Survey of Bundle Methods for Nonsmooth Optimization

Article *in* Optimization Methods and Software · February 2002

DOI: 10.1080/1055678029002782

1 author:

Marko M. Mäkelä
University of Turku
139 PUBLICATIONS **3,167** CITATIONS

Optimization Methods and Software, Vol. 00, pp. 1–29
Reprints available directly from the publisher
Photocopying permitted by license only

SURVEY OF BUNDLE METHODS FOR NONSMOOTH OPTIMIZATION

MARKO M. MÄKELÄ

Department of Mathematical Information Technology,
University of Jyväskylä, P.O. Box 35 (Agora),
FIN-40351 Jyväskylä, Finland

(Received 22 March 2001; In final form 15 November 2001)

Bundle methods are at the moment the most efficient and promising methods for nonsmooth optimization. They have been successfully used in many practical applications, for example, in economics, mechanics, engineering and optimal control. The aim of this paper is to give an overview of the development and history of the bundle methods from the seventies to the present. For simplicity, we first concentrate on the convex unconstrained case with a single objective function. The methods are later extended to nonconvex, constrained and multicriteria cases.

Keywords: Nonsmooth optimization; Nondifferentiable programming; Bundle methods

1 INTRODUCTION

We consider the nonsmooth (nondifferentiable) optimization problem of the form

$$\begin{cases} \text{minimize} & f(x) \\ \text{subject to} & x \in G, \end{cases} \tag{P}$$

where the objective function $f : \mathbb{R}^n \to \mathbb{R}^m$ is not required to have continuous derivatives. We suppose only that f is a locally Lipschitz continuous function on the feasible set $G \subset \mathbb{R}^n$.

*E-mail: makela@mit.jyu.fi

Nonsmooth optimization problems of type (P) arise in many fields of applications, for example, in economics [1], mechanics [2], engineering [3] and optimal control [4]. The source of nonsmoothness can be divided into four classes: inherent, technological, methodological and numerical nonsmoothness.

In inherent nonsmoothness, the original phenomenon to be considered itself contains various discontinuities and irregularities. Typical examples of inherent nonsmoothness are the phase changes of material in the continuous casting of steel (see, e.g., [5]) and piecewise linear tax models in economics (see [6]). Technological nonsmoothness in a model is usually caused by some extra technological constraints. These constraints may cause a nonsmooth dependence between variables and functions although the functions originally were continuously differentiable. That kind of examples are so-called obstacle problems in optimal shape design (see [7]) and discrete feasible sets in product planning. On the other hand, some solution algorithms for constrained optimization may also lead to a nonsmooth problem. Examples of methodological nonsmoothness are the exact penalty function method and the Lagrange decomposition method. Finally, the problems may be analytically smooth but numerically nonsmooth. That is the case with, for instance, noisy input data or so-called "stiff problems", which are numerically unstable and behave like nonsmooth problems.

There are also several approaches to solve nonsmooth optimization problems. The usage of smooth gradient-based methods for nonsmooth problems is a simple approach but may lead to a failure in convergence, in optimality test or in gradient approximation, as was stated in an excellent, review of nonsmooth optimization [6]. On the other hand, the direct methods, for example, the method of Hooke and Jeves (see, e.g., [8] pp. 274–279) and polytope search (see. e.g., [9] pp. 145–148) employing no derivative information, become inefficient when the size of the problem is growing. Different kinds of regularization techniques introduced, for instance, in [7,10] may give satisfactory results in some cases but are not, in general, as efficient as the direct nonsmooth approach, as was noticed in [11].

The methods for nonsmooth optimization can be divided into two main classes: subgradient methods and bundle methods. Both of them are based on the assumption that only the objective function value and one subgradient at each point are available. The history of

[10.1.2002–2:04pm] [1–29] [Page No. 2] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

the subgradient methods (Kiev methods) starts in the 60s, and they are mainly developed in the Soviet Union. The basic idea behind the subgradient methods is to generalize the smooth methods by replacing the gradient by an arbitrary subgradient. Due to this simple structure, they are widely used methods in nonsmooth optimization, although they suffer from some serious drawbacks. Firstly, a nondescent search direction may occur, and thus no standard line search operation can be applied. For this reason, the step sizes have to be chosen a priori. Secondly, the lack of an implementable stopping criterion and the poor rate of the convergence (less than linear) are also disadvantages of the subgradient methods. To overcome the last handicap, in other words, in order to maintain linear convergence the variable metric ideas were adopted to a subgradient context in [12] by introducing two space dilation methods (ellipsoid method and *r*-algorithm). Also some modified ideas have been proposed in [13], where two adaptive variable metric methods, deviating in step size control, were derived. For an excellent overview of the subgradient methods we refer to [12].

In what follows we concentrate on the bundle methods, which are at the moment the most efficient and promising method class for nonsmooth optimization. In Section 2, we give a short review concerning the development of bundle methods. For simplicity, we first restrict ourselves to the convex unconstrained ($G = \mathbb{R}^n$) and single criterion ($m = 1$) version of problem (P). The methods to be presented are extended to nonconvex optimization in Section 3, to constrained optimization in Section 4, and, finally, to multicriteria optimization in Section 5.

2 CONVEX UNCONSTRAINED OPTIMIZATION

To point out the basic ideas and differences between the methods we here consider an unconstrained optimization problem of the form

$$\begin{cases} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathbb{R}^n, \end{cases} \tag{CP}$$

where the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is convex, in other words,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \tag{1}$$

[10.1.2002–2:04pm] [1–29] [Page No. 3] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$. We suppose that at every point $x \in \mathbb{R}^n$ we can compute the function value $f(x)$ and an unspecified subgradient $\xi(x)$ from the subdifferential

$$\partial f(x) = \{\xi(x) \in \mathbb{R}^n | f(y) \geq f(x) + \xi(x)^T(y - x) \quad \text{for all } y \in \mathbb{R}^n\}. \quad (2)$$

For a convex function we have the following necessary and sufficient optimality condition.

THEOREM 2.1 A *convex function* $f : \mathbb{R}^n \to \mathbb{R}$ *attains its global minimum at* x^*, *if and only if*

$$0 \in \partial f(x^*).$$

Proof See, for example, [14] p. 264. □

2.1 General Bundle Algorithm

We first describe a general bundle method that produces a sequence $\{x_k\}_{k=1}^{\infty} \subset \mathbb{R}^n$ converging to a global minimum of f, if it exists. We suppose that, in addition to the current iteration point x_k, we have some trial points $y_j \in \mathbb{R}^n$ (from past iterations) and subgradients $\xi_j \in \partial f(y_j)$ for $j \in J_k$, where the index set J_k is nonempty subset of $\{1, \ldots, k\}$.

The idea behind bundle methods is to approximate f from below by using a piecewise linear function, in other words, we replace f by a so-called *cutting plane model*

$$\hat{f}_k(x) := \max_{j \in J_k} \{f(y_i) + \xi_j^T(x - y_j)\}, \quad (3)$$

which can be written in the equivalent form

$$\hat{f}_k(x) = \max_{j \in J_k} \{f(x_k) + \xi_j^T(x - x_k) - \alpha_j^k\}, \quad (4)$$

with the *linearization error*

$$\alpha_j^k := f(x_k) - f(y_j) - \xi_j^T(x_k - y_j) \quad \text{for all } j \in J_k. \quad (5)$$

[10.1.2002-2:04pm] [1-29] [Page No. 4] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

Note that the convexity implies

$$\hat{f}_k(x) \leq f(x) \quad \text{for all } x \in \mathbb{R}^n \text{ and } \alpha_j^k \geq 0 \quad \text{for all } j \in J_k. \tag{6}$$

The next iterate candidate is then defined by

$$y_{k+1} := x_k + d_k, \tag{7}$$

where the search direction d_k is calculated by

$$d_k := \arg\min_{d \in \mathbb{R}^n} \left\{ \hat{f}_k(x_k + d) + \frac{1}{2} d^T M_k d \right\}. \tag{8}$$

The role of the stabilizing term $1/2\, d^T M_k d$ is to guarantee the existence of the solution d_k and to keep the approximation local enough. The regular and symmetric $n \times n$ matrix M_k is intended to accumulate information about the curvature of f in a ball around x_k.

