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

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 [3] 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 [3].

The optimization problem considered in this section is

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

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

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 $f(x) \geq l_j(x),\ x \in X$.
The objective function f can now be approximated by the piecewise linear function

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

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

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

Picture of function and cutting plane approximation of it

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

$$\min \tilde{m}_k(x) = m_k(x) + \frac{1}{2t_k}\|x - x^k\|^2 \quad \text{s.t.} \quad x \in X,\ t_k > 0. \tag{5}$$

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

This regularization term can be motivated and interpreted in many different ways, c.f. [3]. From different possible regularization terms the most popular in bundle methods is the penalty-like regularization used here.

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

In practice this looks the following: Define first the *model decrease* δ_k which is the decrease of the model for the new iterate x^{k+1} compared to the function value at the current stability center $\hat{x}^k$.

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

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

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

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

Next to the model decrease other forms of decrease measures and variations of these are possible. Some are used in [3, 10].

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

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

where

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

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

Subproblem (5) can now be written as

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

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

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

$$d^k \text{ solves the original subproblem (11) and} \tag{13}$$

$$\xi_k = \max_{j\in J_k} {g^j}^\top d^k - e_j^k = m_k(\hat{x}^k+d^k) - f(\hat{x}^k). \tag{14}$$

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

1.1.2 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 [7].

Subproblem (5) then writes with respect to the serious point $\hat{x}^k$

$$\min_{x\in\mathbb{R}^n} m_k(x) + \mathtt{i}_X(x) + \frac{1}{2t_k}\|x-\hat{x}^k\|^2. \tag{15}$$

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* [2] or *prox-operator*

$$prox_{t,f}(x) = \arg\min_{y\in\mathbb{R}^n}\left\{\check{f}(y) + \frac{1}{2t}\|x-y\|^2\right\}, \quad t>0. \tag{16}$$

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.

1.1.3 Aggregate objects

Look again at a slightly different formulation of the bundle subproblem

$$\min_{\substack{d\in\mathbb{R}^n,\\ \xi\in\mathbb{R}}} \quad \xi + \mathtt{i}_X + \frac{1}{2t_k}\|d\|^2 \tag{17}$$

$$\text{s.t.} \quad {g^j}^\top d - e_j^k - \xi \leq 0, \quad j \in J_k. \tag{18}$$

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 [4] assuming a constraint qualification if the constraint set X makes it necessary [8].

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

$$0 = \nu^k + \frac{1}{t_k}d^k + \sum_{j\in J^k} \alpha_j g^j \tag{19}$$

$$\sum_{j\in J_k} \alpha_j = 1, \tag{20}$$

$$\alpha_j \geq 0,\ j \in J^k, \tag{21}$$

$$ {g^j}^\top d^k - e_j^k - \xi_k \leq 0, \tag{22}$$

$$\sum_{j\in J^k} \alpha_j \left(f(\hat{x}^k) + {g^j}^\top d^k - e_j^k - \xi_k\right) = 0. \tag{23}$$

From condition (19) follows then

$$d^k = t_k\left(G^k + \nu^k\right) \quad \text{with} \quad G^k := \sum_{j\in J^k} \alpha_j g^j \in \partial m_k(x^{k+1}) \tag{24}$$

with the *aggregate subgradient* G^k.

Rewriting condition (23) yields the *aggregate error*

$$E_k := \sum_{j \in J^k} \alpha_j e_j^k = (G^k)^\top d^k + f(\hat{x}^k) - m_k(x^{k+1}). \tag{25}$$

Here relation (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. *[1, 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.

A totally different use of the aggregate objects was proposed by Kiwiel in [5]. The aggregate subgradient can be used to build the *aggregate linearization*

$$a_k(\hat{x}^k + d) := m_k(x^{k+1}) + \langle G^k, d - d^k \rangle. \tag{26}$$

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. This is shown im more detail in reference.

References

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

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

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

[4] 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.

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

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

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

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

[9] 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.

[10] 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.