Contents

References

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 [21]. To provide an easier understanding of the proximal bundle method in [8] 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 [7]. We focus here on the primal approach.

1.1 Derivation of the Bundle Method

This section gives a short summery of the derivations and results of chapter XV in [10] 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 [10].

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.

1.1.1 A Stabilized Cutting Plane 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

$$l_j(x) = f(x^j) + \left\langle g^j, x - x^j \right\rangle$$

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

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

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

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

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 a solution. To tackle this problem a penalty term is introduced to the subproblem. It then reads

$$\min \tilde{m}_k(x) = m_k(x) + \frac{1}{2t_k}\|x - x^k\|^2 \quad \text{s.t.} \quad x \in X,\ t_k > 0. \tag{1.2}$$

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

The regularization term can be motivated and interpreted in many different ways, c.f. [10]. 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^M which is the decrease of the model for the new iterate x^{k+1} compared to the function value at the

current stability center $\hat{x}^k$

$$\delta_k^M = f(\hat{x}^k) - m_k(x^{k+1}) \geq 0. \tag{1.3}$$

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

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

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

Besides the model decrease other forms of decrease measures and variations of these are possible. Some are presented in [10] and [36].

1.1.2 Subproblem Formulations

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 $\hat{x}^k$.

$$\begin{aligned} l_j(x) &= f(x^j) + \left\langle g^j, x - x^j \right\rangle \\ &= f(\hat{x}^k) + \left\langle g^j, x - \hat{x}^k \right\rangle - \left(f(\hat{x}^k) - f(x^j) + \left\langle g^j, x^j - \hat{x}^k \right\rangle \right) \\ &= f(\hat{x}^k) + \left\langle g^j, x - \hat{x}^k \right\rangle - e_j^k \end{aligned}$$

where

$$e_j^k := f(\hat{x}^k) - f(x^j) + \left\langle g^j, x^j - \hat{x}^k \right\rangle \geq 0 \quad \forall j \in J_k$$

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 (1.2) can now be written as

$$\min_{\hat{x}^k+d\in X} \tilde{m}_k(\hat{x}^k+d) = f(\hat{x}^k) + \max_{j\in J_k}\left\{\left\langle g^j, d\right\rangle - e_j^k\right\} + \frac{1}{2t_k}\|d\|^2 \tag{1.4}$$

$$\Leftrightarrow \quad \min_{\substack{\hat{x}^k+d\in X,\\ \xi\in\mathbb{R}}} \xi + \frac{1}{2t_k}\|d\|^2 \quad \text{s.t.} \quad \left\langle g^j, d\right\rangle - e_j^k - \xi \le 0, \quad j\in J_k \tag{1.5}$$

where $d := x - \hat{x}^k$ and 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 (1.5) if and only if

$$\begin{aligned} &d^k \text{ solves the original subproblem (1.4) and} \\ &\xi_k = \max_{j\in J_k} {g^j}^\top d^k - e_j^k = m_k(\hat{x}^k + d^k) - f(\hat{x}^k). \end{aligned} \tag{1.6}$$

The new iterate is 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 [28].

Subproblem (1.2) then reads

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

with respect to the serious point $\hat{x}^k$.

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* [7] or *prox-operator*

$$prox_{t,\check{f}}(x) = \underset{y\in\mathbb{R}^n}{\arg\min}\left\{\check{f}(y) + \frac{1}{2t}\|x-y\|^2\right\}, \quad t > 0. \tag{1.7}$$

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

We look again at a slightly different formulation of the bundle subproblem

$$\begin{aligned} \min_{\substack{d\in\mathbb{R}^n,\\ \xi\in\mathbb{R}}} \quad & \xi + \mathtt{i}_X + \frac{1}{2t_k}\|d\|^2 \\ \text{s.t.} \quad & \left\langle g^j, d\right\rangle - e_j^k - \xi \le 0, \quad j \in J_k. \end{aligned}$$

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 [11] assuming a constraint qualification holds if the constraint set X makes it necessary [32].