A *serious step*

$$x_{k+1} := y_{k+1} \tag{9}$$

is taken if y_{k+1} is significantly better than x_k, in other words,

$$f(y_{k+1}) \leq f(x_k) + m_L v_k, \tag{10}$$

where $m_L \in (0, 1/2)$ is a line search parameter and

$$v_k = \hat{f}_k(y_{k+1}) - f(x_k) \tag{11}$$

is the predicted descent of f at x_k.

Otherwise, a *null step*

$$x_{k+1} := x_k \tag{12}$$

is taken to improve the cutting plane model $\hat{f}_{k+1}$, since we set $J_{k+1} := J_k \cup \{k+1\}$ in both steps.

The iteration is terminated if

$$v_k \geq -\varepsilon_s, \tag{13}$$

where $\varepsilon_s > 0$ is a final accuracy tolerance supplied by the user.

[10.1.2002–2:04pm] [1–29] [Page No. 5] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

Notice that the problem (8) is still a nonsmooth optimization problem. However, due to its piecewise linear nature, it can be rewritten as a (smooth) quadratic programming subproblem

$$\begin{cases} \text{minimize} & v + \frac{1}{2} d^T M_k d \\ \text{subject to} & -\alpha_j^k + \xi_j^T d \leq v \quad \text{for all} \quad j \in J_k. \end{cases} \tag{QP}$$

By dualizing we get an equivalent problem to (QP) when finding multipliers λ_j^k for $j \in J_k$ solving the problem

$$\begin{cases} \text{minimize} & \frac{1}{2}\left[\sum_{j\in J_k} \lambda_j \xi_j\right]^T M_k^{-1} \left[\sum_{j\in J_k} \lambda_j \xi_j\right] + \sum_{j\in J_k} \lambda_j \alpha_j^k \\ \text{subject to} & \sum_{j\in J_k} \lambda_j = 1 \\ \text{and} & \lambda_j \geq 0. \end{cases} \tag{DP}$$

Notice that in (DP) the sign has been changed and maximization has been converted to minimization. For computational reasons it might in some cases be more efficient to solve (QP) instead of (DP). The following theorem establishes the relations between these two problems.

THEOREM 2.2 *The problems* (QP) *and* (DP) *are equivalent, and they have unique solutions* (d_k, v_k) *and* λ_j^k *for* $j \in J_k$, *respectively, such that*

$$d_k = -\sum_{j\in J_k} \lambda_j^k M_k^{-1} \xi_j, \tag{14}$$

$$v_k = -d_k^T M_k d_k - \sum_{j\in J_k} \lambda_j^k \alpha_j^k. \tag{15}$$

Proof See [11] pp. 115–117. □

Notice that from the computational point of view, the choice of the index set is a crucial point. If the methods stores all the previous subgradients, in other words if $J_k = \{1, \ldots, k\}$, it may cause serious computational difficulties in the form of unbounded storage requirement.

In what follows, we shortly present several versions of bundle methods, which are slight modifications of the general bundle

[10.1.2002–2:04pm] [1–29] [Page No. 6] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

algorithm presented above. For more information of the development of bundle methods, we refer to [15], Chapters XIII–XV.

To avoid technical details, we focus on their main differences in the choice of the cutting plane approximation $\hat{f}_k$, the linearization error α_j^k or the stabilizing matrix M_k.

2.2 Cutting Plane Methods

The history of bundle methods originates from the cutting plane idea, which was developed independently in [16,17]. In the original cutting plane method, the curvature matrix was chosen to be

$$M_k \equiv 0. \tag{16}$$

For this reason the direction-finding subproblems (QP) and (DP) are linear programming problems and thus simpler to solve. It is also obvious that if the original objective function f is piecewise linear or almost piecewise linear, then the cutting plane method may converge in a reliable way and rapidly to the exact global optimum.

However, the lack of a stabilizing quadratic term has the effect that (QP) or (DP) may not have a finite optimum. In addition, the convergence of the cutting plane algorithm can be really slow in some cases (see [18]). This is mainly due to the fact that $\|d_k\|$ may be so large that the new trial point y_{k+1} is far from the previous ones y_j, $j = 1, \ldots, k$. Then y_{k+1} lies in the region where the model $\hat{f}_k$ poorly approximates the original objective f. Furthermore, the cutting plane method behaves very badly when f is differentiable [19].

2.3 Conjugate Subgradient Methods

The next step towards bundle methods was made by the method of conjugate subgradients, which was developed in two journal papers, [18,20], in the same issue. The main idea of the conjugate subgradient method was to calculate the search direction d_k as the projection of 0 onto the set

$$\text{conv}\{-\xi_j | j \in J_k\} \tag{17}$$

approximating the set $-\partial f(x_k)$. This dual approach means that each of the previous subgradients $\xi_j \in \partial f(y_j)$ for $j \in J_k$ is related as if it were a subgradient at the current point x_k, in other words, the linearization error (5) was neglected. On the other hand, no curvature of the objective function was taken into account, thus the search direction finding problems (QP) and (DP) were applied with

$$M_k \equiv I \quad \text{and} \quad \alpha_j^k \equiv 0. \tag{18}$$

The neglect of α_j^k has the effect that the cutting plane model $\hat{f}_k$ is a usable approximation to f only if the trial points y_j are close enough to x_k. For this reason, the choice of the index set $J_k \subset \{1, \ldots, k\}$ is a crucial point of the algorithm. Several *subgradient selection strategies* to choose J_k were proposed, for instance, in [18,20–22] to keep the approximation local enough and limit the number of stored subgradients. In [18,20] the authors, for the first time, introduce the *subgradient aggregation strategy*, which requires storing only a limited number of subgradients. The convergence properties have been analyzed in [23]. However, the numerical experiments have shown that the convergence of the conjugate subgradient method is rather slow in practice (see [24]).

2.4 ε-Steepest Descent Methods

The first method really called "bundle method" was the ε-steepest descent method introduced in [25], where the idea was to combine the cutting plane method with the conjugate gradient method. The method was further developed in [26], where the subgradient aggregation strategy was used to limit the number of stored subgradients, and later in [27]. Contrary to the conjugate subgradient method, some attention to the linearization error (5) was paid. In contrast to the standard treatment of the linearization error as a term appearing in the objective function of (DP), this error was controlled by an extra constraint. Since no curvature of the objective function was taken into account yet, in other words, since

$$M_k \equiv I, \tag{19}$$

[10.1.2002–2:04pm] [1–29] [Page No. 8] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

the dual search direction finding problem (DP) has the form

$$\begin{cases} \text{minimize} & \dfrac{1}{2}\left\| \sum\limits_{j \in J_k} \lambda_j \xi_j \right\|^2 \\ \text{subject to} & \sum\limits_{j \in J_k} \lambda_j = 1 \\ & \lambda_j \geq 0 \\ \text{and} & \sum\limits_{j \in J_k} \lambda_j \alpha_j^k \leq \varepsilon k, \end{cases} \tag{DP$'$}$$

where $\varepsilon_k > 0$ is a controlling parameter. This tolerance controls the radius of the ball in which the cutting plane model is thought to be a good approximation to the objective function. The main difficulty in the ε-steepest descent method is the design of convergent rules for automatic updating of ε_k. However, according to the numerical tests of [24], this method works better in general than the conjugate subgradient method.

