothness of the objective function is needed.

Instead of using line search it is also possible to do *proximity control*. This means that the step size parameter t_k is managed in a smart way to ensure the right amount of decrease in the objective function. This method is very helpful in the case of nonconvex objective functions with inexact information as it is predominantly considered in this thesis.

As inexactness can be seen as a kind of slight nonconvexity one could be tempted to think that nonconvex bundle methods are destined to be extended to the inexact case. Indeed, the two existing algorithms [30, 18] that deal with both nonconvexity and inexactness are both extensions of a nonsmooth bundle method. This is however seldom possible for algorithms that employ a line search because for functions with inexact information convergence of this subroutine cannot be proven.

To this end proximity control seems to be a very promising strategy. It is used in many different variations in [1, 14, 17, 19, 20] and [25].

In the beginning bundle methods were mostly explored from the dual point of view. Newer concepts focus also on the primal version of the method. This invokes for example having different model functions for the subproblem.

In [3, 4] the difference function

$$h(d) := f(x^j + d) - f(x^j) \quad j \in J_k \tag{2.5}$$

is approximated to find a 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 new iterates.

In [20] 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:

$$\Phi(y, x) = \phi(y, x) + \frac{1}{2}(y - x)^\top Q(x)(y - x) \tag{2.6}$$

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.

Instead of a line search is uses proximity control to ensure convergence.

Generally for all of this methods convergence to a stationary point is established under the assumptions of a locally Lipschitz objective function and bounded level sets $\{x \in \mathbb{R}^n | f(x) \leq f(\hat{x}^1)\}$. If the method uses a line search additionally semismoothness of the objective function is needed.

In [18] the second order approach of [20] is extended to functions with inexact information. As far a we know this is the only other bundle method that can deal with nonconvexity and inexactness in both the function value and subgradient. It inspires the variable metric variation of the method used by Hare et al. in [30] that is presented in section 4 of this thesis.

3 Proximal bundle method for nonconvex functions with inexact information

introduction

This section focuses on the proximal bundle method presented by Hare et al. in [30]. 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 by Hare and Sagastizábal in [6]: 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.

3.1 Derivation of the Method

"assumptions and notations"

Throughout this section we consider now the optimization problem

$$\min_x f(x) \quad \text{s.t.} \quad x \in X. \tag{3.1}$$

The objective function $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz and (subdifferentially) regular. $X \subseteq \mathbb{R}^n$ is assumed to be a convex compact set.

Definition 3.1. [22] $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ 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 $\bar{x}, f(\bar{x})$.

Both the function value as well as one element of the subdifferential can be provided in an inexact form.

For the function value inexactness is defined straight forwardly: If

$$\|f - f(x)\| \leq \sigma \tag{3.2}$$

then f approximates the value $f(x)$ within σ.

For the subgradients we adopt the notation used in [18] and interpret inexactness in the following way: $g \in \mathbb{R}^n$ approximates a subgradient of $\partial f(x)$ within $\theta \geq 0$ if

$$g \in \partial f(x) + B_\theta(0) := \partial_{[\theta]} f(x) \tag{3.3}$$

where $\partial f(x)$ is the *Clarke subdifferential* of f.

The given definition of the inexactness can be motivated by the relation

$$g \in \partial_{[\theta]} f(x) \Leftrightarrow g \in \partial(f + \theta\| \cdot -x\|)(x)$$

noticed in [28]. It means that the approximation of the subgradient of $f(x)$ is an exact subgradient of a small perturbation of f at x.

Remark: For convex objective functions this approximate subdifferential does not equal the usual convex ε-subdifferential. The two can however be related via

$$\partial_\theta f(x) \subset \partial_{[\theta']} f(x)$$

for a suitable θ'. Generally an explicit relation between θ and θ' is hard to find [18].

Like in the paper it is assumed that the errors are bounded although the bound does not have to be known:

$$|\sigma_j| \leq \bar{\sigma} > 0 \quad \text{and} \quad 0 \leq \theta_j \leq \bar{\theta} \quad \forall j \in J^k. \tag{3.4}$$

In the context of inexact information it is important to make a distinction between the (unknown) exact function value and its approximation. Throughout this section we 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.

