A DESCENT ALGORITHM FOR LARGE-SCALE LINEARLY CONSTRAINED CONVEX NONSMOOTH MINIMIZATION

Krzysztof C. Kiwiel

April 1984
CP-84-15

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

A Descent Algorithm for Large-Scale Linearly Constrained Convex Nonsmooth Minimization

Krzysztof C. Kiwiel
Systems Research Institute, Polish Academy of Sciences,
Newelska 6, 01-447 Warsaw, Poland.

Abstract. A descent algorithm is given for solving a large convex program obtained by augmenting the objective of a linear program with a (possibly nondifferentiable) convex function depending on relatively few variables. Such problems often arise in practice as deterministic equivalents of stochastic programming problems. The algorithm's search direction finding subproblems can be solved efficiently by the existing software for large-scale smooth optimization. The algorithm is both readily implementable and globally convergent.

AMS 1980 subject classification. Primary: 65K05. Secondary: 90C25.

Key words: Nonsmooth optimization, nondifferentiable programming, linear constraints, convex programming, descent methods, large-scale optimization

1. Introduction

This paper presents a method for solving the following problem

$$\begin{aligned} &\text{minimize } \langle c,y\rangle + f(x) \quad \text{over all } (y,x) \in R^M \times R^N \\ &\text{satisfying } Ay + Bx \le b, \end{aligned} \tag{1.1}$$

where $c \in R^M$, A is an $P \times M$-matrix, B is an $P \times N$-matrix, $b \in R^P$, and $f: R^N \longrightarrow R^1$ is a (possibly nondifferentiable) convex function. We suppose that the set of feasible points

$$S = \{(y,x) \in R^{M+N} : Ay + Bx \le b\}$$

is nonempty and bounded, and that at each $(y,x) \in S$ we can compute $f(x)$ and a certain subgradient $g_f(x) \in \partial f(x)$, i.e. an arbitrary element of the subdifferential $\partial f(x)$ of f at x on which we cannot impose any further restrictions.

Problems of the form (1.1) are often encountered in practice, especially as deterministic equivalents of two-stage stochastic programming problems [K1],[NW1],[W1]. In many applications the number M of "linear" variables y_i is much larger than the number N of "nonlinear" variables x_i, and the matrices A and B are sparse (have relatively few nonzero entries). In such cases problem (1.1) can be solved by the existing algorithms for large-scale optimization (e.g. MINOS [MS1]) if f is differentiable. In the nondifferentiable large-scale case, only a few algorithms have been proposed [NW1], and they frequently assume the knowledge of the full subdifferential $\partial f(x)$ at each x.

The method presented in this paper modifies one given in [K3] to make use of the special structure of problem (1.1). It is a feasible point method of descent in the sense of generating successive points in S with nonincreasing objective values.

To deal with nondifferentiability of f, at each iteration a piecewise linear (polyhedral) approximation to f is constructed from at most N+2 subgradients of f calculated previously at certain trial points. A search direction is found by solving a quadratic programming subproblem obtained by replacing f in (1.1) by its polyhedral approximation augmented with a simple quadratic term. Then a line search finds the next approximation to a solution and the next trial point. The two-point line search is employed to detect discontinuities in the gradient of f.

We show that the method is globally convergent under no additional assumptions. We may add that the method will find a solution in a finite number of iterations if f is polyhedral and certain technical conditions are satisfied (see [K2]). From lack of space, we shall pursue this subject elsewhere.

The method is implementable in the sense of requiring bounded storage and a finite number of simple operations per iteration. For problems with large sparse matrices A and B and relatively few nonlinear variables x_i, the method can use MINOS [MS1] for solving its quadratic programming subproblems. In fact, an efficient implementation of the method would require modifying MINOS to exploit the fact that consecutive subproblems retain the original constraints of (1.1), differ only in a few auxiliary linear constraints on x, have simple terms quadratic in x as the only nonlinearities in their objectives, etc. It would be interesting to perform the necessary numerical experimentation, but we have not had the means to do so.

Other descent methods for solving problem (1.1) can be found in [DV1], [K4], [LSB1], [M1], [M2], [PMW1] and [SNH1]. None of their search direction finding subproblems can be solved efficiently by the available software when problem (1.1) is large. Therefore, we hope that our method could compete with the existing algorithms.

The method is derived and stated in Section 2. Its global convergence is established in Section 3, where we also discuss the case of an unbounded feasible set S. Finally, we have a conclusion section.