2.5 Generalized Cutting Plane Methods

To avoid the difficulties of the above methods when handling the linearization error, the idea of the generalized cutting plane method was introduced in [28] and further developed in [29]. The starting point was the classical cutting plane algorithm in a primal form. The stabilizing term $1/2\|d\|^2$ was added to the objective function to guarantee the existence of the solution and to keep the approximation local enough. Thus the search direction was obtained as a solution of (QP) or (DP) with the choice

$$M_k \equiv I. \tag{20}$$

In spite of different backgrounds, the ε-steepest descent method and the generalized cutting plane method have the following connection: if λ_j^k for $j \in J_k$ are optimal multipliers of (DP) with $M_k \equiv I$, then they solve also the problem (DP$'$) with the choice

$$\varepsilon_k = \sum_{j \in J_k} \lambda_j^k \alpha_j^k. \tag{21}$$

[10.1.2002–2:04pm] [1–29] [Page No. 9] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

The storage-saving subgradient selection and aggregation strategies were introduced in [29], Chapter 2. Especially, the aggregation strategy of [29] has been successfully applied in subsequent methods. When compared with the ε-steepest descent method, the difficulty of selecting the approximation tolerance ε_k was avoided in the generalized cutting plane method. Theoretically in smooth case both of the methods reduces to steepest descent method, which is linearly convergent. The numerical tests of [30] support the superiority of the generalized cutting plane method. However, the tests also disclose its sensitivity to the scaling of the objective function (i.e., multiplication of f by a positive constant).

2.6 Diagonal Variable Metric Bundle Methods

The diagonal variable metric idea was the next improvement in the history of the bundle methods. A weighting parameter was added to the quadratic term of the objective functions in (QP) and (DP) in order to accumulate some second-order information about the curvature of f around x_k. Thus the variable metric matrix M_k took the diagonal form

$$M_k = u_k I, \tag{22}$$

with the weighting paramter $u_k > 0$. Based on the proximal point algorithm of [31] and the work of [32], the *proximal bundle* method was derived in [33], where also an adaptive safeguarded quadratic interpolation technique for updating u_k was introduced. A somewhat similar outcome was concluded in [34], where the *bundle trust region* method was developed combining the bundle idea with the classical trust region method of [35,36]. Furthermore, the *diagonal "poor man's" quasi-Newton* method of [37,38] and the *proximal bundle method algorithm* of [39] based on the Moreau–Yosida regularization (see [40–42]) belong to this class of bundle methods.

The methods deviate mainly in the strategies for updating the weights u_k: the bundle trust region method employs the same safeguarded quadratic interpolation technique as the proximal bundle method, while the diagonal quasi-Newton method relies on a curved search technique (instead of line search). In [43] Kiwiel proved that

[10.1.2002–2:04pm] [1–29] [Page No. 10] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

these methods can find an ε-solution, i.e., a point x_k such that $f(x_k)-f(x^*) \leq \varepsilon$, after at most $O(1/\varepsilon^3)$ objective function and subgradient evaluations.

In order to avoid the solution of the time-consuming quadratic subproblem (QP), the quadratic term $1/2d^T M_k d$ in (8) was replaced in [44] by the D-function of a generalized Bregman function (see [45]).

The numerical tests in [11,33,34,38] demonstrate the obvious progress in the convergence speed of the diagonal variable metric bundle methods when compared to the earlier methods.

2.7 Variable Metric Bundle Methods

The development of second-order methods has been fascinating the researchers in nonsmooth optimization during its whole history. Although the tools of nonsmooth analysis (see, e.g., [46]) already exist, their influence in numerical methodology has not been as fruitful as in the first-order case. However, several attempts to employ

$$M_k \text{ as a full matrix} \tag{23}$$

with some updating scheme have been proposed by various authors. Already in his pioneering work [28], Lemaréchal derived a version of the variable metric bundle method utilizing the classical BFGS secant updating formula from smooth optimization (see [9] p. 135). Due to the disappointing numerical results in [24], this idea was buried nearly for two decades. In [47] the space dilation updating scheme of [12] was adopted from the subgradient method context. Later, based on the Moreau–Yosida regularization, BFGS update and the curved search technique, a reversal quasi-Newton method was proposed in [37,38,48], and based on a proximal point approach, variable metric methods were derived in [49–51]. According to very limited numerical experiments (see, e.g., [47,50]), it seems that the variable metric bundle methods work fairly well. However, when the results proportioned to the extra computational efforts needed with the full matrix algebra, they do not offer a substantial advancement in numerical solution process.

The most recent variable metric bundle method using BFGS update was derived in [52]. The idea of the method is to use only three

subgradients (two calculated at x_k and y_{k+1}, one aggregated, containing information from past iterations). This means that the dimension of the normally time-consuming quadratic programming subproblem (DP) is only three and it can be solved with simple calculations. The numerical tests of [52] show that the method is comparable with the diagonal variable metric bundle methods in the number of function evaluations, but the computational time can be significantly smaller.

2.8 Translated Bundle Methods

Another strategy to exploit higher-order information about f is to keep the stabilizing matrix constant, in other words

$$M_k \equiv I, \tag{24}$$

but move the supporting hyperplane by varying the linearization error (5). The first attempts into this direction were made in [53], where the so-called α-function was introduced, and in [54], where the linearization error α_j^k was replaced by the quantity

$$\beta_j^k := f(y_j) - f(x_k). \tag{25}$$

This idea was extended in [47], where the combination

$$\rho_j^k := \max\left\{(1-\theta)\alpha_j^k, \min\{\alpha_j^k, \beta_j^k\}\right\}, \tag{26}$$

with $\theta \in (0,1)$ was utilized. In the same paper [47] also a quadratic-based translation approach was proposed in the form

$$\rho_j^k := \xi_k^T(x_k - y_j). \tag{27}$$

None of the above methods behaves uniformly better than the others in numerical tests of [47]. According to these limited results, the convergence rate of the translated bundle methods places itself between the ε-steepest descent method and the diagonal variable metric bundle methods.

[10.1.2002–2:04pm] [1–29] [Page No. 12] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

2.9 Tilted Bundle Methods

Based on the work [55], the tilted bundle method was developed in [56] in order to accumulate some second-order information and some interior-point features (see, e.g., [57]) to the proximal bundle method of [33]. The matrix M_k was handled in the original diagonal form

$$M_k \equiv u_k I, \tag{28}$$

with the safeguarded quadratic interpolation technique for updating u_k. The cutting plane model (3) was replaced by the so-called *tilted cutting plane model*

$$\check{f}_k(x) := \max_{j \in J_k} \{f(x_k) + (1 - \theta_j^k)\xi_j^T(x - x_k) - \rho_j^k\} \tag{29}$$

with some tilting parameters $\theta_j^k \in [0, 1 - \kappa]$ for $j \in J_k$ and $\kappa \in [0,1]$. The linearization error (5) was replaced by the combination

$$\rho_j^k := \max\{\check{\alpha}_j^k, \kappa \cdot \alpha_j^k\},$$

where the tilted linearization error is given by

$$\check{\alpha}_j^k := f(x_k) - f(y_j) - (1 - \theta_j^k)\xi_j^T(x_k - y_j).$$

Note that if $\kappa = 1$, then $\theta_j^k = 0$ for all $j \in J_k$ and the method reduces to the proximal bundle method of [33] described in Subsection 2.6. The tilted cutting plane is a real "cutting plane", since it cuts off the parts of the epigraph of f while the standard cutting planes cut nothing. The numerical tests in [56] appear rather promising, but the question how to choose the tilting parameters κ and θ_j^k is still open.