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 [29] not the objective function but a convexified version of it is modeled as the objective function of the subproblem.

As already pointed out in 1.2 the bundle subproblem can be formulated by means of the prox-operator (1.16).

The key idea is now to use the relation

$$prox_{T=\frac{1}{\eta}+t,f}(x) = prox_{t,f+\eta/2|\cdot-x|^2}(x). \tag{3.5}$$

This means, that the proximal point of the function f for parameter $T = \frac{1}{\eta} + t$ is the same as calculating the proximal point of the convexified function

$$\tilde{f}(y) = f(y) + \frac{\eta}{2}|y - x|^2 \tag{3.6}$$

with respect to the parameter t [6]. η is therefore called the *convexification parameter* and t is the *prox-parameter*.

The main difference of the method in [30] to the basic bundle method is that the function that is modeled by the cutting plane model s no longer the original objective function f but the convexified version $\tilde{f}$. This results in the following changes:

The linear functions forming the model have a tilted slope, here called the *augmented subgradient* at the iterate x^j

$$s_j^k = g^j + \eta_k \left(x^j - \hat{x}^k\right). \tag{3.7}$$

Additionally they are shifted downwards to keep the linearization error nonnegative. The *augmented linearization error* is therefore defined as

$$0 \leq c_j^k := e_j^k + b_j^k, \quad \text{with} \quad \begin{cases} e_j^k := \hat{f}_k - f_j - \langle g^j, \hat{x}^k - x^j \rangle \\ b_j^k := \frac{\eta_k}{2} \|x^j - \hat{x}^k\|^2 \end{cases} \tag{3.8}$$

and

$$\eta_k \geq \max \left\{ \max_{j \in J_k, x^j \neq \hat{x}^k} \frac{-2e_j^k}{\|x^j - \hat{x}^k\|^2}, 0 \right\} + \gamma. \tag{3.9}$$

The parameter $\gamma \geq 0$ is a safeguarding parameter to keep the calculation numerically stable.

The new model function can therefore be written as

$$M_k(\hat{x}^k + d) := \hat{f}_k + \max_{j \in J_k} \left\{ {s_j^k}^\top d - c_j^k \right\} \tag{3.10}$$

The definition of the aggregate objects follows straightforwardly:

$$S^k := \sum_{j \in J_k} \alpha_j^k s_j^k, \tag{3.11}$$

$$C_k := \sum_{j \in J_k} \alpha_j^k c_j^k \text{ and } A_k(\hat{x}^k + d) := M_k(x^{k+1}) + \left\langle S^k, d - d^k \right\rangle. \tag{3.12}$$

Just as the model decrease

$$\delta^k := C_k + t_k \|S^k + \nu^k\|^2. \tag{3.13}$$

A bundle algorithm that deals with nonconvexity and inexact function and subgradient information can therefor be stated.

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

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

2. Set

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

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

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

Set $\hat{x}^{k+1} = x^{k+1}, \hat{f}^{k+1} = f^{k+1}$ and select $t_{k+1} > 0$.
Otherwise

$$\rightarrow \text{nullstep}$$

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}, 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

The update of t_k can be done in the same way described for the basic bundle method. Similarly the methods to update the bundle index set J^k stay valid. The update conditions (1.32) for the model are now written with respect to the augemented aggregate linearization and the approximate function value $\hat{f}_{k+1}$.

$$\begin{aligned} M_{k+1}(\hat{x}^k + d) &\geq \hat{f}_{k+1} - c_{k+1}^{k+1} + \left\langle s^{k+1}, d \right\rangle \\ M_{k+1}(\hat{x}^k + d) &\geq A_k(\hat{x}^k + d). \end{aligned} \tag{3.14}$$

remark on η_k? how does it behave in my applications???

