Working Paper

A Bundle Method for Minimizing a Sum of Convex Functions with Smooth Weights

Krzysztof C. Kiwiel

WP-94-13
March 1994

IIASA International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 715210 □ Telex: 079 137 iiasa a □ Telefax: +43 2236 71313

A Bundle Method for Minimizing a Sum of Convex Functions with Smooth Weights

Krzysztof C. Kiwiel

WP-94-13
March 1994

International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 715210 □ Telex: 079 137 iiasa a □ Telefax: +43 2236 71313

A bundle method for minimizing a sum of convex functions with smooth weights*

Krzysztof C. Kiwiel†

March 14, 1994

Abstract

We give a bundle method for minimizing a (possibly nondifferentiable and nonconvex) function $h(x) = \sum_{i=1}^m p_i(x)f_i(x)$ over a closed convex set in $\mathbb{R}^n$, where p_i are nonnegative and smooth and f_i are finite-valued convex. Such functions arise in certain stochastic programming problems and scenario analysis. The method finds search directions via quadratic programming, using a polyhedral model of h that involves current linearizations of p_i and polyhedral models of f_i based on their accumulated subgradients. We show that the method is globally convergent to stationary points of h. The method exploits the structure of h and hence seems more promising than general-purpose bundle methods for nonconvex minimization.

Key words. Nondifferentiable optimization, stochastic programming, bundle methods, semismooth functions.

1 Introduction

We present a method for solving the nondifferentiable optimization (NDO) problem

$$\text{minimize} \quad h(x) := \sum_{i=1}^m p_i(x)f_i(x) \quad \text{over all} \quad x \in S, \tag{1.1}$$

where S is a nonempty closed convex set in $\mathbb{R}^n$, $p_i : \mathbb{R}^n \to \mathbb{R}_+$ are nonnegative continuously differentiable and $f_i : \mathbb{R}^n \to \mathbb{R}$ are convex and possibly nondifferentiable, for $i = 1{:}m$ $(= 1, \ldots, m)$. We suppose that at each $x \in S$ we can calculate the gradient $\nabla p_i(x)$ of p_i and an arbitrary subgradient $g_{f_i}(x) \in \partial f_i(x)$ of f_i, $i = 1{:}m$.

The method is an extension of one for the convex case (all p_i constant) given in [Kiw90] and exploits some ideas of [Kiw86] for handling nonconvexity. It is a descent method which finds search directions via quadratic programming (QP) subproblems. Each subproblem is obtained by linearizing each p_i at the current iterate and constructing a polyhedral model of each f_i from its accumulated subgradients. An inexact line search ensures global convergence of the method to stationary points of h over S.

The special convex case of problem (1.1) with constant $p_i(x)$, $i = 1{:}m$, can be solved even in the large-scale case by several methods of varying efficiency; cf. [ErW88, HUL93,

*Research supported by the Polish Academy of Sciences and the International Institute for Applied Systems Analysis, Laxenburg, Austria.

†Systems Research Institute, Newelska 6, 01–447 Warsaw, Poland (kiwiel@ibspan.waw.pl)

Kiw90, Rus86, Rus93b, ScZ92]. In general, problem (1.1) is nonconvex but semismooth [Mif77b], so it could be solved by other general-purpose bundle methods for NDO [Kiw85, Kiw92, Mif82, ScZ92]. However, such algorithms would not be very efficient, since they cannot exploit the special structure of h. In particular, our method uses only the current linearizations of p_i for search direction finding and, hence, does not need any complicated techniques for handling nonconvexity of h. Moreover, when all the weights p_i have small gradients (are almost constant) then our method automatically gets close to its efficient predecessors for the convex case [Kiw90, Rus86, Rus93b].

We should add that problem (1.1) has been suggested to us by A. Ruszczyński [Rus93a] as an important extension of stochastic programming problems (cf. [ErW88]). In classical versions of such problems, each p_i is the (constant) probability of an event (scenario [RoW91]) with cost $f_i(x)$, and one minimizes the expected cost $h(x)$ over all feasible decisions x in S. Our framework allows the probability of a future event to depend on the decision taken at the first stage. It seems that such models could find widespread applications, once suitable software for their solution becomes available.

The paper is organized as follows. In §2 we state our method for the simplest case of $m = 1$. Its global convergence is established in §3. The extension to $m > 1$ is described in §4.

We use the following notation and terminology. $\langle \cdot, \cdot \rangle$ and $|\cdot|$ denote the standard inner product and norm respectively in a given Euclidean space. δ_S is the *indicator* function of S ($\delta_S(x) = 0$ if $x \in S$, ∞ otherwise). For any convex function $f : \mathbb{R}^n \to \mathbb{R}$, $\partial_\epsilon f(x) = \{g : f(y) \geq f(x) + \langle g, y - x \rangle - \epsilon \ \forall y\}$ is the *ϵ-subdifferential* of f at x for each $\epsilon \geq 0$, $\partial f(x) = \partial_0 f(x)$ being the ordinary subdifferential. The mapping $\partial . f(\cdot)$ is locally bounded and upper semicontinuous [Kiw85, HUL93]. Under our assumptions, the function h (cf. (1.1)) has at each x the *Clarke subdifferential* (generalized gradient [Cla83])

$$\partial h(x) = \sum_{i=1}^{m} [p_i(x) \partial f_i(x) + f_i(x) \nabla p_i(x)], \tag{1.2}$$

and h is semismooth [Mif77b]. We say that a point $\bar{x} \in S$ is *stationary for h on S* if $0 \in \partial h(\bar{x}) + \partial \delta_S(\bar{x})$, where $\partial \delta_S$ is the normal cone operator of S; this is a necessary condition for $\bar{x}$ to minimize h over S [Cla83, Mif77b].