2.10 Level Bundle Methods

Since the level sets of the cutting plane model (3) behave in a rather regular way, a new bundle method variant based on the minimization of the stabilizing quadratic term subject to some level set of $\hat{f}_k$ was

[10.1.2002–2:04pm] [1–29] [Page No. 13] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

proposed in [58]. In other words, the search direction finding problem (8) was replaced by

$$d_k := \arg\min_{d\in\mathbb{R}^n}\left\{\frac{1}{2}d^T M_k d \mid \hat{f}_k(x_k+d) \le f_k^{lev}\right\}, \tag{30}$$

where the target level $f_k^{lev} < f(x_k)$ is chosen to ensure $f_k^{lev} \uparrow \inf f$ as $k \to \infty$. The stabilizing matrix was chosen to be constant

$$M_k \equiv I. \tag{31}$$

Bounded storage versions of the level bundle idea were introduced in [59–61] and in [62], where the global convergence was proven without any compactness assumptions, as in [58]. In numerical tests, the level bundle methods worked in a reliable way (see [58,62]) but they lost, for example, when compared with the diagonal variable metric bundle methods.

2.11 Bundle-Newton Method

The most recent advance in the development of the second-order bundle method was made in [63], where the bundle-Newton method was derived. Instead of the piecewise linear cutting plane model (3), they introduced a quadratic model of the form

$$\tilde{f}_k(x) := \max_{j\in J_k}\left\{ f(y_i) + \xi_j^T(x-y_j) + \frac{1}{2}\rho_j(x-y_j)^T M_j(x-y_j)\right\}. \tag{32}$$

where $\varrho_j \in [0,1]$ is a damping parameter. The search direction finding problem (8) was then replaced by the problem

$$d_k := \arg\min_{d\in\mathbb{R}^n}\{\tilde{f}_k(x_k+d)\}. \tag{33}$$

When we compare the bundle-Newton method to the earlier variable metric bundle methods, we can state that the bundle-Newton method is the "real" second-order method, since every part of the model contains

[10.1.2002–2:04pm] [1–29] [Page No. 14] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

the second-order information in the form of the stabilizing matrix M_j. For the approximation

$$M_j \approx \nabla^2 f(y_j) \tag{34}$$

the authors proposed optionally analytic or finite-difference approximations. Under some additional assumptions it can be shown to maintain superlinear convergence rate. Although the operations with fill matrix demand more storage and time, the numerical experiments in [63] seem to be very promising, especially in piecewise quadratic test cases (see [64]).

2.12 Bundle Methods with Inexact Data

There exist also several variants of bundle methods for the case of inexact objective and subgradient evaluations. Suppose that for every point $x \in \mathbb{R}^n$ and $\varepsilon > 0$ we can compute an approximate function value $f_\varepsilon(x)$ satisfying

$$f(x) - \varepsilon \leq f_\varepsilon(x) \leq f(x)$$

and one ε-subgradient ξ_ε from the ε-subdifferential

$$\partial_\varepsilon f(x) = \left\{\xi \in \mathbb{R}^n \mid f(y) \geq f(x) + \xi^T(y - x) - \varepsilon \quad \text{for all } y \in \mathbb{R}^n\right\}. \tag{35}$$

Then the cutting plane model (3) can be replaced by *approximate cutting plane model*

$$\bar{f}_k(x) := \max_{j \in J_k}\left\{f_{\varepsilon_j}(y_j) + \xi_{\varepsilon_j}^T(x - y_j)\right\}, \tag{36}$$

where $\xi_{\varepsilon j} \in \partial_{\varepsilon j} f(y_j)$ for all $j \in J_k$ and $\varepsilon_j \to 0$. In [65] the generalized cutting plane method of [29] was extended for inexact data, in other words

$$M_k \equiv I. \tag{37}$$

Later the proximal bundle method of [33] with

$$M_k = u_k I \tag{38}$$

[10.1.2002–2:04pm] [1–29] [Page No. 15] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

was extended in [66,67] where also some encouraging numerical experience was reported. The latest variant of bundle methods for inexact data was introduced in [68]. As stated in [65], the extension of these methods to the nonconvex case is difficult, since convexity is essential for avoiding line searches and the noisy function f_ε is not upper semidifferentiable (see [41]).

3 NONCONVEX OPTIMIZATION

In this section we consider an unconstrained optimization problem of the form

$$\begin{cases} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathbb{R}^n, \end{cases} \tag{NP}$$

where the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is supposed to be locally Lipschitz continuous, in other words, for any bounded set $S \subset \mathbb{R}^n$ there exists a (Lipschitz) constant $K > 0$ such that

$$|f(x) - f(y)| \leq K\|x - y\| \quad \text{for all } x, y \in S. \tag{39}$$

Note that f needs not to be convex anymore. The subdifferential of f is now defined (see [69]) as

$$\partial f(x) = \operatorname{conv}\left\{ \lim_{i\to\infty} \nabla f(x^i) \mid x^i \to x \quad \text{and} \quad \nabla f(x^i) \text{ exists} \right\}. \tag{40}$$

In order to guarantee the convergence of the methods, f is supposed to be upper semidifferentiable, as defined in [70]. In other words, for any $x \in \mathbb{R}^n, d \in \mathbb{R}^n$ and sequences $\xi_i \subset \mathbb{R}^n$ and $t_i \subset (0, \infty)$ satisfying $\xi_i \in \partial f(x + t_i d)$ and $t_i \downarrow 0$, one has

$$\limsup_{i\to\infty} \xi_i^T d \geq \liminf_{i\to\infty} [f(x + t_i d) - f(x)]/t_i. \tag{41}$$

For the locally Lipschitz continuous function we have the following necessary optimality condition.

[10.1.2002-2:04pm] [1-29] [Page No. 16] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

THEOREM 3.1 *If the locally Lipschitz continuous* $f : \mathbb{R}^n \to \mathbb{R}$ *attains a local minimum at* x^*, *then*

$$0 \in \partial f(x^*). \tag{42}$$

Proof See, for example, [69] p. 38. □

Nonconvexity brings along many difficulties when compared to the convex case. Firstly, since the optimality condition of Theorem 3.1 is not sufficient without some convexity assumption, the methods cannot guarantee even local optimality of the solutions. Only some candidates, called *stationary* points, satisfying the condition (42), are to be looked for.

Secondly, in the convex case, the cutting plane model (3) was an underestimate for the objective function, and the nonnegative linearization error (5) measured how good an approximation the model was to the original problem (see [6]). In the nonconvex case these properties are not valid anymore: α_j^k may have a tiny (or even negative!) value, although the trial point y_j lies far away from the current iteration point x_k and thus the corresponding subgradient ξ_j is useless. Furthermore, adding more cutting planes (more subgradients to the bundle) might make the local approximation properties of the model function $\hat{f}_k$ worse, due to nonconvexity.

For the reasons mentioned above, the main modifications in the methods needed in the nonconvex case concern the linearization error α_j^k. Since the problem is now more complicated, some inclusions, for instance, so-called resetting strategies, and some changes, for example, in the line search procedure, have to be made in order to guarantee the convergence properties (see [29], Chapters 2 and 3). For simplicity, we concentrate on the changes made in the form of the linearization error in what follows, and introduce two strategies, namely, *subgradient deletion rules* and *subgradient locality measures*, to avoid the difficulties caused by nonconvexity.