There exist a subgradient $\nu^k \in \partial \mathtt{i}_X(x^{k+1})$ and Lagrangian multipliers $\alpha_j,\ j \in J^k$ such that

$$0 = \nu^k + \frac{1}{t_k}d^k + \sum_{j\in J^k}\alpha_j g^j \tag{1.8}$$

$$\sum_{j\in J_k}\alpha_j = 1, \tag{1.9}$$

$$\alpha_j \ge 0,\ j \in J^k, \tag{1.10}$$

$$\left\langle g^j, d^k\right\rangle - e_j^k - \xi_k \le 0, \tag{1.11}$$

$$\sum_{j\in J^k}\alpha_j\left(\left\langle g^j, d^k\right\rangle - e_j^k - \xi_k\right) = 0. \tag{1.12}$$

From condition (1.8) follows that

$$d^k = t_k\left(G^k + \nu^k\right) \tag{1.13}$$

with the *aggregate subgradient*

$$G^k := \sum_{j \in J^k} \alpha_j g^j \quad \in \partial m_k(x^{k+1}). \tag{1.14}$$

Rewriting condition (1.12) yields the *aggregate error*

$$E_k := \sum_{j \in J^k} \alpha_j e_j^k = (G^k)^\top d^k + f(\hat{x}^k) - m_k(x^{k+1}). \tag{1.15}$$

Here relation (1.6) 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] 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*

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

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 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 constrained case the stopping condition is written as

$$\delta_k = E^k + t_k \|G^k + \nu^k\|^2 \leq \texttt{tol}.$$

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

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

where (1.14) and (1.15) were used, shows that the new δ_k is only a little variation of the model decrease δ_k^M. If the iterate x^{k+1} does not lie on the boundary of the constraint set X, the vector $\nu^k = 0$ and the expression simplifies to the one stated in (1.3).

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

$$m_{k+1}(\hat{x}^k + d) \geq f(\hat{x}^{k+1}) - e_{k+1}^{k+1} + \left\langle g^{k+1}, d \right\rangle \tag{1.16}$$

$$m_{k+1}(\hat{x}^k + d) \geq a_k(\hat{x}^k + d) \tag{1.17}$$

The first condition means, that the newly computed information is 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 [8]. 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.

An issue of bundle methods is that in spite of the possibility to delete inactive information the bundle can still become very big. Kiwiel therefore proposed a totally different use of the aggregate objects in [15]. The aggregate subgradient can be used to build the *aggregate linearization*

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

This function can be used to avoid memory overflow as it compresses the information of all bundle elements into one affine plane. Adding the function a_k to the cutting plane model preserves the assumptions (1.16) and (1.17) put on the model and can therefore be used instead of or in combination with the usual cutting planes.

This can however impair the speed of convergence if the bundle is kept too small and provides hence less information about the objective function [2].

We have now all the ingredients so that the following basic bundle algorithm can be stated:

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
$$d^k = \arg\min_{d \in \mathbb{R}^n} m_k(\hat{x}^k + d) + \mathtt{i}_X(\hat{x}^k + d) + \frac{1}{2t_k}\|d\|^2$$
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^{k+1} 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 (1.16) and (1.17) are fulfilled.

In practice the choice of t_k can be realized by taking

$$t_{k+1} = \kappa_+ t_k, \quad \kappa_+ > 1 \tag{1.18}$$

at every serious step and

$$t_{k+1} = \max\{\kappa_- t_k, t_{min}\}, \quad \kappa_- < 1 \text{ and } t_{min} > 0 \tag{1.19}$$

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 [25] and [23]. The trust region idea was very much exploited by Schramm and Zowe in [29]. In the case $X = R^n$ the sequence $\{\hat{x}^k\}$ can be unbounded. In this case bounding $t_k < \infty \ \forall k$ preserves the convergence proof [10, Theorem 3.2.2].