3.2 On Different Convergence Results

In terms of usability of the described algorithm it is interesting to see if stronger convergence results are possible if additional assumptions are put on the objective function. This is investigated in the following section.

3.2.1 The Constraint Set

The method above can handle nonconvex objective functions. It is not explicitly assumed, that the function taken has a finite minimum. Therefore the constraint set X has to ensure the boundedness of the sequence $\{\hat{x}^k\}$. This is not necessary if the objective function is assumed to have bounded level sets $\{x \in \mathbb{R}^n | f(x) \leq f(\hat{x}^1)\}$, an assumption commonly used when optimizing nonconvex functions. As the objective function is assumed to be continuous bounded level sets are compact. Additionally the descent test makes sure that $f(\hat{x}^{k+1}) \leq f(\hat{x}^k)$ for all k. The proof holds therefore in the same way as with the set X.

3.2.2 Exact Information and Vanishing Errors

As the presented algorithm was originally designed for nonconvex objective functions where function values as well as subgradients are available in an exact manner, all convergence results stay the same with the error bounds $\bar{\sigma} = \bar{\theta} = 0$. As already indicated previously this is the case because inexactness can be seen as a kind of nonconvexity and no additional concepts had to be added to the method when generalizing it to the inexact setting.

If we additionally require the objective function to be lower-$\mathcal{C}^2$ it can be proven that the sequence $\{\eta_k\}$ is bounded [6]. This is not possible in the case of inexact information even for convex objective functions.

For asymptotically vanishing errors, meaning $\lim_{k\to\infty} \sigma_k = 0$ and/or $\lim_{k\to\infty} \theta_k = 0$ the convergence theory holds equally well with error bounds $\bar{\sigma} = \bar{\theta} = 0$ in [30, Lemma 5]. Still it is difficult if not impossible to show that the sequence $\{\eta_k\}$ is bounded without further assumptions. Under the assumption that f is lower-$\mathcal{C}^2$ and some continuity bounds on the errors

$$\frac{|\sigma_j - \hat{\sigma}_k|}{\|x^j - \hat{x}^k\|^2} \leq L_\sigma, \qquad \frac{\theta_j}{\|x^j - \hat{x}^k\|} \leq L_\theta \quad \forall k \text{ and } \forall j \in J_k$$

boundedness of the sequence $\{\eta_k\}$ can be shown. The question remains however if those assumptions are possible to be assured in practice.

3.2.3 Convex Objective Functions

convex objective function now Let us first recall that as laid down previously the definition of an approximate subgradient does not coincide with the usual ε-subdifferential for convex functions. It is therefore assumed that for convex objective functions the approximate subgradient g^j is taken from the θ_j-subdifferential with $\theta_j < \bar{\theta}$ for all $j \in J$.

This means we consider now inexactness of the following type

$$f_j = f(x^j) - \sigma_j, \sigma_j \leq \bar{\sigma} \tag{3.15}$$

$$g^j \in \mathbb{R}^n \text{ such that } f(\cdot) \geq f^j + \langle g^j, \cdot - x^j \rangle - \theta_j, \ \forall x^j \in \mathbb{R}^n, \theta_j \leq \bar{\theta} \tag{3.16}$$

- convex objective function
  $\rightarrow$ generally better convergence properties possible
  but more or less only on error bound??? $\rightarrow$ different concept of algorithm for convex inexact functions to exploit convexity (contrary to nonconvex obj functions)

To conclude this section we can say: At the moment there exist two fundamentally different approaches to tackle inexactness in various bundle methods depending on if the method is developed for convex or nonconvex objective functions. In the nonconvex case inexactness is only considered in the paper by Hare, Sagastizàbal and Sodolov [30] presented above and Noll [18]. In these cases the inexactness can be seen as an additional nonconvexity. In practice this means that the algorithm can be taken from the nonconvex case with no or only minor changes. This includes that all results of the exact case remain true as soon as function and subgradient are evaluated in an exact way. In case of convex objective functions with inexact information stronger convergence results are possible. However to be able to exploit convexity in order to achieve those results the algorithms look quite different from those designed for nonconvex objective functions and are generally not able to deal with such functions.

