Contents

1 Bundle Methods

When bundle methods were first introduced in 1975 by Claude Lemaréchal and Philip Wolfe they were developed to minimize a convex (possibly nonsmooth) function f for which at least one subgradient at any point x can be computed [4].
To provide an easier understanding of the proximal bundle method in [5] and stress the most important ideas of how to deal with nonconvexity and inexactness first a basic bundle method is shown here.
link to chapter?

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 [1]. I focus here on the primal approach.

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

notation, definitions
already done in previous preliminaries chapter?

1.1 A basic bundle method

This section gives a short summery of the derivations and results of chapter XV in [2] 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 [2].

The optimization problem considered in this section is

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

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

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

explanation

1.1.1 Derivation of the bundle method

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

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

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

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

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

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

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

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

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

Some nice sentences to explain the term a little bit more and to lead over to the next paragraph.
To understand the deeper motivation of this term see [2]. For this introduction it suffices to see that due to the regularization term the subproblem is now strongly convex and

therefore always uniquely solvable.

The second major step towards the bundle algorithm is the introduction of a so called *stability center* or *serious point* $\hat{x}^k$. It is the iterate that yields the "best" approximation of the optimal point up to the k'th iteration (not necessarily the best function value though).
The updating technique for $\hat{x}^k$ is crucial for the convergence of the method: If the next iterate yields a decrease of f that is "big enough", namely bigger than a fraction of the decrease suggested by the model function for this iterate, the stability center is moved to that iterate. If this is not the case, the stability center remains unchanged.
In practice this looks the following:
Define first the *nominal decrease* δ_k which is the decrease of the model for the new iterate x^{k+1} compared to the function value at the current stability center $\hat{x}^k$.

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

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

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

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

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

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

citation for this???!!!

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

$$= f(\hat{x}^k) + {g^j}^\top (x - \hat{x}^k) - e_j^k \tag{9}$$

where

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

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

Subproblem (5) can now be written as

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

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

where the constant term $f(\hat{x}^k)$ was discarded for the sake of simplicity.
If X is a polyhedron this is a quadratic optimization problem that can be solved using standard methods of nonlinear optimization. The pair (ξ_k, d^k) solves (12) if and only if d^k solves the original subproblem (11) and $\xi_k = f(\hat{x}^k) + \max_{j \in J_k} {g^j}^\top d^k - e_j^k$. The new iterate is then given by $x^{k+1} = \hat{x}^k + d^k$. citation!!!

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

$$prox_{t,f}(x) = \arg\min_y \left\{ \check{f}(y) + \frac{1}{2t}\|x - y\|^2 \right\}, \quad t > 0 \tag{13}$$

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

1.1.2 Aggregate objects

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

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

Subproblem (5) then writes as

$$\min_{\hat{x}^k + d \in R^n, \xi \in \mathbb{R}} \xi + \mathbb{I}_X + \frac{1}{2t_k}\|d\|^2 \quad \text{s.t.} \quad {g^j}^\top d - e_j^k - \xi \leq 0, \quad j \in J_k \tag{14}$$

check if f also not put into subproblem before

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

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

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

Lemma 1.1. *The optimization problem (14) has for $t_k > 0$ a unique solution given by*

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

Furthermore

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

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

where

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

Comment on the inequality missing

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

Explain aggregation process in more detail

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

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

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

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

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

Finally use Lemma ??? in [2]

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

to reformulate the nominal decrease δ_k:

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

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

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

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

The following basic bundle algorithm can now be stated:

Reformluate equations, model function
introduce aggregate expressions
say something to J-update, say something to t-update
see if all abbreviations ($f_j, g^j, ...$) are introduced
introduce prox-operator and proximal points

algorithm

Basic bundle method

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

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

1. Calculate
$$d^k = \arg\min_{d\in\mathbb{R}^n} m_k(\hat{x}^k + d) + \mathbb{I}_X + \frac{1}{2t_k}\|d\|^2$$
and the corresponding Lagrange multiplier $\alpha_j^k,\ j \in J_k$. say how model m_k looks here. include $\mathbb{I}_X$

2. Set
$$G^k = \sum_{j\in J_k} \alpha_j^k g_j^k, \quad E_k = \sum_{j\in J_k} \alpha_j^k e_j^k, \quad \text{and} \quad \delta_k = E_k + t_k\|G^k + \nu^k\|^2$$
If $\delta_k \leq \texttt{tol} \rightarrow$ STOP.