3.1 Subgradient Deletion Rules

As mentioned above with the conjugate subgradient methods, the linearization error (5) was neglected and some deletion rules were needed

[10.1.2002–2:04pm] [1–29] [Page No. 17] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

to reduce the past subgradient information in order to localize the approximations. For this reason they were used also for nonconvex problems in [21] and [22], where, for example, the following kind of deletion rules were proposed

$$J_k := \{1 \leq j \leq k \mid \|x_k - y_j\| \leq \delta_k\}, \tag{43}$$

where $\delta_k > 0$ tends to zero. Without any special deletion rules the translated bundle method introduced in [54] was proposed to be suitable also for the nonconvex case in [71], since the substitute linearization error (25) is always nonnegative, in other words

$$\beta_j^k = f(y_j) - f(x_k) \geq 0. \tag{44}$$

More complicated deletion rules were derived in [29], Chapter 4, for generalized cutting plane methods. The linearization error (5) was replaced by its absolute value, in other words,

$$\beta_j^k := |\alpha_j^k| = |f(x_k) - f(y_j) - \xi_j^T(x_k - y_j)| \quad \text{for all} \quad j \in J_k \tag{45}$$

and the algorithm was reset (deleting the old subgradient information) whenever

$$\|d_k\| \leq m_S \cdot \max_{j \in J_k}\{s_j^k\}, \tag{46}$$

where $m_S > 0$ is a reset tolerance supplied by the user and

$$s_j^k := \|x_j - y_j\| + \sum_{i=j}^{k-1} \|x_{i+1} - x_i\| \tag{47}$$

is the distance measure estimating

$$\|x_k - y_j\| \tag{48}$$

without the need to store the trial points y_j.

Later, in [73] Kiwiel introduces his *restricted step* proximal bundle method, where the search direction finding problem (8) was replaced by

$$d_k := \arg \min_{d \in \mathbb{R}^n} \left\{ \hat{f}_k(x_k + d) + \frac{1}{2} d^T M_k d \mid \|d\| \leq \delta_k \right\}, \tag{49}$$

[10.1.2002–2:04pm] [1–29] [Page No. 18] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

where $\delta_k > 0$ tends to zero and

$$M_k = u_k I, \tag{50}$$

with the weighting parameter $u_k > 0$. The resetting test (46) was replaced by

$$u_k \|d_k\| \leq m_S \cdot \max_{j \in J_k} \{s_j^k\} - \sum_{j \in J_k} \lambda_j^k \beta_j^k, \tag{51}$$

where λ_j^k for $j \in J_k$ are the optimal multipliers of (DP) and β_j^k defined by (45).

3.2 Subgradient Locality Measures

Next, we introduce another popular strategy to avoid the difficulties caused by nonconvexity. In order to add some localizing information to the model, the linearization error (5) was replaced by the so-called subgradient locality measure

$$\beta_j^k =: \max\left\{\alpha_j^k, \gamma\|x_k - y_j\|^2\right\} \tag{52}$$

in [26] for the ε-steepest descent method. The distance measure parameter $\gamma \geq 0$ can be set to zero when f is convex. The authors also proposed to use the distance measure (47) avoiding the storing of the trial points y_j, in other words, to replace (52) by

$$\beta_j^k =: \max\left\{\alpha_j^k,\ \gamma(s_j^k)^2\right\}. \tag{53}$$

In [53] the subgradient locality measure, in the form (52) and in [29], Chapter 3, in the form

$$\beta_j^k =: \max\{|\alpha_j^k|,\ \gamma(s_j^k)^2\}, \tag{54}$$

was introduced for the generalized cutting plane method. Furthermore, in [34] the subgradient locality measure, in the form (52) and in [11] in the form (54), was proposed for a diagonal variable metric bundle

[10.1.2002–2:04pm] [1–29] [Page No. 19] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

method (bundle trust region and proximal bundle method, respectively). In [63] the generalization of (54) was introduced in the form

$$\beta_j^k =: \max\left\{|\alpha_j^k|,\ \gamma(s_j^k)^\omega\right\} \tag{55}$$

with $\omega \geq 1$ for the bundle-Newton method and in [72] for the variable metric bundle method.

The latest modification of the subgradient locality measure was proposed in [73] of the form

$$\beta_j^k =: \max\left\{|\alpha_j^k|,\ \gamma_1(s_j^k)^2,\ \gamma_2\|\xi_j\|s_j^k\right\} \tag{56}$$

with $\gamma_1, \gamma_2 \geq 0$. The numerical tests in [73], comparing this Levenberg–Marquardt type-method with the restricted-step method (49), show the slight superiority of the latter method based on the subgradient deletion rules. Especially, in the worst cases, the restricted-step method did not need as many objective functions and subgradient calculations as the Levenberg–Marquardt-type method.

4 CONSTRAINED OPTIMIZATION

In this section we consider a constrained optimization problem of the form

$$\begin{cases} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) \leq 0 \quad \text{for} \quad i = 1, \ldots, m, \end{cases} \tag{CCP}$$

where the objective function $f : \mathbb{R}^n \to \mathbb{R}$ and the constraint functions $g_i : \mathbb{R}^n \to \mathbb{R}$ are supposed to be convex. In the nonconvex case the methods of the previous section can be used.

The problem (CCP) is said to satisfy the *Slater constraint qualification* if

$$g(y) < 0 \tag{57}$$

for some $y \in \mathbb{R}^n$, where $g : \mathbb{R}^n \to \mathbb{R}$ is the *total constraint function* defined by

$$g(x) := \max_{i=1,\ldots,m} g_i(x). \tag{58}$$

For the convex constrained problem we have the following necessary and sufficient Karush–Kuhn–Tucker optimality condition.

THEOREM 4.1 *If the problem* (CCP) *is convex and satisfies the Slater constraint qualification, then it attains its global minimum at* x^* *if and only if there exist Lagrange multipliers* $0 \leq \mu \in \mathbb{R}^m$, *such that* $\mu_{igi}(x^*) = 0$ *for all* $i = 1, \ldots, m$ *and*

$$0 \in \partial f(x^*) + \sum_{i=1}^{m} \mu_i \partial g_i(x^*). \tag{59}$$

Proof See [74] pp. 56–57. □

4.1 Constraint Linearization Method

In bundle methods the most popular way to handle the constraints of (CCP) is to find an unconstrained minimum of the *improvement function*

$$H(x; y) := \max\{ f(x) - f(y),\ g(x)\}. \tag{60}$$

The cutting plane model of the improvement function linearizing both the objective and the constraint functions is defined by

$$\hat{H}_k(x) := \max\{ \hat{f}_k(x) - f(x_k),\ \hat{g}_k(x)\}, \tag{61}$$

where

$$\hat{g}_k(x) := \max_{j \in J_k}\{g(y_j) + (\xi_j^9)^T(x - y_j)\} \tag{62}$$

and $\xi_j^9 \in \partial g(y_j)$. The search direction finding problem (8) is then replaced by

$$d_k := \arg\min_{d \in \mathbb{R}^n}\{\hat{H}_k(x_k + d) + \frac{1}{2} d^T M_k d\}. \tag{63}$$

[10.1.2002–2:04pm] [1–29] [Page No. 21] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

This approach was employed in the context of the generalized cutting plane method, in other words,

$$M_k \equiv I, \tag{64}$$

for instance, in [53]. Storage saving versions utilizing subgradient selection and aggregation strategies were introduced in [29], Chapters 5 and 6, and for linear constraints with nonconvex objective function in [75]. In [11] this constraint linearization approach was used with the proximal bundle method, in other words,

$$M_k = u_k I \tag{65}$$

with $u_k > 0$.