We shall use the following notation and terminology. R^M and R^N denote the M- and N-dimensional Euclidean spaces with the usual inner products $\langle\cdot,\cdot\rangle$ and the associated norms $|\cdot|$, respectively. We use x_i to denote the i-th component of the vector x. Superscripts are used to denote different vectors, e.g. x^1 and x^2. All vectors are column vectors. However, for convenience a column vector in R^{M+N} is sometimes denoted by (y,x) even though y and x are column vectors in R^M and R^N, respectively. For any $x \in R^N$ and $\varepsilon \geq 0$,

$$\partial_\varepsilon f(x) = \{ g \in R^N : f(\tilde{x}) \geq f(x) + \langle g, \tilde{x}-x \rangle - \varepsilon \text{ for all } x \in R^N \}$$

denotes the ε-subdifferential of f at x. We denote by $\partial f(x)$ the set $\partial_0 f(x)$, i.e. the ordinary subdifferential. Note that f is continuous and the mapping $(x,\varepsilon) \longmapsto \partial_\varepsilon f(x)$ is locally bounded, because f is real-valued and convex on R^N (see, e.g. [DV1]).

2. The Method

Given a starting point $z^1 = (y^1,x^1) \in S$, the algorithm described below generates sequences of points $z^k = (y^k,x^k)$ in S, search directions $d^k = (d_y^k,d_x^k)$ in $R^M \times R^N$ and stepsizes t_L^k in $\{0,1\}$, related by $z^{k+1} = z^k + t_L^k d^k$ for k=1, 2,... . The sequence $\{z^k\}$ is intended to converge to a solution of problem

(1.1). The algorithm is a method of descent in the sense that $F(z^{k+1}) \le F(z^k)$ if $z^{k+1} \ne z^k$, where

$$F(z) = \langle c, y\rangle + f(x) \quad \text{for all} \quad z=(x,y) \in S$$

is the objective function of problem (1.1). Also a sequence of trial points $\tilde{z}^k = (\tilde{y}^k, \tilde{x}^k) \in S$ is generated by the formula $\tilde{z}^{k+1} = z^k + d^k$ for $k=1, 2, \ldots$, $\tilde{z}^1 = z^1$.

The algorithm calculates subgradients $g^j = g_f(\tilde{x}^j) \in f(\tilde{x}^j)$ of f at the trial points $\tilde{z}^j = (\tilde{y}^j, \tilde{x}^j)$. With each such subgradient we associate the following linearization of f at $\tilde{x}^j$

$$f_j(x) = f(\tilde{x}^j) + \langle g^j, x-\tilde{x}^j\rangle \quad \text{for all} \quad x \in R^N,$$

which can be expressed at iteration $k \ge j$ as

$$f_j(x) = f_j^k + \langle g^j, x-x^k\rangle$$

with $f_j^k = f_j(x^k)$, for all $x \in R^N$ and $j=1,\ldots,k$. At the k-th iteration, the algorithm uses the following polyhedral approximation to f

$$\hat{f}^k(x) = \max\{f_j(x) : j \in J^k\} \quad \text{for all} \quad x \in R^N,$$

where the set $J^k \subset \{1,\ldots,k\}$ has at most N+2 elements ($|J^k| \le N+2$). By convexity, $f(x) \ge f_j(x)$ for all $x \in R^N$, so $\hat{f}^k$ is a lower polyhedral approximation to f:

$$f(x) \ge \hat{f}^k(x) \quad \text{for all } x \in R^N,$$

$$f(\tilde{x}^j) = \hat{f}^k(\tilde{x}^j) \text{ for all } j \in J^k,$$

and the function

$$\hat{F}^k(z) = \langle c, y\rangle + \hat{f}^k(x) \quad \text{for all } z=(y,x) \in R^{M+N}$$

is a lower polyhedral approximation to F.

Since we want to find a feasible direction of descent for F at $z^k = (y^k, x^k)$, we shall find $d^k = (d_y^k, d_x^k)$ to

$$\text{minimize } \hat{F}^k(z^k+d) + \tfrac{1}{2}|d_x|^2 \text{ over all } d=(d_y,d_x)$$
$$\text{satisfying } z^k+d \in S, \tag{2.1}$$

where the penalty term $|d_x|^2/2$ serves to keep $\tilde{x}^{k+1} = x^k + d_x^k$ in the region where $\hat{f}^k$ is a close approximation to f, so that $\hat{F}^k(\cdot)$ is close to $F(\cdot)$ at $\tilde{z}^{k+1} = z^k + d^k$. Clearly, d^k may be found from the solution $(d_y^k, d_x^k, u^k) \in R^{M+N+1}$ to the following k-th quadratic programming subproblem

$$\begin{aligned} &\text{minimize } \langle c,d_y\rangle + u + \tfrac{1}{2}|d_x|^2 \text{ over all } (d_y,d_x,u) \in R^{M+N+1} \\ &\text{satisfying } f_j^k + \langle g^j,d_x\rangle \le u \text{ for } j \in J^k, \\ &\qquad A(y^k + d_y) + B(x^k + d_x) \le b. \end{aligned} \tag{2.2}$$