- stronger convergence results possible because of exploitation of convexity
- changes in the algorithm because if convexity should be exploited: inexactness cannot be treated as nonconvexity
-
- *inexact*: $0 \in (\partial f(\bar{x}) + \partial I_D(\bar{x})) + B_{\bar{\theta}}(0)$

if f lower-$\mathcal{C}^1$:

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

4 Noll Part

4.1 Introduction

4.2 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:

$$0 \in \partial M_k(x^{k+1}) + \partial \mathtt{i}_D(x^{k+1}) + \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k \tag{4.1}$$

$$\Rightarrow \quad S^k(+\nu^k) = -\left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k \tag{4.2}$$

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

$$\begin{aligned}
\delta_k &= \hat{f}_k - M_k(x^{k+1}) - (\nu^k)^\top d^k && (4.3)\\
&= \hat{f}_k - A_k(x^{k+1}) - (\nu^k)^\top d^k && (4.4)\\
&= C_k - (S^k)^\top d^k - (\nu^k)^\top d^k && (4.5)\\
&= C_k - (S^k + \nu^k)^\top d^k && (4.6)\\
&= C_k + (d^k)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k && (4.7)
\end{aligned}$$

4.3 important assumptions

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

$$\text{if} \quad norm\left(\frac{y^k {y^k}^\top}{{y^k}^\top d^k}\right) > 10^{??} \tag{4.8}$$

$$\text{set } \frac{y^k {y^k}^\top}{threshold} \tag{4.9}$$

$$threshold = norm\left(y^k {y^k}^\top\right)/10^{??} \tag{4.10}$$

$$\text{end} \tag{4.11}$$

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 $\xi > 0$.

$$\min_{\hat{x}+d\in D} M^k(\hat{x}^k + d^k) + d^\top \frac{1}{2}\left(Q + \frac{1}{t_k}\mathbb{I}\right) d \tag{4.12}$$

4.4 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 → 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 [30] stays the same; no Q involved

Theorem 4.1. *Theorem 6 in [30] $\rightarrow$ 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 [30] hold.*

The proof is very similar to the one stated in [30] 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

$$\hat{f}_{k+1} \leq \hat{f}_k - m\delta_k \tag{4.13}$$

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 finite 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 (4.13), one obtains

$$0 \leq m \sum_{k=1}^{l} \delta_k \leq \sum_{k=1}^{l} \left(\hat{f}_k - \hat{f}_{k+1} \right), \tag{4.14}$$

so letting $l \to \infty$,

$$0 \leq m \sum_{k=1}^{\infty} \delta_k \leq \hat{f}_1 - \underbrace{\lim_{k \to \infty} \hat{f}_k}_{\neq \pm\infty}. \tag{4.15}$$

As a result,

$$\sum_{k=1}^{\infty} \delta_k = \sum_{k=1}^{\infty} \left(C^k + (d^k)^\top \left(Q + \frac{1}{t_k} \mathbb{I} \right) d^k \right) < \infty \tag{4.16}$$

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

$$C_k \to 0 \quad \text{and} \quad (d^k)^\top \left(Q + \frac{1}{t_k} \mathbb{I} \right) d^k \to 0. \tag{4.17}$$

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

$$S^k + \nu^k = \left(Q + \frac{1}{t_k} I \right) d^k \to 0 \quad \text{for} \quad k \in K. \tag{4.18}$$

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

□

Remark: If one assumes that the set $\Omega = \{x \in \mathbb{R}^n | f(x) \leq f(x^1) + 2\bar{\sigma}\}$ 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 4.2. *[30] 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 [30] 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 (4.12) by

$$\Psi_k := M_k(x^{k+1}) + \left(d^k\right)^\top \frac{1}{2}\left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k. \tag{4.19}$$

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

$$A_k(\hat{x}) = M_k(x^{k+1}) - \langle S^k, d^k \rangle. \tag{4.20}$$

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.