2 The method

To simplify notation, we now consider the case of $m = 1$ (extensions to $m > 1$ are deferred till §4). Thus we wish to minimize $h(x) = p(x) f(x)$ over $x \in S$, where $p : \mathbb{R}^n \to \mathbb{R}_+$ is continuously differentiable and $f : \mathbb{R}^n \to \mathbb{R}$ is convex. Given $y \in S$ and $g_f(y) \in \partial f(y)$, let

$$\bar{f}(x; y) = f(y) + \langle g_f(y), x - y \rangle , \tag{2.1}$$

$$\alpha_f(x, y) = f(x) - \bar{f}(x; y) \geq 0 \tag{2.2}$$

denote the value at x of the *linearization* of f computed at y and its *error* at x respectively ($\alpha_f \geq 0$ by convexity). The method generates a sequence $\{x^k\}_{k=1}^{\infty}$ in S that should converge to a minimizer of $h + \delta_S$, and *trial points* $\{y^k\} \subset S$ at which linearizations of f are computed. Let $f^j(\cdot) = \bar{f}(\cdot; y^j)$ and $g_f^j = g_f(y^j)$ for all j. To deal with nondifferentiability of f and $h = pf$, at iteration k the method uses their polyhedral models

$$\check{f}^k(x) = \max\{f^j(x^k) + \left\langle g_f^j, x - x^k \right\rangle : j \in J^k\}, \tag{2.3a}$$

$$\check{h}^k(x) = p(x^k) \check{f}^k(x) + f(x^k) \left\langle \nabla p(x^k), x - x^k \right\rangle , \tag{2.3b}$$

where $J^k \subset \{1\!:\!k\}$, $k \in J^k$. The kth *search direction* from $x^k \in S$ is chosen as

$$d^k = \arg\min\{\check{h}^k(x^k + d) + u^k|d|^2/2 : x^k + d \in S\}, \tag{2.4}$$

where the weight $u^k > 0$ should keep $x^k + d^k$ in the region where $\check{h}^k$ is a close approximation to h. The *predicted descent*

$$v^k = \check{h}^k(x^k + d^k) - h(x^k) \tag{2.5}$$

is employed by a line search to find the next x^{k+1} and y^{k+1}.

Note that (2.4) can be solved by finding (d^k, v^k) to

$$\begin{array}{ll} \text{minimize} & u^k|d|^2/2 + v \quad \text{over all } (d,v) \in \mathbb{R}^{n+1} \\ \text{satisfying} & -p(x^k)\alpha_j^k + \left\langle p(x^k)g_f^j + f(x^k)\nabla p(x^k), d\right\rangle \le v, \quad j \in J^k, \\ & x^k + d \in S, \end{array} \tag{2.6}$$

where $\alpha_j^k = f(x^k) - f^j(x^k) \ge 0$ (cf. (2.2)). Denote the Lagrange multipliers of (2.6) by λ_j^k, $j \in J^k$. Let $\check{y}^{k+1} = x^k + d^k = \arg\min_S \check{h}^k$. As in [Kiw90], using the fact $p(x^k) \ge 0$, from (2.3) and the optimality condition $0 \in \partial[\check{h}^k + u^k|\cdot - x^k|^2/2 + \delta_S](\check{y}^{k+1})$ for (2.4) we deduce the existence of $\tilde{g}_f^k \in \partial\check{f}^k(\check{y}^{k+1})$, $\tilde{g}_h^k = p(x^k)\tilde{g}_f^k + f(x^k)\nabla p(x^k) \in \partial\check{h}^k(\check{y}^{k+1})$ and $\tilde{g}_S^k \in \partial\delta_S(\check{y}^{k+1})$ such that the *aggregate linearizations* $\tilde{f}^k(\cdot) = \check{f}^k(\check{y}^{k+1}) + \left\langle \tilde{g}_f^k, \cdot - \check{y}^{k+1}\right\rangle$, $\tilde{h}^k(\cdot) = \check{h}^k(\check{y}^{k+1}) + \left\langle \tilde{g}_h^k, \cdot - \check{y}^{k+1}\right\rangle$ and $\tilde{\delta}_S^k(\cdot) = \left\langle \tilde{g}_S^k, \cdot - \check{y}^{k+1}\right\rangle$ minorize f, $\check{h}^k$ and δ_S respectively and $p(x^k)\tilde{g}_f^k + f(x^k)\nabla p(x^k) + \tilde{g}_S^k + u^k d^k = 0$. Moreover, letting

$$\tilde{g}^k = \tilde{g}_h^k + \tilde{g}_S^k = p(x^k)\tilde{g}_f^k + f(x^k)\nabla p(x^k) + \tilde{g}_S^k = -u^k d^k, \tag{2.7}$$