Moreover,

$$u^k = \hat{f}^k(x^k + d_x^k),$$

so we may interpret

$$\begin{aligned} v^k &= \hat{F}^k(z^k + d^k) - F(z^k) \\ &= \langle c,d_y^k\rangle + u^k - f(x^k) \end{aligned} \tag{2.3}$$

as an approximate derivative of F at z^k in the direction d^k.

It will be convenient to describe the P linear constraints of problem (1.1) in terms of P affine functions $h_i: R^M \cdot R^N \longrightarrow R^1$ such that

$$S = \{(y,x) \in R^{M+N} : h_i(y,x) \le 0 \text{ for } i \in I\},$$

where $I = \{1,\dots,P\}$. Then subproblem (2.2) takes on the form

$$\begin{aligned} &\text{minimize } \langle c,d_y\rangle + u + \tfrac{1}{2}|d_x|^2 \text{ over all } (d_y,d_x,u) \in R^{M+N+1} \\ &\text{satisfying } f_j^k + \langle g^j, d_x\rangle \le u \text{ for } j \in J^k, \\ &\qquad h_i^k + \langle \nabla_y h_i, d_y\rangle + \langle \nabla_x h_i, d_x\rangle \le 0 \text{ for } i \in I \end{aligned} \tag{2.4}$$

with $h_i^k = h_i(y^k,x^k)$ for $i \in I$, since

$$h_i(y^k+d_y, x^k+d_x) = h_i(y^k,x^k) + \langle \nabla_y h_i, d_y\rangle + \langle \nabla_x h_i, d_x\rangle$$

for all (d_y,d_x), because each h_i is affine.

Having motivated the search direction finding subproblems, we shall now state the method in detail, commenting on its rules in what follows.

Algorithm 2.1.

Step 0 (Initialization). Select a starting point $z^1 = (y^1, x^1) \in S$, a final accuracy tolerance $\varepsilon_s \geq 0$ and a line search parameter $m \in (0,1)$. Set $J^1 = \{1\}$, $\tilde{z}^1 = (\tilde{y}^1, \tilde{x}^1) = z^1$, $g^1 = g_f(\tilde{x}^1)$ and $f_1^1 = f(\tilde{x}^1)$. Set the counters k=1, l=0 and k(0) = 1.

Step 1 (Direction finding). Find the solution (d_y^k, d_x^k, u^k) to subproblem (2.4), and Lagrange multipliers λ_j^k, $j \in J^k$, and μ_i^k, $i \in I$, of (2.4) such that the set

$$\hat{J}^k = \{j \in J^k : \lambda_j^k \neq 0\}$$

satisfies $|\hat{J}^k| \leq N+1$. Set $d^k = (d_y^k, d_x^k)$ and compute v^k by (2.3).

Step 2 (Stopping criterion). If $v^k \geq -\varepsilon_s$, terminate; otherwise, continue.

Step 3 (Line search). Set $\tilde{z}^{k+1} = (\tilde{y}^{k+1}, \tilde{x}^{k+1}) = z^k + d^k$. If

$$F(\tilde{z}^{k+1}) \leq F(z^k) + mv^k, \tag{2.5}$$

set $t_L^k = 1$ (serious step), set $k(l+1) = k+1$ and increase l by 1; otherwise, i.e. if (2.5) does not hold, set $t_L^k = 0$ (null step). Set $z^{k+1} = (y^{k+1}, x^{k+1}) = z^k + t_L^k d^k$.

Step 4 (Linearization updating). Set $J^{k+1} = \hat{J}^k \cup \{k+1\}$. Set $g^{k+1} = g_f(\tilde{x}^{k+1})$,

$$f_{k+1}^{k+1} = f(\tilde{x}^{k+1}) + \langle g^{k+1}, x^{k+1} - \tilde{x}^{k+1} \rangle, \tag{2.6}$$

$$f_j^{k+1} = f_j^k + \langle g^j, x^{k+1} - x^k \rangle \quad \text{for } j \in \hat{J}^k. \tag{2.7}$$

Increase k by 1 and go to Step 1.