4.2 Exact Penalty Function Method

Another strategy in nonsmooth constrained optimization is to employ the exact penalty function (see [76])

$$e(x; c) := f(x) + \sum_{j=1}^{m} c_i \cdot \max\{g_i(x), 0\}. \tag{66}$$

The cutting plane model of the exact penalty function is then defined by

$$\hat{e}_k(x; c_k) := \hat{f}_k(x) + \sum_{j=1}^{m} c_i^k \cdot \max\{\hat{g}_i^k(x), 0\}, \tag{67}$$

where

$$\hat{g}_i^k(x) := \max_{j \in J_k}\{ g_i(y_j) + (\xi_{i,j}^9)^T (x - y_j)\} \tag{68}$$

and $\xi_{i,j}^9 \in \partial g_i(y_j)$. The search direction is then obtained as

$$d_k := \arg\min_{d \in \mathbb{R}^n} \left\{ \hat{e}_k(x_k + d) + \frac{1}{2} d^T M_k d \right\}. \tag{69}$$

The exact penalty technique described above was used with the generalized cutting plane method in [77,78] and with the proximal bundle

[10.1.2002–2:04pm] [1–29] [Page No. 22] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

method in [79]. For similar approaches, see also [22,80]. The numerical tests in [79] indicate that the exact penalty function method is less sensitive to scaling than the constraint linearization method.

5 MULTICRITERIA OPTIMIZATION

Next we consider a constrained multicriteria optimization problem of the form

$$\begin{cases} \text{minimize} & f(x) = (f_1(x), \ldots, f_s(x))^T \\ \text{subject to} & g_i(x) \leq 0 \quad \text{for} \quad i = 1, \ldots, m, \end{cases} \tag{MCP}$$

where the objective functions $f_i : \mathbb{R}^n \to \mathbb{R}$ and the constraint functions $g_i : \mathbb{R}^n \to \mathbb{R}$ are supposed to be convex.

A feasible point $x^* \in \mathbb{R}^n$ is said to be *Pareto optimal*, if there does not exist another feasible point $x \in \mathbb{R}^n$ such that $f_i(x) \leq f_i(x^*)$ for all $i = 1, \ldots, s$, and $f_j(x) < f_j(x^*)$ for at least one j. Furthermore, $x^* \in \mathbb{R}^n$ is said to be *weakly Pareto optimal*, if there does not exist another feasible point $x \in \mathbb{R}^n$ such that $f_i(x) < f_i(x^*)$ for all $i = 1, \ldots, s$.

For the convex multicriteria problem we have the following necessary and sufficient Karush–Kuhn–Tucker optimality condition.

THEOREM 5.1 *If the problem* (MCP) *is convex and satisfies the Slater constraint qualification, then it attains the weakly Pareto optimal solution at* x^* *if and only if there exist Lagrange multipliers* $0 \leq \lambda \in \mathbb{R}^s$ *with* $\lambda \neq 0$ *and* $0 \leq \mu \in \mathbb{R}^m$ *such that* $\mu_i g_i(x^*) = 0$ *for all* $i = 1, \ldots, m$ *and*

$$0 \in \sum_{j=1}^{8} \lambda_j \partial f_j(x^*) + \sum_{i=1}^{m} \mu_i \partial g_i(x^*). \tag{70}$$

Proof See [81] pp. 50–51. □

In multicriteria optimization, instead of a unique optimal solution, there exists a set of (weakly) Pareto optimal solutions. However, the aim of most of the methods is to look for only one final solution.

[10.1.2002–2:04pm] [1–29] [Page No. 23] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

5.1 Function Scalarization Method

Most methods for multicriteria optimization are based on scalarization of the problem, in other words, the multicriteria problem is replaced by a single-criterion one (see [82]). This scalarization can be realized in very many ways [81]. One possibility is to find the closest point to the ideal criterion vector

$$z_j^* = \min_{g(x)\leq 0} f_j(x) \tag{71}$$

for all $j = 1, \ldots, s$ with respect to some metric. For example, we can use *weighted L_p-metric* ($1 \leq p < \infty$) by forming a convex combination of the objective functions

$$\begin{cases} \text{minimize} & \left(\sum_{j=1}^{s} w_j (f_j(x) - z_j^*)^p \right)^{1/p} \\ \text{subject to} & g_i(x) \leq 0 \quad \text{for} \quad i = 1, \ldots, m. \end{cases} \tag{L_p}$$

or weighted L_∞-metric

$$\begin{cases} \text{minimize} & \max_{1\leq j\leq s} [w_j (f_j(x) - z_j^*)] \\ \text{subject to} & g_i(x) \leq 0 \quad \text{for} \quad i = 1, \ldots, m, . \end{cases} \tag{L_∞}$$

where w_j are some weighting coefficients satisfying $w_j > 0$ and $\sum_{j=1}^{s} w_j = 1$. The solution of the problem (L_p) is Pareto optimal and the solution of the problem (L_∞) weakly Pareto optimal for the original problem (MCP) (see [81], pp. 97–98). After scalarization, the problem (L_p) or (L_∞) can be solved by anyone of the single-criteria bundle methods described before.

5.2 Function Linearization Method

Another approach is to generalize the idea of the constraint linearization method to the multicriteria case. This function linearization method treats the multicriteria problem (MCP) as such, without employing any artificial single objective; scalarization occurs inside the bundle method. The improvement function (60) reads in this case

$$H(x; y) := \max_{j=1,\ldots,s} \{ f_j(x) - f_j(y),\ g(x) \}. \tag{72}$$

If the problem (MCP) is convex and satisfies the Slater constraint qualification, then it attains a weakly Pareto optimal solution at x^* if and only if (see [83])

$$x^* = \arg\min_{x \in \mathbb{R}^n} H(x, x^*). \tag{73}$$

The problem is linearized analogously and the cutting plane model takes the form

$$\hat{H}_k(x) := \max_{j=1,\ldots,s} \{\hat{f}_j^k(x) - f_j(x_k),\ \hat{g}_k(x)\}. \tag{74}$$

Then the search direction finding problem (8) can be written as

$$d_k := \arg\min_{d \in \mathbb{R}^n} \left\{ \hat{H}_k(x_k + d) + \frac{1}{2} d^T M_k d \right\} \tag{75}$$

The function linearization techniques described above were employed in the context of the generalized cutting plane method, in other words,

$$M_k \equiv I, \tag{76}$$

in [84] for unconstrained convex problems, and in [85] for constrained convex and nonconvex problems. In [83], this approach was used with the proximal bundle method, in other words,

$$M_k = u_k I \tag{77}$$

with $u_k > 0$. We should add that the function linearization method may be used as a black box optimizer when designing interactive multicriteria optimization algorithms (see [83,86–88]).

Acknowledgements

The author would like to thank the referees for their extremely careful reading and their valuable suggestions for improving the paper. This research was supported by the grant 8583 of the Academy of Finland.

References

[1] J. Outrata, M. Kočvara and J. Zowe (1998). *Nonsmooth Approach to Optimization Problems With Equilibrium Constraints. Theory, Applications and Numerical Results*. Kluwer Academic Publishers, Dordrecht.

[2] J.J. Moreau, P.D. Panagiotopoulos and G. Strang (Eds.) (1988). In: *Topics in Nonsmooth Mechanics*. Birkhäuser Verlag, Basel.

[3] E.S. Mistakidis and G.E. Stavroulakis (1998). *Nonconvex Optimization in Mechanics. Smooth and Nonsmooth Algorithms, Heuristics and Engineering Applications by the F.E.M*. Kluwer Academic Publisher, Dordrecht.

[4] F.H. Clarke, Yu. S. Ledyaev, R.J. Stern and P.R. Wolenski (1998). *Nonsmooth Analysis and Control Theory*. Springer, New York.