$\tilde{\alpha}_f^k = f(x^k) - \tilde{f}^k(x^k) \ge 0$ (cf. $f \ge \tilde{f}^k$), $\tilde{\alpha}_h^k = p(x^k)\tilde{\alpha}_f^k \ge 0$, $\tilde{\alpha}_S^k = -\tilde{\delta}_S^k(x^k) = \left\langle \tilde{g}_S^k, d^k\right\rangle \ge 0$ (cf. $0 = \delta_S(x^k) \ge \tilde{\delta}_S(x^k)$) and

$$\tilde{\alpha}^k = \tilde{\alpha}_h^k + \tilde{\alpha}_S^k = p(x^k)\tilde{\alpha}_f^k + \tilde{\alpha}_S^k \ge 0, \tag{2.8}$$

we have

$$\tilde{g}_f^k \in \partial_{\tilde{\alpha}_f^k} f(x^k), \tag{2.9}$$

$$-v^k = u^k|d^k|^2 + \tilde{\alpha}^k. \tag{2.10}$$

Indeed, (2.9) follows from $f \ge \tilde{f}^k$, and (2.10) from $v^k = \tilde{h}^k(\check{y}^{k+1}) - h(x^k) = \tilde{h}^k(x^k) - h(x^k) + \left\langle \tilde{g}_h^k, d^k\right\rangle = -\tilde{\alpha}_h^k - \tilde{\alpha}_S^k + \left\langle \tilde{g}^k, d^k\right\rangle$ (cf. (2.5), (2.7)). Thus $v^k \le 0$. If $v^k = 0$ then either $p(x^k) > 0$, $\tilde{\alpha}_f^k = 0$ and $\tilde{g}_f^k \in \partial f(x^k)$ (cf. (2.8)-(2.10)), or $p(x^k) = 0$, and $d^k = 0$ (cf. (2.7)) imply $\tilde{g}_S^k \in \partial\delta_S(x^k)$ and $0 \in \partial h(x^k) + \partial\delta_S(x^k)$, so x^k is stationary and the method may stop. Further, we note that for $\hat{J}^k = \{j \in J^k : \lambda_j^k \ne 0\}$, the selected model

$$\hat{f}^k(x) = \max\{f^j(x) : j \in \hat{J}^k\} \tag{2.11}$$

may *a posteriori* replace $\check{f}^k$ in (2.3) without changing (2.4)-(2.5), since $\hat{f}^k(\check{y}^{k+1}) = \check{f}^k(\check{y}^{k+1})$ and $\tilde{g}_f^k = \sum_{j\in\hat{J}^k}\lambda_j^k g_f^j \in \partial\hat{f}^k(\check{y}^{k+1}) \subset \partial\check{f}^k(\check{y}^{k+1})$, using $\lambda_j^k \ge 0$, $\lambda_j^k[\check{f}^k(\check{y}^{k+1}) - f^j(\check{y}^{k+1})] = 0$, $j \in J^k$, $\sum_j \lambda_j^k = 1$. Thus $\hat{f}^k$ incorporates all the active linearizations, and the inactive ones may be dropped to save storage.

We may now state the method in detail.

Algorithm 2.1.

Step 0 (*Initiation*). Select an initial point $x^1 \in S$, a final stationarity tolerance $\epsilon_{\text{opt}} \ge 0$, positive linesearch parameters κ_L, κ_R and κ_v satisfying $\kappa_L + \kappa_v < \kappa_R < 1$, a stepsize bound $\bar{t} \in (0, 1]$, lower and upper bounds for weights $0 < u_{\min} \le u_{\max}$, an initial weight $u^1 \in [u_{\min}, u_{\max}]$ and the maximum number of stored subgradients $M \ge n + 2$. Set $y^1 = x^1$, $J^1 = \{1\}$, $f^1 = f(y^1)$, $g_f^1 = g_f(y^1)$. Set the counters $k = 1$, $l = 0$ and $k(0) = 1$.

Step 1 (*Direction finding*). Find the solution (d^k, v^k) of (2.6) and its multipliers λ_j^k such that the set $\hat{J}^k = \{j \in J^k : \lambda_j^k \neq 0\}$ satisfies $|\hat{J}^k| \leq M - 1$.

Step 2 (*Stopping criterion*). If $v^k \geq -\epsilon_{\text{opt}}$, terminate; otherwise, continue.

Step 3 (*Line search*). By a line search procedure as given below, find two stepsizes $0 \leq t_L^k \leq t_R^k \leq 1$ such that $x^{k+1} = x^k + t_L^k d^k$ and $y^{k+1} = x^k + t_R^k d^k$ satisfy

$$h(x^{k+1}) \leq h(x^k) + \kappa_L t_L^k v^k, \tag{2.12}$$

and either a *descent step* is taken: $t_L^k = t_R^k > 0$ and either $t_L^k \geq \bar{t}$ or

$$\begin{aligned} \kappa_v |v^k| \leq p(x^k)\alpha_f(x^k, x^{k+1}) + [p(x^{k+1}) - p(x^k)] \left\langle g_f(x^{k+1}), d^k \right\rangle \\ + [f(x^{k+1}) \left\langle \nabla p(x^{k+1}), d^k \right\rangle - f(x^k) \left\langle \nabla p(x^k), d^k \right\rangle], \end{aligned} \tag{2.13}$$

or a *null step* occurs: $t_L^k = 0$ (i.e., $x^{k+1} = x^k$) and

$$-p(x^{k+1})\alpha_f(x^{k+1}, y^{k+1}) + \left\langle p(x^{k+1}) g_f(y^{k+1}) + f(x^{k+1}) \nabla p(x^{k+1}), d^k \right\rangle \geq \kappa_R v^k. \tag{2.14}$$

If $t_L^k > 0$, set $k(l+1) = k+1$ and increase the counter of descent steps l by 1.