A few remarks on the algorithm are in order.

For problems of interest to us, subproblems (2.4) will have relatively few nonlinear variables ($N \ll M$) and large, but sparse, constraint matrices. Such subproblems can be solved by MINOS [M1] in a finite number of iterations; moreover, MINOS will automatically produce at most N+1 nonzero Lagrange multipliers λ_j^k for the first constraints of (2.4), since these constraints involve only N+1 variables.

In Step 2 we always have

$$F(z) \geq F(z^k) + v^k - |v^k|^{1/2} \, |x-x^k| \quad \text{for all} \quad z=(y,x) \in S, \tag{2.8}$$

and hence

$$F(z^k) \leq \min\{F(z) : z \in S\} - v^k +$$
$$+ |v^k|^{1/2} \max\{ |x-x^k| : z = (y,x) \in S, \; F(z) \leq F(z^k)\}.$$

This will be proved in the next section. The above estimates justify the stopping criterion of the method.

Step 3 is always entered with $v^k < 0$. The trial point $\tilde{z}^{k+1}$ is accepted as the next iterate z^{k+1} only if this decreases significantly the objective value. Otherwise the algorithm stays at $z^{k+1} = z^k$ (a null step), but the new subgradient information collected at $\tilde{z}^{k+1}$ will aid in finding a better next search direction, since $k+1 \in J^{k+1}$. Of course, $\{z^k\} \subset S$, because $\tilde{z}^{k+1} = z^k + d^k \in S$ for all k.

We may add that if there are no linear variables in problem (1.1) (M=0), then Algorithm 2.1 becomes similar to the method of [K3].

3. Convergence

In this section we show that the algorithm generates a minimizing sequence $\{z^k\} \subset S$, i.e. $F(z^k) \downarrow \min\{F(z) : z \in S\}$; moreover, there exists $\bar{z} = (\bar{y},\bar{x})$ in the set of solutions of problem (1.1)

$\bar{Z} = \text{Arg min}\{F(z) : z \in S\}$
such that $x^k \longrightarrow \bar{x}$ and $y^k \xrightarrow{K} \bar{y}$ for some infinite set $K \subset \{1,2,\ldots\}$.
We assume, of course, that the final accuracy tolerance ε_s is set to zero. Our analysis will dwelve on the results in [K2], [K3].

We start by analyzing the following dual to the k-th subproblem (2.4)

$$\underset{\lambda,\mu}{\text{minimize}} \ \frac{1}{2}\Big|\sum_{j\in J^k}\lambda_j g^j + \sum_{i\in I}\mu_i \nabla_x h_i\Big|^2 + \sum_{j\in J^k}\lambda_j \alpha_j^k - \sum_{i\in I}\mu_i h_i^k,$$

$$\text{subject to} \quad \lambda_j \geq 0 \ \text{ for } j\in J^k, \ \sum_{j\in J^k}\lambda_j = 1, \tag{3.1}$$

$$\mu_i \geq 0 \ \text{ for } i\in I,$$

$$c + \sum_{i\in I}\mu_i \nabla_y h_i = 0,$$

where

$$\alpha_j^k = f(x^k) - f_j^k \quad \text{for } j\in J^k. \tag{3.2}$$

<u>Lemma 3.1</u>. (i) The Lagrange multipliers (λ^k, μ^k) of (2.4) solve (3.1) and yield the unique part (d_x^k, u^k) of the solution (d_y^k, d_x^k, u^k) of (2.4) by

$$d_x^k = -p_x^k, \tag{3.3}$$

$$u^k = \tilde{f}_p^k + \langle p_f^k, d_x^k\rangle, \tag{3.4}$$

where

$$p_x^k = p_f^k + \sum_{i\in I}\mu_i^k \nabla_x h_i, \tag{3.5}$$

$$(p_f^k, \tilde{f}_p^k) = \sum_{j\in J^k}\lambda_j^k (g^j, f_j^k). \tag{3.6}$$

(ii) The optimal value w^k of (3.1) satisfies

$$w^k = \frac{1}{2}|p_x^k|^2 + \tilde{\alpha}^k, \tag{3.7}$$

and one has

$$v^k = -\{|p_x^k|^2 + \tilde{\alpha}^k\}, \tag{3.8}$$

where

$$\tilde{\alpha}^k = \tilde{\alpha}_p^k + \alpha_h^k , \tag{3.9}$$

$$\tilde{\alpha}_p^k = f(x^k) - \tilde{f}_p^k , \tag{3.10}$$

$$\alpha_h^k = - \sum_{i \in I} \mu_i^k h_i^k . \tag{3.11}$$

Proof. (i) Observe that the feasible set of subproblem (2.1) is nonempty and bounded, since so is S by assumption, and that its objective is convex in d and strongly convex in d_x. Hence the first assertion can be deduced from convex duality theory as in [W2],[K2],[K3].