In general it can be shown that if f posses global minima and the basic bundle algorithm generates the sequence $\{\hat{x}^k\}$ this sequence converges to a minimizer of problem (1.1) (c.f [10]).

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 [20] and [14]. It then took over 25 years that bundle methods were again generalized to the use of inexact information, first works on this subject being [9, 16] and [31].

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 therefore no additional strategies are used to cope with the noise. 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 [36]. Here the most important aspects of that paper are reviewed.

2.1.1 Different Types of Inexactness

Throughout this section we consider the optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \tag{2.1}$$

where the function $f : \mathbb{R}^n \to \mathbb{R}$ is a finite convex function. The function values and one subgradient at each point x are given by an inexact oracle. It is reasonable to define very different kinds of inexactness and further assumptions can be put on the noise to reach stronger convergence results. However, generally inexact information for convex objective functions is defined in the following way:

$$f_x = f(x) - \sigma_x, \quad \sigma_x \leq \bar{\sigma} \tag{2.2}$$

$$g_x \in \mathbb{R}^n \text{ such that } f(\cdot) \geq f_x + \langle g_x, \cdot - x \rangle - \theta_x, \quad \theta_x \leq \bar{\theta}. \tag{2.3}$$

From this follows because of

$$f(\cdot) \geq f(x) + \langle g_x, \cdot - x \rangle - (\sigma_x + \theta_x) \tag{2.4}$$

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

Different convergence results for the applied bundle methods are possible depending on if the bounds $\bar{\sigma}$ and $\bar{\theta}$ are unknown, known or even controllable.

In case of controllability of $\bar{\sigma}$ and $\bar{\theta}$ it may be possible to drive them to zero as the iterations increase $\lim_{k\to\infty} \sigma_k = 0$ and $\lim_{k\to\infty} \theta_k = 0$. We talk then of *asymptotically vanishing errors*. This case 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 only asymptotically exact for descent steps [13]. This observation was the motivation for the partly inexact bundle methods presented in [13] and [36]. 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. [36].

In view of good convergence properties oracle that only underestimate the true function, so called *lower oracles*, are also very interesting. 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 new steps into the method to cope with the inexactness, such as noise attenuation [36, 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 [9] and [16]. 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. A more detailed description is given in [36].

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 defined in (2.2) and (2.3)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 $\bar{\sigma}$-solution of the problem meaning that

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

with f^* being an exact solution of problem (2.1).

Generally for asymptotically vanishing errors it is possible to construct bundle methods very similar to the basic bundle method that converge to the exact minimum of the problem. For more detailed results refer to [36].

2.2 Nonconvex Bundle Methods with Exact Information

In the nonconvex case the optimization problem is the following:

$$\min_{x \in \mathbb{R}^n} f(x). \tag{2.5}$$

This time $f : \mathbb{R}^n \to \mathbb{R}$ is a finite, locally Lipschitz function. It is neither expected to be convex nor differentiable.

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 set of functions 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* [15, 20]. Here the linearization error is essentially replaced by the nonnegative number

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

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.

2.2.1 Proximity Control

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 [8, 23] 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, 19, 22, 24, 25] and [30].

2.2.2 Other Concepts

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

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

is approximated to find a descent direction of f. The negative linearization errors are addressed by using 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 [25] 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:

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

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 [23] the second order approach of [25] 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 [8] that is presented in section 4 of this thesis.

3 Proximal bundle method for nonconvex functions with inexact information

This section focuses on the proximal bundle method presented by Hare et al. in [8]. 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 [7]: 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

Throughout this section we consider 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. [27, Theorem 7.25] $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ 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 $\bar{x}, f(\bar{x})$.

3.1.1 Inexactness

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

$$\|f_x - f(x)\| \leq \sigma_x$$

then f approximates the value $f(x)$ within σ_x. This is a little bit different from the definition in (2.2). In the convex case it follows from (2.4) that $\bar{\sigma} \geq \sigma_x \geq -\theta_x \geq -\bar{\theta}$ and therefore $f_x \in [f(x) - \bar{\theta}, f(x) + \bar{\sigma}]$.