[5] K. Miettinen, M.M. Mäkelä and T. Männikkö (1998). Optimal control of continuous casting by nondifferentiable multiobjective optimization. *Computational Optimization and Applications*, **11**, 177–194.

[6] C. Lemaréchal (1989). In: G.L. Nemhauser, A.H.G. Rinnooy Kan and M.J. Todd (Eds.), *Nondifferentiable Optimization in Optimization*, pp. 529–572. North-Holland, Amsterdam.

[7] J. Haslinger and P. Neittaanmäki (1996). *Finite Element Approximation for Optimal Shape, Material and Topology Design*. J. Wiley & Sons, Chichester.

[8] M.S. Bazaraa and C.M. Shetty (1979). *Nonlinear Programming, Theory and Algorithms*. John Wiley & Sons, New York.

[9] D.P. Bertsekas (1995). *Nonlinear Programming*. Athena Scientific, Belmont.

[10] K.L. Teo and C.J. Goh (1988). On constrained optimization problems with nonsmooth cost functionals. *Applied Mathematics and Optimization*, **18**, 181–190.

[11] M.M. Mäkelä and P. Neittaanmäki (1992). *Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control*. World Scientific Publishing Co., Singapore.

[12] N.Z. Shor (1985). *Minimization Methods for Non-differentiable Functions*. Springer-Verlag, Berlin.

[13] S.P. Uryas'ev (1991). New variable metric algorithms for nondifferentiable optimization problems. *Journal of Optimization Theory and Application*, **71**, 359–388.

[14] R.T. Rockafellar (1970). *Convex Analysis*. Princeton University Press, Princeton, New Jersey.

[15] C. Lemaréchal and J.-B. Hiriart-Urruty (1993). *Convex Analysis and Minimization Algorithms II*. Springer Verlag, Berlin.

[16] E.W. Cheney and A.A. Goldstein (1959). Newton's method for convex programming and tchebycheff approximation. *Numerische Mathematik*, **1**, 253–268.

[17] J.E. Kelley (1960). The cutting plane method for solving convex programs. *SIAM J.*, **8**, 703–712.

[18] P. Wolfe (1975). A method of conjugate subgradients for minimizing nondifferentiable functions. *Mathematical Programming Study*, **3**, 145–173.

[19] Y.-X Yuan (1987). Some results in nonsmooth optimization. *Journal of Computational Mathematics*, **5**, 74–88.

[20] C. Lemaréchal (1975). An Extension of davidon methods to non differentiable problems. *Mathematical Programming Study*, **3**, 95–109.

[21] R. Mifflin (1977). An algorithm for constrained optimization with semismooth functions. *Mathematics of Operations Research*, **2**, 191–207.

[22] E. Polak, D.Q. Mayne and Y. Wardi (1983). On the extension of constrained optimization algorithms from differentiable to nondifferentiable problems. *SIAM Journal on Optimal Control and Optimization*, **21**, 179–203.

[23] J.L. Goffin (1977). On the convergence rates of subgradient optimization methods. *Mathematical Programming*, **13**, 329–347.

[10.1.2002–2:04pm] [1–29] [Page No. 26] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

[24] C. Lemaréchal (1982). Numerical experiments in nonsmooth optimization. In: E.A. Nurminski (Ed.), *Progress in Nondifferentiable Optimization*, pp. 61–84. IIASA-report, Laxemburg, Austria.
[25] C. Lemaréchal (1976). Combining Kelley's and conjucate gradient methods. In: *Abstracts of IX International Symposium on Mathematical Programming*, Budapest, Hungary.
[26] C. Lemaréchal, J.-J. Strodiot and A. Bihain (1981). On a bundle algorithm for nonsmooth optimization. In: O.L. Mangasarian, R.R. Mayer and S.M. Robinson, (Eds.), *Nonlinear Programming*, pp. 245–281. Academic Press, New York 4.
[27] J.-J.Strodiot, V.H. Nguyen and N. Heukemes (1983). ε-Optimal solutions in nondifferentiable convex programming and some related questions. *Mathematical Programming*, **25**, 307–328.
[28] C. Lemaréchal (1978). *Nonsmooth Optimization and Descent Methods*. IIASA-report, Laxemburg, Austria.
[29] K.C. Kiwiel (1985). *Method of Descent for Nondifferentiable Optimization*. Springer-Verlag, Berlin.
[30] M.M. Mäkelä (1989). *Methods and Algorithms for Nonsmooth Optimization*, *Reports on Applied Mathematics and Computing*, No. 2. University of Jyväskylä, Department of Mathematics.
[31] R.T. Rockafellar (1976). Monotone operators and the proximal point algorithm. *SIAM Journal on Optimal Control and Optimization*, **14**, 877–898.
[32] A.Auslender (1987). Numerical methods for nondifferentiable convex optimization. *Mathematical Programming Study*, **30**, 102–126.
[33] K.C. Kiwiel (1990). Proximity control in bundle methods for convex nondifferentiable optimization. *Mathematical Programming*, **46**, 105–122.
[34] H. Schramm and J. Zowe (1992). A version of the bundle idea for minimizing a nonsmooth functions: conceptual idea, convergence analysis, numercial results. *SIAM Journal on Optimization*, **2**, 121–152.
[35] K. Levenberg (1944). A Method for the solution of certain nonlinear problems in least squares. *Quart. Appl. Math.*, **2**, 164–166.
[36] D.W. Marquardt (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, **11**, 431–441.
[37] C. Lemaréchal and C. Sagastizábal (1994). An approach to variable metric bundle methods. In: J. Henry and J.P. Yvon (Eds.), *Lecture Notes in Control and Information Sciences*, Vol. 197, pp. 144–162. Springer-Verlag, New York.
[38] C. Lemaréchal and C. Sagastizábal (1997). Variable metric bundle methods: from conceptual to implementable forms. *Mathematical Programming*, **76**, 393–410.
[39] J.R. Birge, L. Qi and Z. Wei (1998). Convergence analysis of some methods for minimizing a nonsmooth convex function. *Journal of Optimization Theory and Applications*, **97**, 357–383.
[40] J.J. Moreau (1965). Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, **93**, 273–299.
[41] K. Yosida (1964). *Functional Analysis*. Springer-Verlag, Berlin.
[42] C. Lemaréchal and C. Sagastizábal (1997). Practical aspects of the moreu-yosida regularization: theoretical preliminaries. *SIAM Journal on Optimization*, **7**, 367–385.
[43] K.C. Kiwiel (2000). Efficiency of proximal bundle methods. *Journal of Optimization Theory and Applications*, **104**, 589–603.
[44] K.C. Kiwiel (1999). A bundle bregman proximal method for convex nondifferentiable minimization. *Mathematical Programming*, **85**, 241–258.
[45] K.C. Kiwiel (1997). Proximal minimization methods with generalized bregman functions. *SIAM Journal on Optimal Control and Optimization* **35**, 1142–1168.
[46] R.A. Poliquin and R.T. Rockafellar (1996). Generalized hessian properties of regularized nonsmooth functions. *SIAM Journal on Optimization*, **6**, 1121–1137.