$$\begin{aligned}
\Psi_{k+1} &= M_k(x^{k+2}) + \frac{1}{2}\left(d^{k+1}\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^{k+1} && (4.21)\\
&\geq A_k(x^{k+2}) + \frac{1}{2}\left(d^{k+1}\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^{k+1} && (4.22)\\
&= M_k(x^{k+1}) + \langle S^k, x^{k+2} - x^{k+1} \rangle + \frac{1}{2}\left(d^{k+1}\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^{k+1} && (4.23)\\
&= \Psi_k - \frac{1}{2}\left(d^k\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k + \frac{1}{2}\left(d^{k+1}\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^{k+1} && (4.24)
\end{aligned}$$

$$- \left(d^k\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right)\left(d^{k+1} - d^k\right) - \langle \nu^k, x^{k+2} - x^{k+1}\rangle \tag{4.25}$$

$$\geq \Psi_k + \frac{1}{2}\left(d^{k+1} - d^k\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right)\left(d^{k+1} - d^k\right) \tag{4.26}$$

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

$$\|d^{k+1} - d^k\| \to 0, \quad k \to \infty. \tag{4.27}$$

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

$$\hat{f} = \delta_k + M_k(\hat{x}) - C_k - \left(d^k\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right)\left(d^k\right) \tag{4.28}$$

$$= \delta_k + M_k(x^{k+1}) - \langle S^k, d^k\rangle - \left(d^k\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right)\left(d^k\right) \tag{4.29}$$

$$= \delta_k \tag{4.30}$$

$$\geq \delta_k + M_k(\hat{x} + d^k) \tag{4.31}$$

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

$$\delta^{k+1} \leq \hat{f} - M_{k+1}(\hat{x} + d^{k+1}). \tag{4.32}$$

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

$$-\hat{f}_{k+1} + c_{k+1}^{k+1} - \left\langle s_{k+1}^{k+1}, d^{k+1}\right\rangle \geq -M_{k+1}(\hat{x} + d^{k+1}). \tag{4.33}$$

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

$$m\delta_k + \left\langle s_{k+1}^{k+1}, d^k - d^{k+1}\right\rangle \geq \hat{f} - M_{k+1}(\hat{x} + d^{k+1}). \tag{4.34}$$

Combining this relation with ??? yields

$$0 \leq \delta_{k+1} \leq m\delta_k + \left\langle s_{k+1}^{k+1}, d^k - d^{k+1} \right\rangle . \tag{4.35}$$

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 (4.27) and the boundedness of $\{\eta_k\}$ using [21, Lemma 3, p.45] it follows from (4.35) that

$$\lim_{k\to\infty} \delta_k = 0. \tag{4.36}$$

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

$$\xi \left(d^k\right)^\top d^k \leq \left(d^k\right)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k \to 0 \tag{4.37}$$

□

Remark: To the boundedness of Q and the resulting inequalities:

$$A \succ B \Leftrightarrow A - B \succ 0 \Leftrightarrow A - B \text{ is positive definite}$$

Respectively for $A \prec B$
For a real symmetric matrix A and a vector $d \in \mathbb{R}^n$ the following result holds:

$$A \prec \xi\mathbb{I} \Rightarrow Ad < \xi d$$

Proof: A real and symmetric $\Rightarrow$ it is orthogonally diagonalizeable:

$$\begin{aligned} \exists \lambda_i \in \mathbb{R} \text{ eigenvalues} \quad v^i \in \mathbb{R}^n \text{ eignvektors} & \\ Av^i = \lambda_i v^i & \\ \text{and} \quad \exists \alpha_i \in \mathbb{R}: \quad d = \sum_i \alpha_i v^i & \\ \Rightarrow Ad = A\sum_i \alpha_i v^i = \sum_i \alpha_i A v^i = \sum_i \alpha_i \lambda_i v^i & \\ A \prec \xi\mathbb{I} \Leftrightarrow \max_i \lambda_i < \xi & \\ \Rightarrow Ad < \xi \sum_i \alpha_i v^i = \xi d & \end{aligned} \tag{4.38}$$

References

[1] P. Apkarian, D. Noll, and O. Prot. A trust region spectral bundle method for nonconvex eigenvalue optimization. *SIAM Journal on Optimization*, 19(1):281–306, jan 2008.