As the 'normal' ε-subdifferential is not defined for nonconvex functions we adopt the notation used in [23] and interpret inexactness in the following way: $g \in \mathbb{R}^n$ approximates a subgradient of $\partial f(x)$ within $\theta \geq 0$ if

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

where $\partial f(x)$ is the Clarke subdifferential defined in Preliminaries section 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 [33]. It means that the approximation of the subgradient of $f(x)$ is an exact subgradient of a small perturbation of f at x. $\partial_{[\varepsilon]} f(x)$ is also known as the *Fréchet ε-subdifferential* of $f(x)$

Definition 3.2. [12] Let $f : \mathbb{R}^n \to \mathbb{R}$, $\varepsilon > 0$. The *Fréchet ε-subdifferential* of f at x is

defined by

$$\partial_{[\varepsilon]} f(x) := \left\{ g \in \mathbb{R}^n | \liminf_{\|h\| \to 0} \frac{f(x+h) - f(x) - \langle g, h \rangle}{\|h\|} \geq -\varepsilon \right\}.$$

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 [23].

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

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

For ease of notation we write from now on f_j instead of f_{x^j} for the approximation of the function value at the j'th iterate in the bundle J. The approximation at the k'th stability center reads $\hat{f}_k$.

3.1.2 Nonconvexity

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 [35] 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 (1.7).

The key idea is to use the relation

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

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

$$\tilde{f}(y) = f(y) + \frac{\eta}{2}|y - x|^2 \tag{3.2}$$

with respect to the parameter t [7]. η is therefore called the *convexification parameter* and t the *prox-parameter*.

The main difference of the method in [8] 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:

In addition to downshifting the linear functions forming the model they have a tilted slope. This is because instead of subgradients of the original objective f subgradients of the function $\tilde{f}$ are taken. We call them *augmented subgradients*. At the iterate x^j it is given by

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

Downshifting is done in a way that keeps the linearization error nonnegative. The *augmented linearization error* is therefore defined as

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

and

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

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

The new model function can therefore be written as

$$M_k(\hat{x}^k + d) := \hat{f}_k + \max_{j \in J_k} \left\{ \left\langle s_j^k, d \right\rangle - c_j^k \right\}.$$

3.1.3 Aggregate Objects

The definition of the *augmented aggregate subgradient* S^k, *error* C_k and *linearization* A_k follows straightforwardly:

$$S^k := \sum_{j \in J_k} \alpha_j^k s_j^k, \tag{3.3}$$

$$C_k := \sum_{j \in J_k} \alpha_j^k c_j^k \tag{3.4}$$

$$A_k(\hat{x}^k + d) := M_k(x^{k+1}) + \left\langle S^k, d - d^k \right\rangle. \tag{3.5}$$

Just as the model decrease

$$\delta^k := C_k + t_k \|S^k + \nu^k\|^2.$$

As the model function is convex even for nonconvex objective functions it is still minorized by the aggregate linearization. It holds

$$A_K(\hat{x}^k + d) \leq M_k(\hat{x}^k + d). \tag{3.6}$$

The update of t_k can be done in the same way described in (1.18) and (1.19) for the basic bundle method. Similarly the methods to update the bundle index set J^k stay valid. The update conditions (1.16) and (1.17) for the model are now written with respect to the augmented aggregate linearization and the approximate function value $\hat{f}_{k+1}$.

$$M_{k+1}(\hat{x}^k + d) \geq \hat{f}_{k+1} - c_{k+1}^{k+1} + \left\langle s^{k+1}, d \right\rangle \tag{3.7}$$

$$M_{k+1}(\hat{x}^k + d) \geq A_k(\hat{x}^k + d). \tag{3.8}$$

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

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

$$G^k = \sum_{j \in J_k} \alpha_j^k s_j^k, \quad \nu^k \in \partial \mathtt{i}_X(x^{k+1})$$
$$C_k = \sum_{j \in J_k} \alpha_j^k c_j^k,$$
$$\delta_k = C_k + t_k \|G^k + \nu^k\|^2.$$

   If $\delta_k \leq \mathtt{tol} \rightarrow$ 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.