(ii) (3.7) follows immediately from the preceding formulae and the fact that λ^k is feasible for (3.1). Next, since μ^k is feasible in (3.1) ,

$$c + \sum_{i \in I} \mu_i^k \nabla_y h_i = 0,$$

while the Kuhn-Tucker conditions for (2.4) yield

$$\sum_{i \in I} \mu_i^k [h_i^k + \langle \nabla_y h_i, d_y^k \rangle + \langle \nabla_x h_i, d_x^k \rangle] = 0,$$

so

$$\langle c, d_y^k \rangle = \langle \sum_{i \in I} \mu_i^k \nabla_x h_i, d_x^k \rangle + \sum_{i \in I} \mu_i^k h_i^k . \tag{3.12}$$

Therefore, by (2.3) and (3.3) - (3.6),

$$\begin{aligned} v^k &= \langle c, d_y^k \rangle + u^k - f(x^k) = \\ &= \langle \sum_{i \in I} \mu_i^k \nabla_x h_i + p_f^k, d_x^k \rangle + \sum_{i \in I} \mu_i^k h_i^k + \tilde{f}_p^k - f(x^k) = \\ &= \langle p_x^k, - p_x^k \rangle - \alpha_h^k - \tilde{\alpha}_p^k \end{aligned}$$

and (3.8) follows, completing the proof.

We may now verify relation (2.8) .

Lemma 3.2. If Algorithm 2.1 did not stop before the k-th iteration, then

$$F(z) \geq F(z^k) + \langle p_x^k, x-x^k\rangle - \tilde{\alpha}^k \quad \text{for all} \quad z=(y,x)\in S. \tag{3.13}$$

Moreover, $\tilde{\alpha}_p^k \geq 0$, $\alpha_h^k \geq 0$, $\tilde{\alpha}^k \geq 0$, $v^k \leq 0$ and relation (2.8) holds.

Proof. As in [K2] and [K3], (3.6) , (3.10) and the fact that λ_j^k, $j \in J^k$, form a convex combination yield

$$f(x) \geq f(x^k) + \langle p_f^k, x-x^k\rangle - \tilde{\alpha}_p^k \quad \text{for all } x \in R^N. \tag{3.14}$$

Let (y,x) be any point in S. Then

$$0 \geq h_i(y,x) = h_i^k + \langle \nabla_y h_i, y-y^k\rangle + \langle \nabla_x h_i, x-x^k\rangle$$

for each $i \in I$, and, since μ^k is feasible in (3.1), we obtain

$$0 \geq \sum_{i\in I} \mu_i^k h_i^k - \langle c, y-y^k\rangle + \langle \sum_{i\in I} \mu_i^k \nabla_x h_i, x-x^k\rangle. \tag{3.15}$$

Adding this inequality to (3.14) and rearranging with the help of (3.5), (3.9) and (3.11), we get

$$\langle c,y\rangle + f(x) \geq \langle c,y^k\rangle + f(x^k) + \langle p_x^k, x-x^k\rangle - \tilde{\alpha}^k, \tag{3.16}$$

which proves (3.13). Setting $(y,x) = (y^k,x^k)$ in (3.14)-(3.16) we obtain $\tilde{\alpha}_p^k \geq 0$, $\alpha_h^k \geq 0$ and $\tilde{\alpha}^k \geq 0$. Then (3.8) yields $v^k \leq 0$, $|p_x^k| \leq |v^k|^{1/2}$ and $\tilde{\alpha}^k \leq -v^k$, so (3.13) and the Cauchy-Schwarz inequality imply (2.8), as desired.

We may now justify the stopping criterion.

Lemma 3.3. If Algorithm 2.1 terminates at the k-th iteration, then z^k solves problem (1.1).

Proof. Since the algorithm stops only if $0 \geq v^k \geq -\varepsilon_s = 0$, i.e. $v^k = 0$, z^k is optimal by (2.8) .

From now on we suppose that the algorithm generates an infinite sequence $\{z^k\}$.