[10.1.2002–2:04pm] [1–29] [Page No. 27] FIRST PROOFS I:/T&F Gordon & Breach/Goms/GOMS-11027.3d Optimization Methods and Software (GOMS) Paper: GOMS-11027 Keyword

[47] M. Gaudiso and M. F. Monaco (1992). Variants to the cutting plane approach for convex nondifferentiable optimization. *Optimization*, **25**, 65–75.
[48] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal and C. Sagastizábal (1995). A family of variable metric proximal methods. *Mathematical Programming*, **68**, 15–47.
[49] R. Mifflin (1996). A quasi-second-order proximal bundle alogrithm. *Mathematical Programming*, **73**, 51–72.
[50] X. Chen and M. Fukushima (1999). Proximal quasi-newton methods for nondifferentiable convex optimization. *Mathematical Programming*, **85**, 313–334.
[51] J.V. Burke and M. Qian (2000). On the super-linear convergence of the variable metric proximal point algorithm using broyden and bfgs matrix secant updating. *Mathematical Programming*, **88**, 157–181.
[52] L. Lukšan and J. Vlček (1999). Globally convergent variable metric method for convex nonsmooth unconstrained minimization. *Journal of Optimization Theory and Applications*, **102**, 593–613.
[53] R. Mifflin (1982). A Modification and an extension of lemaréchal's algorithm for nonsmooth minimization. *Mathematical Programming Study*, **17**, 77–90.
[54] M. Gaudioso and M.F. Monaco (1982). A bundle type approach to the unconstrained minimization of convex nonsmooth functions. *Mathematical Programming*, **23**, 216–226.
[55] V.N. Tarasov and N.K. Popva (1984). A modification of the cutting-plane method with accelerated convergence. In: V.F. Demyanov and D. Pallaschke (Eds.), *Nondifferentiable Optimization: Motivations and Applications*, pp. 284–290. Springer-Verlag, Berlin.
[56] K.C. Kiwiel (1991). A tilted cutting plane proximal bundle method for convex nondifferentiable optimization. *Operations Research Letters*, **10**, 75–81.
[57] N. Megiddo (1989) Progress in mathematical programming. *Interior Point and Related Methods*. Springer-Verlag, New York.
[58] C. Lemaréchal, A. Nemirovskii and Yu. Nesterov (1995). New variants of bundle methods. *Mathematical Programming*, **69**, 111–147.
[59] U.Brännlund (1994). A descent method with relaxation type step. In: J. Henry and J.P. Yvon (Eds.), *Lecture Notes in Control and Information Sciences*, Vol. 197. pp. 177–186. Springer-Verlag, New York.
[60] K.C. Kiwiel (1995). Proximal level bundle methods for convex nondifferentiable optimization, saddle-point problems and variational inequalities. *Mathematical Programming*, **69**, 89–109.
[61] U. Brännlund, K.C. Kiwiel and P.O. Lindberg (1995). A descent proximal level bundle method for convex nondifferentiable optimization. *Operations Research Letters*, **17**, 121–126.
[62] U. Brännlund, K.C. Kiwiel and P.O. Lindberg (1996). Preliminary computational experience with a descent level method for convex nondifferentiable optimization. In: J. Doležal and J. Fidler (Eds.), *System Modelling and Optimization*, pp. 387–394. Chapman and Hall, London.
[63] L. Lukšan and J. Vlček (1998). A bundle-newton method for nonsmooth unconstrained minimization. *Mathematical Programming*, **83**, 373–391.
[64] M.M. Mäkelä, M. Miettinen, L. Lukšan and J. Vlček (1999). Comparing nonsmooth nonconvex bundle methods in solving hemivariational inequalities. *Journal of Global Optimization*, **14**, 117–135.
[65] K.C. Kiwiel (1985). An algorithm for nonsmooth convex minimization with errors. *Mathematics of Computation*, **45**, 173–180.
[66] K.C. Kiwiel (1991). Approximations in decomposition of large-scale convex programs via a nondifferentiable optimization method. In: Ü. Jaaksoo and V.I. Utkin (Eds.), *Proceedings of the 11th Triennal IFAC World Congress*. pp. 161–166. Pergamon Press, Oxford.

[67] K.C. Kiwiel (1995). Approximations in proximal bundle methods and decomposition of convex programs. *Journal of Optimization Theory and Applications*, **84**, 529–548.
[68] C.I. Fábián (2000). Bundle-type methods for inexact data. In *Proceedings of the XXIV Hungarian Operations Research Conference, CEJOR Cent. Eur. J. Oper. Res.*, **8**, 35–55.
[69] F.H. Clarke (1983). *Optimization and Nonsmooth Analysis*. J. Wiley & Sons, New York.
[70] A. Bihain (1984). Optimization of upper semidifferentiable functions. *Journal of Optimization Theory and Applications*, **4**, 545–568.
[71] M. Gaudioso and M. F. Monaco (1988). *Some Techniques for Finding the Search Direction in Nonsmooth Minimization Problems*. Dipartimento di Sistemi, Universita' della Calabria, Report 75.
[72] J. Vlček and L. Lukšan (1999). *Globally Convergent Variable Metric Method for Nonconvex Nondifferentiable Unconstrained Minimization*. University of Jyväskylä, Department of Mathematical Information Technology, Report B 8.
[73] K.C. Kiwiel (1996). Restricted step and levenberg-marquardt techniques in proximal bundle methods for nonconvex nondifferentiable optimization. *SIAM Journal on Optimization*, **6**, 227–249.
[74] R.T. Rockafellar (1981). *The Theory of Subgradients and its Applications to Problems of Optimization. Convex and Nonconvex Functions*. Heldermann Verlag, Berlin.
[75] K.C. Kiwiel (1986). A method of linearizations for linearly constrained nonconvex nonsmooth minimization. *Mathematical Programming*, **34**, 175–187.
[76] B.N. Pshenichnyi and Yu. M. Danilin (1978). *Numerical Methods for Extremal Problems*. Mir, Moscow.
[77] K.C. Kiwiel (1985). An exact penalty function method for nonsmooth constrained convex minimization problems. *IMA Journal of Numercial Analysis*, **5**, 111–119.
[78] K.C. Kiwiel (1987). A constraint linearization method for nondifferentiable convex minimization. *Numerische Mathematik*, **51**, 395–414.
[79] K.C. Kiwiel (1991). Exact penalty functions in proximal bundle methods for constrained convex nondifferentiable minimization. *Mathematical Programming*, **52**, 285–302.
[80] A. Auslender, J.T. Crouzeix and P. Fedit (1987). Penalty-proximal methods in convex programming. *Journal of Optimization Theory and Applications*, **55**, 1–21.
[81] K. Miettinen (1999). *Nonlinear Multiobjective Optimization*. Kluwer Academic Publishers, Boston.
[82] A.P. Wierzbicki (1977). Basic properties of scalarization functionals for multiobjective optimization. *Matematische Operations for Schung und Statistik, Ser. Optimization*, **8**, 55–60.
[83] K. Meittinen and M.M. Mäkelä (1995). Interactive bundle-based method for nondifferentiable multiobjective optimization: NIMBUS. *Optimization*, **34**, 231–246.
[84] K.C. Kiwiel (1985). A descent method for nonsmooth convex multiobjective minimization. *Large Scale Systems*, **8**, 119–129.
[85] S. Wang (1989). Algorithms for multiobjective and nonsmooth optimization. In: P. Kleinschmidt, F.J. Radermacher, W. Schweitzer and H. Wildermann (Eds.), *Methods of Operations Research*, Vol. 58, pp. 131–142. Athenäum Verlag, Frankfurt am Main.
[86] H. Mukai (1980). Algorithms for multicriterion optimization. *IEEE Transactions on Automatic Control*, ac-**25**, 177–186.
[87] K. Miettinen and M.M. Mäkelä (2000). Interactive multiobjective optimization system WWW-NIMBUS on the internet. *Computers and Operations Research*, **27**, 709–723.
[88] K. Miettinen and M.M. Mäkelä (1999). Comparative evaluation of some interactive reference point-based methods for multi-objective optimization. *Journal of the Operational Research Society*, **50**, 949–959.