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 constraint set X ensures 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.

Another possibility is to bound the step sizes t_k also from above. Then the sequence $\{\hat{x}^k\}$ also stay bounded and the proof still holds. In [36] another stopping criterion is psoposed that ensures convergence even for unbounded sequences $\{\hat{x}^k\}$. Is this also possible in my case or only for convex???

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 [7]. 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 $\lim_{k\to\infty} \theta_k = 0$ the convergence theory holds equally well with error bounds $\bar{\sigma} = \bar{\theta} = 0$ in [8, 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.
remark on η_k? how does it behave in my applications???

3.2.3 Convex Objective Functions

An obvious gain when working with convex objective functions is that the approximate stationarity condition of [8, Lemma 5 (iii)] is now an approximate optimality condition. If one takes the error definitions (2.2) and (2.3) that are available in the convex case and assumes $X = \mathbb{R}^n$ statement (22) in [8] therefore means that

$$0 \in \partial_{\bar{\sigma}+\bar{\theta}} f(\bar{x}).$$

Thus $\bar{x}$ is $(\bar{\sigma} + \bar{\theta})$-optimal.

This follows from the definition of S^k in (3.3) and local Lipschitz continuity of the ε-subdifferential [27, Proposition 12.68].

- (iii) in Lemma 5 für conv eps-subdiff umschreiben
- beweis für $\bar{\sigma}$-optimalität
-

bounded t_k instead of D? better????

- 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 [8] presented above and Noll [23]. 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 different from those designed for nonconvex objective functions and are generally not able to deal with such functions.

4 Variable Metric Bundle Method

introduction

A way to extend the proximal bundle method is to use an arbitrary metric $\frac{1}{2}\langle d, W_k d\rangle$ with a symmetric and positive definite matrix W_k instead of the Euclidean metric for

the stabilization term $\frac{1}{2t_k}\|d\|^2$. Methods doing so are called *variable metric bundle methods*. This section combines the method of Hare et al. presented in section 3 with the second order model function used by Noll in [23] to a metric bundle method suitable for nonconvex functions with noise.

This section therefore starts by explaining the ideas from [23] used to extend the method presented above. It then gives an explicit strategy how to update the metric during the steps of the algorithm and concludes with a convergence proof for the developed method.

Names and definitions of the objects are the ones used in section 3.

4.1 The Main Ingredients to the Method

explanation

As already mentioned in section 1 the stabilization term can be interpreted in many different ways. In the context of this section we can understand it as a pretty rough approximation of the curvature of the objective function. Of course bundle methods are designed to work with non differentiable objectives so it cannot be expected that the function provides any kind of curvature. However, if it does, incorporating it into the method could speed up convergence.

4.1.1 Variable Metric Bundle Methods

Variable metric bundle methods use an approach that can be motivated by the thoughts stated above. Instead of using the Euclidean norm for the stabilization term $\frac{1}{2}\|d\|^2$ the metric is derived from as symmetric and positive definite matrix W_k. As the name of the method suggests, this matrix can vary over the iterations of the algorithm. The subproblem in the k'th iteration therefore reads

$$\min_{\hat{x}^k+d} M_k(\hat{x}^k + d) + \frac{1}{2}\langle d, W_k d\rangle .$$

some more explanation

stabilization term mimics curvature → more sophisticated in variable metric bundle methods → explain those; examples → Noll also kind of variable metric bundle method → explain noll again (ref to chapter 2) and relate to variable metric

Paper that use variable metric bundle methods: [17, 18, 6, 34]

4.1.2 Second Order Model

Describe what is needed from Noll → describe how to really ensure desired properties (Vlcek) → explain other changes like δ_k → say what is not changed

This is also used by Noll et al. in [25] and Noll in [23] for nonconvex functions with exact and inexact information respectively. In the original papers the motivation is to approximate the original objective function by a quadratic model

$$\Phi(x, \hat{x}) = \phi(x, \hat{x}) + \frac{1}{2}\langle x - \hat{x}, Q(\hat{x})(x - \hat{x})\rangle \tag{4.1}$$

with the convex and possibly nonsmooth first order model $\phi(\cdot, \hat{x})$ and the quadratic but not necessarily convex second order part $\frac{1}{2}\langle \cdot - \hat{x}, Q(\hat{x})(\cdot - \hat{x})\rangle$. This model is then again approximated to solve the bundle subproblem.