Step 4 (*Linearization updating*). Select $\tilde{J}^k$ such that $\hat{J}^k \subset \tilde{J}^k \subset J^k$ and $|\tilde{J}^k| \leq M-1$, set $J^{k+1} = \tilde{J}^k \cup \{k+1\}$, $g_f^{k+1} = g_f(y^{k+1})$, $f_{k+1}^{k+1} = \bar{f}(x^{k+1}; y^{k+1})$ and $f_j^{k+1} = f_j^k + \left\langle g_f^j, x^{k+1} - x^k \right\rangle$ for $j \in \tilde{J}^k$ (so that $\alpha_j^{k+1} = f(x^{k+1}) - f_j^{k+1}$, $j \in J^{k+1}$).

Step 5 (*Weight updating*). If $x^{k+1} \neq x^k$, select $u^{k+1} \in [u_{\min}, u_{\max}]$; otherwise, either set $u^{k+1} = u^k$ or choose $u^{k+1} \in [u^k, u_{\max}]$.

Step 6. Increase k by 1 and go to Step 1.

A few comments on the method are in order. If S is described by finitely many linear inequalitites then Step 1 may use the QP methods of [Kiw89, Kiw94], which can solve efficiently sequences of related subproblems (2.6). Step 2 is justified by stationarity estimates following from (2.7)–(2.10), i.e., $\tilde{\alpha}^k$ and $u^k|d^k|$ measure how far the null vector is from $\partial h(x^k) + \partial \delta_S(x^k)$. Step 3 is entered with $v^k < 0$ and $x^k + d^k \in S$, but d^k need not be a descent direction for h at x^k. Whenever descent occurs, criteria (2.12)–(2.13) make t_L^k sufficiently large so that $h(x^{k+1})$ is significantly better than $h(x^k)$. On the other hand, each null step collects a new linearization of f to modify significantly the next direction finding subproblem (cf. (2.6) and (2.14)). At Step 4 one may let $J^{k+1} = J^k \cup \{k+1\}$ and then, if necessary, drop from J^{k+1} an index $j \in J^k \setminus \hat{J}^k$ with the largest error α_j^{k+1}. Step 5 may use the weight updating procedure of [Kiw90].

The following procedure may be used at Step 3, with $x = x^k$, $d = d^k$, $v = v^k$.

Procedure 2.2 (*line search*).

(i) Set $t_L = 0$ and $t = t_U = 1$. Choose $\kappa \in (\kappa_L + \kappa_v, \kappa_R)$.

(ii) If $h(x + td) \leq h(x) + \kappa t v$ set $t_L = t$, otherwise $t_U = t$.

(iii) If $h(x + td) \leq h(x) + \kappa_L t v$ and either $t \geq \bar{t}$ or $p(x)\alpha_f(x, x + td) + [p(x + td) - p(x)] \langle g_f(x + td), d \rangle + [f(x + td) \langle \nabla p(x + td), d \rangle - f(x) \langle \nabla p(x), d \rangle] \geq -\kappa_v v$, set $t_L^k = t_R^k = t_L$ and return.

(iv) If $t \leq \bar{t}$ and $-p(x)\alpha_f(x, x+td) + \langle p(x) g_f(x + td) + f(x) \nabla p_i(x), d \rangle \geq \kappa_R v$ set $t_R^k = t$, $t_L^k = 0$ and return.

(v) Choose $t \in [t_L + 0.1(t_U - t_L), t_U - 0.1(t_U - t_L)]$ and go to (ii).

Lemma 2.3. *Procedure 2.2 exits with t_L^k and t_R^k satisfying the requirements of Step 3.*

Proof. If the search does not terminate, there exists $\hat{t}$ such that $t_L \uparrow \hat{t}$ and $t_U \downarrow \hat{t}$. We consider two cases. First, if $\hat{t} > 0$ then, since $t_L \uparrow \hat{t}$, $t_U \downarrow \hat{t}$, $\kappa v < \kappa_L v < 0$, and h is continuous, we eventually have $h(x + td) \leq h(x) + \kappa_L t v$ at step (iii), with $t = t_U$ for infinitely many such t. Therefore, such t satisfy

$$h(x + td) > h(x) + \kappa t v, \tag{2.15a}$$

$t < \bar{t}$ and $p(x)\alpha_f(x, x+td) + [p(x+td) - p(x)]\,\langle g_f(x+td), d\rangle + [f(x+td)\,\langle \nabla p(x+td), d\rangle - f(x)\,\langle \nabla p(x), d\rangle] < -\kappa_v v$; hence, since also

$$-p(x)\alpha_f(x, x+td) + \langle p(x) g_f(x+td) + f(x)\nabla p_i(x), d\rangle < \kappa_R v,$$

we have

$$\langle p(x+td) g_f(x+td) + f(x+td)\nabla p(x+td), d\rangle < (\kappa_R - \kappa_v) v. \tag{2.15b}$$