We shall need the following consequence of relation (2.8) and the continuity of F.

Lemma 3.4. Suppose that there exist an infinite set $K \subset \{1,2,...\}$ and a point $\bar{z} \in S$ such that $z^k \xrightarrow{K} z$ and $v^k \xrightarrow{K} 0$. Then $\bar{z} \in \bar{Z}$.

Note that the rules of the algorithm imply

$$z^k = z^{k(l)} \quad \text{if} \quad k(l) \leq k < k(l+1) \ , \tag{3.17}$$

where we let $k(l+1) = +\infty$ if the number l of serious steps stays bounded, i.e. if $z^k = z^{k(l)}$ for some fixed l and all $k \geq k(l)$. Our first convergence result deals with the case of infinitely many serious steps.

Lemma 3.5. Suppose that there exist an infinite set $L \subset \{1,2,...\}$ and a point $\bar{z} \in S$ such that $z^{k(l)}{}_{l \in L} \longrightarrow \bar{z}$. Then $z^k \xrightarrow{K} \bar{z}$ and $v^k \xrightarrow{K} 0$ for $K = \{k(l+1)-1 : l \in L\}$.

Proof: This follows from (2.5) as in [K3].

Our next result deals with the case of a finite number of serious steps.

Lemma 3.6. Suppose that $z^k = z^{k(l)}$ for some fixed l and all $k \geq k(l)$. Then $w^k \downarrow 0$ and $v^k \longrightarrow 0$.

Proof. Suppose $z^k = z^{k(l)} = \bar{z} = (\bar{y},\bar{x})$ for all $k \geq k(l)$. We shall show that w^k vanishes by demonstrating that w^{k+1} is less than a fraction of w^k after each null step.

(i) Choose $k \geq k(1)$, so that $t_L^k=0$, $z^{k+1} = z^k$ and $\tilde{z}^{k+1} = z^{k+1}+d^k$. By the line search rules , $F(\tilde{z}^{k+1})-F(z^{k+1}) > mv^k$, hence

$$\langle c,d_y^k\rangle - \alpha_{k+1}^{k+1} + \langle g^{k+1}, d_x^k\rangle = \langle c,d_y^k\rangle - [f(x^{k+1})-f(\tilde{x}^{k+1}) -$$

$$- \langle g^{k+1}, x^{k+1}-\tilde{x}^{k+1}\rangle] + \langle g^{k+1}, d_x^k\rangle =$$

$$= - F(z^{k+1}) + F(\tilde{z}^{k+1}) > mv^k$$

from (3.2) and (2.6) . Expressing $\langle c,d_y^k\rangle$ in the above inequality via (3.12) and letting

$$\hat{g}^{k+1} = g^{k+1} + \sum_{i \in I} \mu_i^k \nabla_x h_i, \tag{3.18}$$

$$\hat{\alpha}^{k+1} = \alpha_{k+1}^{k+1} + \alpha_h^k = \alpha_{k+1}^{k+1} - \sum_{i \in I} \mu_i^k h_i^k, \tag{3.19}$$

we obtain

$$- \hat{\alpha}^{k+1} + \langle \hat{g}^{k+1}, d_x^k\rangle > mv^k \qquad \text{for all } k \geq k(1). \tag{3.20}$$

(ii) For each $\nu \in [0,1]$ and any fixed $k > k(1)$, define the multipliers

$$\lambda_k(\nu) = \nu \quad , \quad \lambda_j(\nu) = (1-\nu)\lambda_j^{k-1} \text{ for } j \in \hat{J}^{k-1},$$

$$\mu_i(\nu) = \mu_i^{k-1} \qquad \text{for } i \in I,$$

and check that they are feasible in (3.1), since $\{\lambda_j^{k-1}\}_{j \in \hat{J}^{k-1}}$ form a convex combination from Lemma 3.1, while $J^k = \hat{J}^{k-1} \cup \{k\}$. Moreover,

$$\sum_{j \in J^k} \lambda_j(\nu)\, g^j + \sum_{i \in I} \mu_i(\nu)\, \nabla_x h_i = (1-\nu)\, p_x^{k-1} + \nu\, \hat{g}^k ,$$

$$\sum_{j \in J^k} \lambda_j(\nu)\, \alpha_j^k - \sum_{i \in I} \mu_i(\nu)\, h_i^k = (1-\nu)\, \tilde{\alpha}^{k-1} + \nu\, \hat{\alpha}^k.$$