[2] Welington de Oliveira and Claudia Sagastizábal. Bundle methods in the xxist century: A bird's-eye view. *Pesquisa Operacional*, 34(3):647–670, dec 2014.

[3] A. Fuduli, M. Gaudioso, and G. Giallombardo. A dc piecewise affine model and a bundling technique in nonconvex nonsmooth minimization. *Optimization Methods and Software*, 19(1):89–102, 2004.

[4] A. Fuduli, M. Gaudioso, and G. Giallombardo. Minimizing nonconvex nonsmooth functions via cutting planes and proximity control. *SIAM Journal on Optimization*, 14(3):743–756, 2004.

[5] Carl Geiger and Christian Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben*. Sp, 2002.

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

[7] Michael HintermÃ$\frac{1}{4}$ller. A proximal bundle method based on approximate subgradients. *Computational Optimization and Applications*, 20:245–266, 2001.

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

[9] 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.

[10] K. C. Kiwiel. Bundle methods for convex minimization with partially inexact oracles. Technical report, Systems Research Institute, Polish Academy of Sciences, 2010.

[11] Krzysztof C. Kiwiel. *Methods of Descent for Nondifferentiable Optimization*. Springer, 1985.

[12] Krzysztof C. Kiwiel. An aggregate subgradient method for nonsmooth and nonconvex minimization. *Journal of Computational and Applied Mathematics*, 14(3):391–400, 1986.

[13] Krzysztof C. Kiwiel. A proximal bundle method with approximate subgradient linearizations. *SIAM Journal on Optimization*, 16(4):1007–1023, jan 2006.

[14] A. S. Lewis and S. J. Wright. A proximal method for composite minimization. *Mathematical Programming*, 158(1-2):501–546, aug 2015.

[15] 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.

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

[17] Dominikus Noll. Cutting plane oracles to minimize non-smooth non-convex functions. *Set-Valued and Variational Analysis*, 18(3-4):531–568, sep 2010.

[18] Dominikus Noll. Bundle method for non-convex minimization with inexact subgradients and function values. In *Computational and Analytical Mathematics*, pages 555–592. Springer Nature, 2013.

[19] Dominikus Noll and Pierre Apkarian. Spectral bundle method for non-convex maximum eigenvalue functions: first-order methods. *Mathematical Programming*, 104(2-3):701–727, jul 2005.

[20] Dominikus Noll, Olivier Prot, and Aude Rondepierre. A proximity control algorithm to minimize non-smooth and non-convex functions. *Pacific Journal of Optimization*, 4(3):571–604, 2012.

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

[22] R. Tyrrell Rockafellar and Roger J. B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 3rd edition, 2009.

[23] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1996.

[24] Helga Schramm and Jochem Zowe. A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results. *SIAM Journal on Optimization*, 2(1):121–152, feb 1992.

[25] Helga Schramm and Jochem Zowe. A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results. *SIAM Journal on Optimization*, 2(1):121–152, feb 1992.

[26] M. V. Solodov. Aon approximations with finite rprecision in bundle methods for nonsmooth optimization. *Journal of Optimization Theory and Applications*, 119(1):151–165, 2003.

[27] Mikhail V. Solodov. *Constraint Qualifications*. Wiley Encyclopedia of Operations Research and Management Science, 2011.

[28] Jay S. Treiman. Clarke's gradients and ε-subgradients in banach spaces. *Transactions of the American Mathematical Society*, 294(1):65–65, jan 1986.

[29] Claudia Sagastizàbal Warren Hare. Computing proximal points of nonconvex functions. *Mathematical Programming*, 116:221–258, 2009.

[30] 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.

[31] 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.