Secondly, if $\hat{t} = 0$ (i.e., $t \downarrow 0$), then we have (2.15a) for all $t = t_U$, and (2.15b) for small t, since $t \downarrow 0$, $f(x + td) \to f(x)$, $\langle g_f(x+td), d\rangle$ is bounded, $\alpha_f(x, x+td) \to 0$, $p(x+td) \to p(x)$, $\nabla p(x+td) \to \nabla p(x)$, while $-v > 0$, $\kappa_v > 0$. Thus in both cases (2.15) holds for infinitely many $t \downarrow \hat{t}$, so a contradiction can be established as in the proofs of [Mif77a, Thm 4.1] or [Kiw85, Lem. 3.3.3] between the semismoothness of h and the fact that $v < 0$ and $\kappa < \kappa_R - \kappa_v$. Therefore, the search terminates. □

3 Convergence

In this section we show that each accumulation point of $\{x^k\}$ is stationary for h on S. We assume, of course, that the tolerance $\epsilon_{\text{opt}} = 0$. Then (cf. §2) upon termination $0 \in \partial h(x^k) + \partial \delta_S(x^k)$. Hence we may suppose that the algorithm does not terminate.

We first show that $|v^k|$ measures the stationarity of x^k.

Lemma 3.1. *Suppose there exists a point $x^\infty \in S$ and an infinite set $K \subset \{1, 2, \ldots\}$ such that $x^k \xrightarrow{K} x^\infty$ and $v^k \xrightarrow{K} 0$. Then $0 \in \partial h(x^\infty) + \partial \delta_S(x^\infty)$.*

Proof. Since $-v^k = u^k |d^k|^2 + \tilde{\alpha}^k \xrightarrow{K} 0$ (cf. (2.10)), $u^k \in [u_{\min}, u_{\max}]$ (cf. Step 5) and $\tilde{\alpha}^k \geq p(x^k)\tilde{\alpha}_f^k \geq 0$ (cf. (2.8)) for all k, we have $d^k \xrightarrow{K} 0$, $p(x^k)\tilde{\alpha}_f^k \xrightarrow{K} 0$. Hence if $p(x^\infty) > 0$ then (cf. continuity of p) $\tilde{\alpha}_f^k \xrightarrow{K} 0$, so we may use $\tilde{g}_f^k \in \partial_{\tilde{\alpha}_f^k} f(x^k)$ (cf. (2.9)) and local boundedness and upper semicontinuity of $\partial_\cdot f(\cdot)$ to deduce the existence of $\tilde{g}_f^\infty \in \partial f(x^\infty)$ and an infinite set $K' \subset K$ such that $\tilde{g}_f^k \xrightarrow{K'} \tilde{g}_f^\infty$. Then the limit of $-u^k d^k - p(x^k)\tilde{g}_f^k - f(x^k)\nabla p(x^k) = \tilde{g}_S^k \in \partial \delta_S(x^k + d^k)$ (cf. (2.7)) as $k \to \infty$, $k \in K'$, yields $-p(x^\infty)\tilde{g}_f^\infty - f(x^\infty)\nabla p(x^\infty) \in \partial \delta_S(x^\infty)$ by continuity and closedness of S, so $0 \in \partial h(x^\infty) + \partial \delta_S(x^\infty)$. Next, if $p(x^\infty) = 0$, for each k let $\tilde{x}^k = x^k + \tilde{g}_f^k / |\tilde{g}_f^k|$ if $\tilde{g}_f^k \neq 0$; otherwise pick any $\tilde{x}^k$ with $|\tilde{x}^k - x^k| = 1$. Multiplying the subgradient inequality $f(\tilde{x}^k) - f(x^k) + \tilde{\alpha}_f^k \geq \left\langle \tilde{g}_f^k, \tilde{x}^k - x^k \right\rangle$ (cf. (2.9)) by $p(x^k) \geq 0$, we get $|p(x^k)\tilde{g}_f^k| \leq p(x^k)[f(\tilde{x}^k) - f(x^k)] + p(x^k)\tilde{\alpha}_f^k \xrightarrow{K} 0$, since p and f are continuous, $x^k \xrightarrow{K} x^\infty$, $p(x^k)\tilde{\alpha}_f^k \xrightarrow{K} 0$ and $|\tilde{x}^k - x^k| = 1$ for all k. Thus $p(x^k)\tilde{g}_f^k \xrightarrow{K} 0 = p(x^\infty)\tilde{g}_f^\infty$ for any $\tilde{g}_f^\infty \in \partial f(x^\infty)$, and the preceding argument yields $0 \in \partial h(x^\infty) + \partial \delta_S(x^\infty)$. □

Note that, by construction (cf. Step 3),

$$x^k = x^{k(l)} \quad \text{if } k(l) \le k < k(l+1), \tag{3.1}$$

where we set $k(l+1) = \infty$ if the number l of descent steps stays fixed.

Lemma 3.2. (i) *Let* $w^k = u^k|d^k|^2/2 + \tilde{\alpha}_p^k$. *Then* $v^k \le -w^k \le v^k/2$.
(ii) *If* $x^{k+1} = x^k$ *then* $0 \le w^{k+1} \le w^k - u^k|d^{k+1} - d^k|^2/2$.
(iii) *If* $k = k(l)$ *then* $w^k \le |p(x^k)g_f(x^k) + f(x^k)\nabla p(x^k)|^2/2u^k$ *with* $u^k \ge u_{\min}$.
(iv) $|d^k| \le |p(x^{k(l)})g_f(x^{k(l)}) + f(x^{k(l)})\nabla p(x^{k(l)})|/(u^k u_{\min})^{1/2}$.