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

4. Compute $f(x^{k+1}),\ g^{k+1}$.
If
$$f^{k+1} \leq \hat{f}^k - m\delta_k \quad \rightarrow \text{serious step.}$$
Set $\hat{x}^{k+1} = x^{k+1},\ f(\hat{x}^{k+1}) = f(x^{k+1})$ and select suitable $t_{k+1} > 0$.
Otherwise $\rightarrow$ nullstep.
Set $\hat{x}^{k+1} = \hat{x}^k,\ f(\hat{x}^{k+1}) = f(x^{k+1})$ and choose t_{k+1} in a suitable way.

5. Select new bundle index set $J_{k+1} = \{j \in J_k | \alpha_j^{k+1} \neq 0\} \cap k+1$, calculate e_j for $j \in J_{k+1}$ and update the model m_k.

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

explanation to t_k update. $\rightarrow$ include at which point??? This simple idea has however some major drawbacks [3]:

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

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

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

2 Noll Part

2.1 Introduction

2.2 Keywords

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

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

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

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

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

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

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

$$\begin{aligned}
\delta_k &= \hat{f}_k - M_k(x^{k+1}) - (\nu^k)^\top d^k && (23)\\
&= \hat{f}_k - A_k(x^{k+1}) - (\nu^k)^\top d^k && (24)\\
&= C_k - (S^k)^\top d^k - (\nu^k)^\top d^k && (25)\\
&= C_k - (S^k + \nu^k)^\top d^k && (26)\\
&= C_k + (d^k)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k && (27)
\end{aligned}$$

2.3 Algorithm

Nonconvex proximal bundle method with inexact information

Select parameters $m \in (0,1), \gamma > 0$ and a stopping tolerance $\mathtt{tol} \geq 0$.
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 → 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 [5] stays the same; no Q involved

Theorem 2.1. *Theorem 6 in [5] → 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 there exist* $\bar{x}$ *and* $\bar{S}$ *such that* $\hat{x}^k \to \bar{x}$, $S^k \to \bar{S}$ *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 [5] hold.*

The proof is very similar to the one stated in [5] 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

$$\hat{f}_{k+1} \leq \hat{f}_k - m\delta_k \tag{28}$$

where $m,\ \delta_k > 0$. From this follows that the sequence $\{\hat{f}_k\}$ is nonincreasing. Since $\{\hat{x}^k\} \subset D$ the sequence is by the fact that f is ??????? which assumption says f bounded below??? and $|\sigma_k| < \bar{\sigma}$ the sequence $\{f(\hat{x}^k) + \sigma_k\} = \{\hat{f}_k\}$ is bounded below. Together with the fact that $\{\hat{f}_k\}$ is nonincreasing one can conclude that it converges.
Using (28), one obtains

$$0 \leq m \sum_{k=1}^{l} \delta_k \leq \sum_{k=1}^{l} \left(\hat{f}_k - \hat{f}_{k+1}\right), \tag{29}$$

so letting $l \to \infty$,

$$0 \leq m \sum_{k=1}^{\infty} \delta_k \leq \hat{f}_1 - \underbrace{\lim_{k\to\infty} \hat{f}_k}_{\neq \pm\infty}. \tag{30}$$

As a result,

$$\sum_{k=1}^{\infty} \delta_k = \sum_{k=1}^{\infty} \left(C^k + (d^k)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k\right) < \infty \tag{31}$$

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

$$C_k \to 0 \quad \text{and} \quad (d^k)^\top \left(Q + \frac{1}{t_k}\mathbb{I}\right) d^k \to 0. \tag{32}$$

As $\liminf_{k\to\infty} t_k > 0$ need: eigenvalues of Q bounded!!! □

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.

References

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

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

[3] 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.

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

[5] 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.