This follows from (3.6), (3.5), (3.18), (3.2) , (3.19) , (3.9)-(3.11), (2.7) and the fact that $x^k = x^{k-1}$. Next, define the function

$$Q^k(\nu) = \frac{1}{2}|(1-\nu)\, p_x^{k-1} + \nu \hat{g}^k|^2 + (1-\nu)\tilde{\alpha}^{k-1} + \nu\hat{\alpha}^k,$$
$$\nu \in [0,1]. \qquad (3.21)$$

Since w^k is the optimal value of (3.1), the preceding relations yield

$$w^k \le \min\{\, Q^k(\nu) : \nu \in [0,1]\}. \qquad (3.22)$$

In particular, by (3.7),

$$w^k \le Q^k(0) = \frac{1}{2}|p_x^{k-1}|^2 + \tilde{\alpha}^{k-1} = w^{k-1}. \qquad (3.23)$$

(iii) It follows from (3.23) that $|p_x^k|^2/2 + \tilde{\alpha}^k = w^k \le w^{k(1)}$ for each $k \ge k(1)$. Hence $|p_x^k|$, $\tilde{\alpha}^k$, $\tilde{\alpha}_p^k$ and α_h^k are uniformly bounded, since $\tilde{\alpha}^k$ is the sum of nonnegative $\tilde{\alpha}_p^k$ and α_h^k. Then (3.14) and the local boundedness of ε-subdifferentials imply boundedness of $\{p_f^k\}$, so also

$$\sum_{i \in I} \mu_i^k \nabla_x h_i = p_x^k - p_f^k$$

must be uniformly bounded. Since $g^k \in \partial f(\tilde{x}^k)$ with $\tilde{x}^k = x^{k-1} + d_x^{k-1}$ $= \bar{x} - p_x^{k-1}$ for $k > k(1)$, $\{g^k\}$ is bounded from the local boundedness of ∂f. Summing up, we deduce the existence of a constant C satisfying

$$\max\{\, |p_x^{k-1}|\ , |\hat{g}^k|, \tilde{\alpha}^{k-1}, 1\} \le C \quad \text{for all } k \ge k(1). \qquad (3.24)$$

(iv) It is shown in the proof of Theorem 3.5 in [K2] that (3.3), (3.7), (3.8) , (3.20)-(3.22) and the fact that $m \in (0,1)$ yield

$$0 \le w^{k+1} \le w^k - (1-m)^2 (w^k)^2 / 8C^2 \quad \text{for each } k \ge k(1).$$

We conclude that $w^k \downarrow 0$. Then $v^k \longrightarrow 0$ from (3.7), (3.3) and the fact that $\tilde{\alpha}^k \ge 0$ for all k. This completes the proof.

Combining Lemmas 2.3, 3.5 and 3.6 with (3.17), we deduce

<u>Theorem 3.7.</u> Every accumulation point of an infinite sequence $\{z^k\}$ generated by Algorithm 2.1 solves problem (1.1).

Corollary 3.8. If Algorithm 2.1 constructs an infinite sequence $\{z^k\}$, then a subsequence of $\{z^k\}$ converges to a solution of problem (1.1), and $\{z^k\}$ minimizes F on S, i.e. $\{z^k\} \subset S$ and $F(z^k) \downarrow \min \{F(z) : z \in S\}$.

Proof. Since $\{z^k\}$ stays in the closed and bounded set S, it has a feasible accumulation point, which is optimal by Theorem 3.7. Therefore the desired conclusion follows from the monotonicity of $\{F(z^k)\}$ and the continuity of F.

The above result may be strengthened as follows.

Corollary 3.9. There exist $\bar{z} = (\bar{y},\bar{x}) \in Z$ and an infinite set $K \subset \{1,2,...\}$ such that $x^k \to \bar{x}$ and $y^k \xrightarrow{K} \bar{y}$.

The proof of the above corollary will follow immediately from our subsequent results.

Up till now we have assumed that S is bounded, although this was only necessary for the proof of Corollary 3.8. Suppose now that S is unbounded. Then, of course, we can no longer guarantee that the k-th subproblem (2.1) has a solution. However, by convex quadratic programming theory (see,e.g.[PD1]), we know that either d^k exists or there is a sequence $\{\tilde{d}^i\}$ such that $z^k + \tilde{d}^i \in S$ for all i and

$$\langle c,\tilde{d}_y^i\rangle + \max \{ f_j^k + \langle g^j, \tilde{d}_x^i\rangle : j \in J^k\} + \tfrac{1}{2} |\tilde{d}_x^i|^2 \longrightarrow -\infty$$

as $i \to \infty$. This yields $\langle c,\tilde{d}_y^i\rangle \to -\infty$ with bounded $\{\tilde{d}_x^i\}$ (otherwise the quadratic term would dominate), so $F(x^k + \tilde{d}^i) \to -\infty$ as $i \to \infty$, implying that F is unbounded from below on S. Such a situation could be detected by the quadratic programming routine, e.g. MINOS.