Proof. (i) This follows from (2.10) and (2.8).

(ii) Let $\hat{h}^k(\cdot) = p(x^k)\hat{f}^k(\cdot) + f(x^k)\left\langle \nabla p(x^k), \cdot - x^k \right\rangle$, $\check{\phi}^k(\cdot) = \check{h}^k(x^k + \cdot) + u^k|\cdot|^2/2 + \delta_S(x^k + \cdot)$, $\hat{\phi}^k(\cdot) = \hat{h}^k(x^k + \cdot) + u^k|\cdot|^2/2 + \delta_S(x^k + \cdot)$ and (cf. (2.4))

$$\eta^k = \min \check{\phi}^k = \check{h}^k(x^k + d^k) + u^k|d^k|^2/2. \tag{3.2}$$

By the choice (2.11) of $\hat{f}^k = \max_{j \in \hat{J}^k} f^j$, $d^k = \arg\min \hat{\phi}^k$ and $\hat{h}^k(x^k + d^k) = \check{h}^k(x^k + d^k)$, so $\eta^k = \min \hat{\phi}^k$ and the strong convexity of $\hat{\phi}^k$ implies (cf. [Roc76])

$$\hat{\phi}^k(d) \ge \eta^k + u^k|d - d^k|^2/2 \quad \forall d \in \mathbb{R}^n. \tag{3.3}$$

If $x^{k+1} = x^k$, then $\check{f}^{k+1} \ge \hat{f}^k$ (cf. $J^{k+1} \supset \hat{J}^k$) and $u^{k+1} \ge u^k$ (cf. Step 5), so $\check{\phi}^{k+1} \ge \hat{\phi}^k$ and

$$\eta^{k+1} \ge \eta^k + u^k|d^{k+1} - d^k|^2/2 \tag{3.4}$$

from (3.2)–(3.3) and $p(x^k) \ge 0$. But $w^k = h(x^k) - \eta^k$, since $\eta^k = \check{h}^k(x^k + d^k) + u^k|d^k|^2/2$ (cf. (3.2)), $\check{h}^k(x^k + d^k) = h(x^k) + v^k$ (cf. (2.5)) and $-w^k = v^k + u^k|d^k|^2/2$ (cf. (2.10)), so $w^{k+1} \le w^k - u^k|d^{k+1} - d^k|^2/2$ from (3.4) and $h(x^{k+1}) = h(x^k)$ (cf. Step 3).

(iii) If $k = k(l)$ then, since $y^k = x^k$ (cf. Step 3), $k \in J^k$ (cf. Step 4) and $\check{f}^k(\cdot) \ge f^k(\cdot) = f(x^k) + \left\langle g_f^k, \cdot - x^k \right\rangle$ (cf. (2.3a)), (3.2) yields

$$\eta^k \ge \min_d \{p(x^k)f(x^k) + \left\langle p(x^k)g_f^k + f(x^k)\nabla p(x^k), d \right\rangle + u^k|d|^2/2\},$$

so $w^k = h(x^k) - \eta^k \le |p(x^k)g_f(x^k) + f(x^k)\nabla p(x^k)|^2/2u^k$, where $u^k \ge u_{\min}$ (cf. Step 5).

(iv) Using $|d^k| \le (2w^k/u^k)^{1/2}$ (cf. part (i) and (2.8)), apply parts (ii)–(iii). □

Lemma 3.3. *If* $B \subset S$ *is bounded then there exists* $c < \infty$ *such that if* $x^k \in B$ *then* $|d^k| \le c/(u^k)^{1/2} \le c/(u_{\min})^{1/2}$ *and* $|g_f^{k+1}| \le c$.

Proof. Use Lemma 3.2(iv), (3.1), the facts $u^k \ge u_{\min}$ (cf. Step 5), $y^{k+1} = x^k + t_R^k d^k$ with $t_R \le 1$ (cf. Step 3) for all k, and local boundedness of f, p, g_f and ∇p. □

We may now consider the case of a finite number of descent steps.

Lemma 3.4. *If* $x^k = x^{k(l)} = x^\infty$ *for some fixed* l *and all* $k \ge k(l)$, *then* $v^k \to 0$.

Proof. By the algorithm's rules and Lemma 3.2(ii), $u^{k+1} \ge u^k$ and $w^{k+1} \le w^k$ for all large k, and $|d^{k+1} - d^k| \to 0$. Let $\bar{v} = \limsup_{k \to \infty} v^k$ and $K \subset \{1, 2, \ldots\}$ satisfy $v^k \xrightarrow{K} \bar{v}$.

Let $k \geq k(l)$ and $\epsilon^k = \left\langle p(x^k)g_f^{k+1} + f(x^k)\nabla p(x^k), d^k \right\rangle - p(x^k)\alpha_{k+1}^{k+1} - v^k$. Then, by (2.6) with $x^{k+1} = x^k$, $k+1 \in J^{k+1}$ and $v = v^{k+1}$,

$$\begin{aligned}\epsilon^k &= \left\langle p(x^k)g_f^{k+1} + f(x^k)\nabla p(x^k), d^{k+1} \right\rangle - p(x^k)\alpha_{k+1}^{k+1} - v^k \\ &\quad - \left\langle p(x^k)g_f^{k+1} + f(x^k)\nabla p(x^k), d^{k+1} - d^k \right\rangle \\ &\leq v^{k+1} - v^k + [p(x^k)|g_f^{k+1}| + |f(x^k)\nabla p(x^k)|]|d^{k+1} - d^k|,\end{aligned}$$

so $\limsup_{k\in K} \epsilon^k \leq 0$ by Lemma 3.3. But (2.14) holds for all large k, so $\epsilon^k \geq \kappa_R v^k - v^k = (1-\kappa_R)|v^k|$ with $\kappa_R \in (0,1)$ imply $\bar{v} = 0$. Then $w^k \downarrow 0$ and $v^k \to 0$ by Lemma 3.2(i,ii). □

It remains to analyze the case of an infinite number of descent steps.