Here we take the same idea but interpret it more in the sense of a stabilization. The objective function is still first convexified and then this convexification approximated by a cutting plane model. The matrix $Q(\hat{x})$ is added into the stabilization so that the k'th bundle subproblem is

$$\min_{\hat{x}^k + d \in X} M_k(\hat{x}^k + d) + \frac{1}{2}\langle d, \left(Q + \frac{1}{t_k}\mathbb{I}\right) d\rangle \tag{4.2}$$

...???

In [25] and [23] it is not specified how the matrix $Q(\hat{x}$ is to be chosen. For convergence it is necessary that the eigenvalues of $Q(\hat{x}$ are bounded. We adopt here the notation in [23] and say

$$\exists q > 0 \text{ such that } -q\mathbb{I} \prec Q(\hat{x}) \prec q\mathbb{I}. \tag{4.3}$$

Here $A \prec B$ means that the matrix $B - A$ is positive definite.

Additionally the matrix $Q(\hat{x} + \frac{1}{t_k}\mathbb{I}$ has to be positive definite.

There are no other conditions put on $Q(\hat{x}$.

It is not stated explicitly how to form such a matrix $Q(\hat{x})$. Here we take an idea developed by Vlček and Lukšan for a variable metric bundle method in [34].

The matrix $Q(\hat{x}$ is formed by a BFGS-Update from the last known - which? how many? all from bundle?? subgradients.

- Q only updated in serious steps - why?
-

possible to do also SR1-update in null steps???

4.1.3 how to get Q and ensure all conditions

4.1.4 the new stopping criterion

There are some minor changes that have to be made compared to the algorithm proposed by Hare et al. the biggest being the stopping condition.

In the same way as for (???) from the optimality condition

$$0 \in \partial M_k(x^{k+1}) + \partial \mathtt{i}_D(x^{k+1}) + \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k \tag{4.4}$$

$$\tag{4.5}$$

follows that

$$S^k + \nu^k = -\left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k. \tag{4.6}$$

From this the model decrease (???) can be recovered using (???), (??) and (4.5):

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

As the changes in the algorithm concern only the stabilization and the model decrease d_k all relations that were obtained for the different prats of the model M_k in section 3 are still valid.

4.2 Keywords

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

$$\text{if} \quad \text{norm}\left(\frac{y^k {y^k}^\top}{{y^k}^\top d^k}\right) > 1/3C \tag{4.8}$$

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

$$\zeta = \frac{\text{norm}\left(y^k {y^k}^\top\right)}{1/3C} \tag{4.10}$$

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

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 $\xi > 0$.

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

4.3 Algorithm

same form as Hare algorithm (nullstep)
add Q calculation

Nonconvex Variable Metric 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

$$G^k = \sum_{j\in J_k} \alpha_j^k s_j^k,$$

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

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 $t_{k+1} > 0$.
   Calculate $Q(\hat{x}^k)$ ... 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.

4.4 Convergence Analysis

In this section the convergence properties of the new method are analyzed. We do this the same way it is done by Hare et al. in [8].

In the paper all convergence properties are first stated in [8, Lemma 5]. It is then shown that all sequences generated by the method meet the requirements of this lemma which we repeat here for convenience.

Lemma 4.1 ([8, Lemma 5]) *Suppose that the cardinality of the set* $\{j \in J^k | \alpha_j^k > 0\}$ *is uniformly bounded in* k.