Lemma 3.5. *Suppose there exist $x^\infty \in S$ and an infinite set $L \subset \{1,2,\ldots\}$ such that $x^{k(l)} \xrightarrow{L} x^\infty$. Then $v^k \xrightarrow{K} 0$, where $K = \{k(l+1) - 1 : l \in L\}$.*

Proof. Suppose $v^k \leq \bar{v}$ for some $\bar{v} < 0$ and all large $k \in K$. Since $x^k \xrightarrow{K} x^\infty$ and $h(x^{k+1}) \leq h(x^k) + \kappa_L t_L^k v^k \leq h(x^k)$ (cf. (2.12)) for all k, $h(x^k) \downarrow h(x^\infty)$ by continuity of h and $t_L^k v^k \to 0$. Then $t_L^k \xrightarrow{K} 0$ and $|x^{k+1} - x^k| \leq t_L^k |d^k| \xrightarrow{K} 0$, since $\{d^k\}_{k\in K}$ is bounded (cf. Lemma 3.3). Thus both $\{x^k\}_K$ and $\{x^{k+1}\}_K$ converge to x^∞, so the right side of (2.13) vanishes as $k \to \infty$, $k \in K$, due to the continuity of f, p and ∇p, the boundedness of $\{d^k\}_K$ and $\{g_f^k\}_K$ (cf. Lemma 3.3), and properties of α_f (cf. [Mif82]). But the left side of (2.13) is at least $\kappa_v|\bar{v}| > 0$ for large $k \in K$, a contradiction. Therefore, $v^k \xrightarrow{K} 0$. □

Combining (3.1) with Lemmas 3.1 and 3.4–3.5, we deduce our main result.

Theorem 3.6. *Every accumulation point of $\{x^k\}$ is stationary for h on S.* □

Remark 3.7. If the set $\{x \in S : h(x) \leq h(x^1)\}$ is bounded and $\epsilon_{\text{opt}} > 0$, then the algorithm will terminate in a finite number of iterations, producing an approximately stationary point x^k with $-v^k \leq \epsilon_{\text{opt}}$. This follows from the proofs of Lemmas 3.4–3.5.

Remark 3.8. Theorem 3.6 still holds if, to save storage, one employs aggregation as in [Kiw85, Kiw86, Kiw90]. Briefly, *subgradient aggregation* boils down to replacing an arbitrary linearization f^j by the aggregate linearization $\tilde{f}^k$ (cf. the derivation of (2.9)) and selecting J^{k+1} so that $\{j, k+1\} \subset J^{k+1}$, e.g., $J^{k+1} = \{j, k+1\}$.

Remark 3.9. The preceding convergence results remain valid if we only assume that p is nonnegative and continuous on S, ∇p is continuous on S, f is continuous on S and $\partial f(\cdot)$ is locally bounded on S. The last assumption may be replaced by the requirement that g_f be bounded on S (then g_f^k are bounded, and so are their aggregates $\tilde{g}_f^k$, as required in the proof of Lemma 3.1). In particular, g_f is bounded if f is polyhedral and finite-valued on S. Such relaxed assumptions carry over to the extension presented in §4.

4 The method for the general case of $m > 1$

Algorithm 2.1 extends easily to the case of $h = \sum_{i=1}^m p_i f_i$ with $m > 1$. Then the linearizations $\bar{f}_i(x;y) = f_i(y) + \langle g_{f_i}(y), x - y\rangle$ and errors $\alpha_{f_i}(x,y) = f_i(x) - \bar{f}_i(x;y)$ of f_i (cf. (2.1),

(2.2)) are employed in the models

$$\check{f}_i^k(x) = \max\{f_i^j(x^k) + \left\langle g_{f_i}^j, x - x^k \right\rangle : j \in J_i^k\}, \tag{4.1a}$$

$$\check{h}_i^k(x) = p_i(x^k)\check{f}_i^k(x) + f_i(x^k)\left\langle \nabla p_i(x^k), x - x^k \right\rangle, \tag{4.1b}$$

$$\check{h}^k(x) = \sum_{i=1}^m \check{h}_i^k(x), \tag{4.1c}$$

with $f_i^j(\cdot) = \bar{f}_i(\cdot; y^j)$, $g_{f_i}^j = g_{f_i}(y^j)$, $j \in J_i^k \subset \{1\!:\!k\}$, $i = 1\!:\!m$. Accordingly, d^k and $v^k = \sum_{i=1}^m v_i^k$ can be computed by finding $(d^k, v_1^k, \ldots, v_m^k)$ to

$$\begin{array}{ll} \text{minimize} & u^k|d|^2/2 + \sum_{i=1}^m v_i \quad \text{over all } (d, v_1, \ldots, v_m) \in \mathbb{R}^{n+m} \\ \text{satisfying} & -p_i(x^k)\alpha_{ij}^k + \left\langle p_i(x^k) g_{f_i}^j + f_i(x^k)\nabla p_i(x^k), d \right\rangle \le v_i, \quad j \in J_i^k,\ i = 1\!:\!m, \\ & x^k + d \in S, \end{array} \tag{4.2}$$

where $\alpha_{ij}^k = f_i(x^k) - f_i^j(x^k)$. The Lagrange multipliers λ_{ij}^k of (4.2) may be used for selecting $\hat{J}_i^k = \{j \in J_i^k : \lambda_{ij}^k \neq 0\}$ such that $\sum_{i=1}^m |\hat{J}_i^k| \le M - m$, where $M \ge n + 2m$ (cf. [Kiw89, Kiw94]). Thus, for dense $g_{f_i}^j$ and $\nabla p_i(x^k)$, the algorithm requires storage of order $n(M+m) \ge n(n+3m)$ (plus the QP workspace, which can be of order $\min\{m,n\}^2/2$; cf. [Kiw94]). The storage requirements can be reduced to about $3mn$ locations via subgradient aggregation (cf. [Kiw90]), at the cost of slower convergence. One easily extends the argument that provided relations (2.7)–(2.9), which become

$$\tilde{g}^k = \sum_{i=1}^m [p_i(x^k)\tilde{g}_{f_i}^k + f_i(x^k)\nabla p_i(x^k)] + \tilde{g}_S^k = -u^k d^k,$$

$$\tilde{\alpha}^k = \sum_{i=1}^m p_i(x^k)\tilde{\alpha}_{f_i}^k + \tilde{\alpha}_S^k \ge 0,$$

$$\tilde{g}_{f_i}^k \in \partial_{\tilde{\alpha}_{f_i}^k} f_(x^k), \quad i = 1\!:\!m.$$

Of course, the line search criteria (2.13)–(2.14) are replaced by

$$\kappa_v|v^k| \le \sum_{i=1}^m \Big\{ p_i(x^k)\alpha_{f_i}(x^k, x^{k+1}) + [p_i(x^{k+1}) - p_i(x^k)]\left\langle g_{f_i}(x^{k+1}), d^k \right\rangle + [f_i(x^{k+1})\left\langle \nabla p_i(x^{k+1}), d^k \right\rangle - f_i(x^k)\left\langle \nabla p_i(x^k), d^k \right\rangle] \Big\},$$

$$\sum_{i=1}^m \left\{ -p_i(x^{k+1})\alpha_{f_i}(x^{k+1}, y^{k+1}) + \left\langle p_(x^{k+1}) g_{f_i}(y^{k+1}) + f_i(x^{k+1})\nabla p_i(x^{k+1}), d^k \right\rangle \right\} \ge \kappa_R v^k,$$

and corresponding changes occur in Procedure 2.2.

It is easy to verify all the convergence results of §3 for this extension of Algorithm 2.1.

References

[Cla83] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[ErW88] Yu. Ermoliev and R. J.-B. Wets, eds., *Numerical Techniques for Stochastic Optimization*, Springer-Verlag, Berlin, 1988.

[HUL93] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.

[Kiw85] K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Mathematics 1133, Springer-Verlag, Berlin, 1985.

[Kiw86] ———, *A method for minimizing the sum of a convex function and a continuously differentiable function*, J. Optim. Theory Appl. **48** (1986) 437–449.

[Kiw89] ———, *A dual method for certain positive semidefinite quadratic programming problems*, SIAM J. Sci. Statist. Comput. **10** (1989) 175–186.

[Kiw90] ———, *Proximity control in bundle methods for convex nondifferentiable minimization*, Math. Programming **46** (1990) 105–122.

[Kiw92] ———, *A restricted step proximal bundle method for nonconvex nondifferentiable optimization*, in Nonsmooth Optimization, Methods and Applications, F. Giannessi, ed., Gordon and Breach, Philadelphia, 1992, pp. 175–188.

[Kiw94] ———, *A Cholesky dual method for proximal piecewise linear programming*, Numer. Math. **?** (1994). To appear.

[Mif77a] R. Mifflin, *An algorithm for constrained optimization with semismooth functions*, Math. Oper. Res. **2** (1977) 191–207.

[Mif77b] ———, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim. **15** (1977) 959–972.

[Mif82] ———, *A modification and an extension of Lemaréchal's algorithm for nonsmooth minimization*, Math. Programming Stud. **17** (1982) 77–90.

[Roc76] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim. **14** (1976) 877–898.

[RoW91] R. T. Rockafellar and R. J.-B. Wets, *Scenarios and policy aggregation in optimization under uncertainty*, Math. Oper. Res. **16** (1991) 119–147.

[Rus86] A. Ruszczyński, *A regularized decomposition method for minimizing a sum of polyhedral functions*, Math. Programming **35** (1986) 309–333.

[Rus93a] ———, *Private communication*, Nov. 1993. IIASA, Laxenburg, Austria.

[Rus93b] ———, *Regularized decomposition of stochastic programs: Algorithmic techniques and numerical results*, WP-93-21, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1993.

[ScZ92] H. Schramm and J. Zowe, *A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results*, SIAM J. Optim. **2** (1992) 121–152.