(i) If $C^k \rightarrow 0$ *as* $k \rightarrow \infty$, *then*

$$\sum_{j \in J^k} \alpha_j^k \|x^j - \hat{x}^k\| \rightarrow 0 \text{ as } k \rightarrow \infty.$$

(ii) If additionally for some subset $K \subset \{1, 2, \dots\}$,

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

then we also have

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

(iii) If in addition $S^k + \nu^k \to 0$ *as* $K \in k \to \infty$, *then* $\bar{x}$ *satisfies the approximate stationarity condition*

$$0 \in (\partial f(\bar{x}) + \partial \mathtt{i}_X(\bar{x})) + B_{\bar{\theta}}(0). \tag{4.13}$$

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

$$f(y) \geq f(\bar{x}) - (\bar{\theta} + \varepsilon)\|y - \bar{x}\| - 2\bar{\sigma}, \quad \text{for all } y \in X \cup B_\rho(\bar{x}). \tag{4.14}$$

As the neither the stabilization nor the descent test is involved in the proof of Lemma 4.1 it is the same as in [8].

We prove now that also the variable metric version of the algorithm fulfills all requirements of Lemma 4.1. The proof is divided into two parts. The first case covers the case of infinitely many serious steps, the second one considers infinitely many null steps.

For both proofs the following lemma is needed:

Lemma 4.2 *For a symmetric matrix* $A \in \mathbb{R}^{n \times n}$ *and a vector* $d \in \mathbb{R}^n$ *the following result holds:*

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

Proof: As the matrix A is real and symmetric it is orthogonally diagonalizeable. There exist eigenvalues $\lambda_i \in \mathbb{R}, i = \{1, ..., n\}$ and corresponding eigenvectors $v^i \in \mathbb{R}^n, i = \{1, ..., n\}$ that satisfy the equations

$$Av^i = \lambda_i v^i \quad i = \{1, ..., n\}.$$

The eigenvectors v^i generate a basis for $\mathbb{R}^n$ so any vector $d \in \mathbb{R}^n$ can be written as

$$d = \sum_i \alpha_i v^i$$

for $\alpha_i \in \mathbb{R}^n, i = \{1, ..., n\}$.

This yields

$$Ad = A\sum_i \alpha_i v^i = \sum_i \alpha_i \lambda_i v^i. \tag{4.15}$$

Plugging the assumption $A \prec \xi\mathbb{I}$ which is equivalent to $\max_i \lambda_i < \xi$ into (4.15) we get relation (4.4) by

$$Ad < \xi \sum_i \alpha_i v^i = \xi d.$$

□

Theorem 4.3 (c.f.[8, Theorem 6]) *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 every 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 4.1 hold.

The proof is very similar to the one stated in [8] but minor changes have to be made due to the different formulation of the nominal decrease δ_k.

Proof: At each serious step we have

$$\hat{f}_{k+1} \leq \hat{f}_k - m\delta_k \tag{4.16}$$

where $m,\ \delta_k > 0$. From this follows that the sequence $\{\hat{f}_k\}$ is nonincreasing. Since $\{\hat{x}^k\} \subset X$ and f is continuous the sequence $f(\hat{x}^k)$ is bounded. With $|\sigma_k| < \bar{\sigma}$ 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 (4.16), one obtains

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

so letting $l \to \infty$,

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

This yields

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

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

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

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

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

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

□

For the case of infinitely many null steps we need result (31) from [8]. It only depends on the definitions of the augmented linearization error and subgradient.

Whenever x^{k+1} is as declared a null step, the relation

$$-c_{k+1}^{k+1} + \left\langle s_{k+1}^{k+1}, x^{k+1} - \hat{x}^k \right\rangle \geq -m\delta_k \tag{4.17}$$

holds.

Theorem 4.4 (c.f. [8, Theorem 7]) *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 4.1 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

$$\Psi_k := M_k(x^{k+1}) + \left(d^k\right)^\top \frac{1}{2}\left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k. \tag{4.18}$$

It is first shown that the sequence $\{\Psi_k\}$ is bounded above. From definition (3.5) follows

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

Using (4.6) for the second equality, the subgradient inequality for $\nu^k \in \partial \mathtt{i}_D$ in the first inequality and (3.6) for the second inequality one obtains

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

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. For this we obtain

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

where the first inequality comes from (3.6) and the fact that $t_{k+1} \leq t_k$ for null steps.The second equality follows from (3.5), the third equation by (4.6) and (4.18) and the last

inequality holds y $\nu^k \in \partial \mathtt{i}_X(x^{k+1})$.

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

$$\|d^{k+1} - d^k\| \to 0, \quad k \to \infty. \tag{4.19}$$

By definition (4.7) and the fact that the augmented aggregate error can be expressed as

$$C_k = \hat{f} - M_k(x^{k+1}) + \left\langle S^k, d^k \right\rangle$$

by the KKT conditions follows

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

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

$$\delta^{k+1} \leq \hat{f} - M_{k+1}(\hat{x} + d^{k+1}). \tag{4.20}$$

By assumption (3.7) on the model, written for $d = d^{k+1}$,

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

In the nullstep case $\hat{f}_{k+1} = \hat{f}$ so adding condition (4.17) to the inequality above, one obtains that

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

Combining this relation with (4.20) yields

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

Because $m \in (0,1)$ and $\left\langle s_{k+1}^{k+1}, d^k - d^{k+1} \right\rangle \to 0$ as $k \to \infty$ due to (4.19) and the boundedness of $\{\eta_k\}$ using [26, Lemma 3, p.45] it follows from (4.4) that

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

From formulation (4.7) of the model decrease follows that $C_k \to 0$ as $k \to \infty$. Since $Q + \frac{1}{t_k}\mathbb{I} \succ \xi\mathbb{I}$ due to $\liminf_{k\to\infty} > 0$ and the bounded eigenvalues of Q we have

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

This means that $d^k \to 0$ for $k \to \infty$ and therefore $\lim_{k\to\infty} x^k = \hat{x}$. It also follows that $\|S^k +$
$vu^k\| \to 0$ as $k \to \infty$. Passing to some subsequence if necessary we can conclude that S^k converges to some $\bar{S}$ and as $\hat{x}^k = \bar{x}$ for all k all requirements of Lemma 4.1 are fulfilled.

□

Remark: All results deduced in section 3.2 are still valid for this algorithm.

Rechtschreibfehler, Namen, Stil überprüfen

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] Napsu Haarala, Kaisa Miettinen, and Marko M. Mäkelä. Globally convergent limited memory bundle method for large-scale nonsmooth optimization. *Mathematical Programming*, 109(1):181–205, 2007.

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

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

[9] Michael HintermÃ¼ller. A proximal bundle method based on approximate subgradients. *Computational Optimization and Applications*, 20:245–266, 2001.

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

[11] 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.

[12] Alejandro Jofré, Dinh The Luc, and Michel Théra. ε-subdifferential and ε-monotonicity. *Nonlinear Analysis: Theory, Methods & Applications*, 33(1):71–90, jul 1998.

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

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

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

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

[17] Claude Lemaréchal and Claudia Sagastizábal. *An approach to variable metric bundle methods*, pages 144–162. Springer Berlin Heidelberg, Berlin, Heidelberg, 1994.

[18] Claude Lemaréchal and Claudia Sagastizábal. Variable metric bundle methods: From conceptual to implementable forms. *Mathematical Programming*, 76(3):393–410, 1997.

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

[20] 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.

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

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

[23] 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.

[24] 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.

[25] 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.

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

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

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

[29] 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.

[30] 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.

[31] 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.

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

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

[34] J. Vlček and L. Lukšan. Globally convergent variable metric bundle method for nonconvex nondifferentiable unconstrained minimization. *Journal of Optimization Theory and Applications*, 111(2):407–430, 2001.

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

[36] 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.