Let us, therefore, suppose that the algorithm generates a (possibly unbounded) infinite sequence $\{z^k\}$ even if S is unbounded. We still have the following result.

Lemma 3.10. (i) Suppose $F(z^k) \geq F(\tilde{z})$ for some $\tilde{z} \in S$ and all k. Then there exists $(\hat{y},\hat{x}) \in S$ such that $x^k \longrightarrow \hat{x}$.
(ii) One has $F(z^k) \downarrow \inf\{F(z) : z \in S\}$.

Proof. (i) If $\tilde{z} = (\tilde{y},\tilde{x}) \in S$ and $F(\tilde{z}) \leq F(z^k)$, then (3.13) yields $\langle p_x^k, \tilde{x}-x^k\rangle \leq \tilde{\alpha}^k$. Hence one may use (2.5), (3.3) and (3.8) as in the proofs of Lemma 3.6 and Theorem 3.7 in [K2] to deduce that $\{x^k\}$ converges.
(ii) Suppose, for contradiction purposes, that $F(z^k) \geq F(\tilde{z}) + \varepsilon$ for some fixed $\tilde{z} = (\tilde{y},\tilde{x}) \in S$, $\varepsilon > 0$ and all k. By the first assertion, $\{x^k\}$ is bounded. Moreover, one may reason as in the proof of Lemma 3.9 in [K2] to show that a subsequence of $\{v^k\}$ tends to zero. Hence, by (2.8), we have $F(z^k) < F(\tilde{z}) + \varepsilon$ for some large k, a contradiction. Therefore, $\{z^k\}$ minimizes F on S.

We conclude from Lemma 3.10 that Corollary 3.9 holds if S is bounded. Also one can use the proof of Lemma 3.10 to show that if problem (1.1) has a solution (e.g. the set $\{z \in S : F(z) \leq F(z^1)\}$ is bounded) and the final accuracy tolerance ε_s is positive, then the algorithm will terminate in a finite number of iteration after finding an approximate solution to problem (1.1).

4. Conclusions

We have presented an implementable and globally convergent method of descent for solving large-scale linearly constrained convex nonsmooth minimization problems with relatively few nonlinear variables, such as those arising in stochastic programming [NW1]. The method seems to be unique among the existing algorithms in that its direction finding subproblems can be solved by the existing software for large-scale smooth optimization [MS1]. Therefore, we hope that the method should prove useful in calculations.

We may add that the method can be extended to the case when both the nonlinear part of the objective and its subgradients can be evaluated only approximately. Also more efficient line searches (see [M2]) can be employed. These extensions, as well as finite termination in the piecewise linear case, will be discussed elsewhere.

Acknowledgment

I would like to thank Larry Nazareth and Roger Wets for the discussions we have had on the subject of this work.

References

[DV1] V.F. Demyanov and L.V. Vasiliev, Nondifferentiable Optimization, (in Russian) , Nauka, Moscow, 1981 .

[K1] P.Kall, "Stochastic programming", European Journal of Operations Research 10 (1982) 125-130.

[K2] K.C. Kiwiel, "An aggregate subgradient method for nonsmooth convex minimization", Mathematical Programming 27 (1983) 320-341.

[K3] K.C. Kiwiel, "An algorithm for linearly constrained convex nondifferentiable minimization problems", Journal of Mathematical Analysis and Applications (to appear).

[K4] K.C. Kiwiel, "A linearization algorithm for constrained nonsmooth minimization", in: P. Thoft-Christensen, e.d., System Modelling and Optimization, Lecture Notes in Control and Information Sciences, Springer, Berlin (to appear).

[LSB1] C. Lemarechal, J.-J. Strodiot and A. Bihain, "On a bundle algorithm for nonsmooth minimization, in: O.L. Mangasarian, R.R. Meyer and S.M. Robinson, eds., Nonlinear Programming 4, Academic Press, New York, 1981, pp.245-281.

[M1] R. Mifflin, "An algorithm for constrained optimization with semismooth functions", Mathematics of Operations Research 2 (1977) 191-207 .

[M2] R. Mifflin, "A modification and an extension of Lemarechal's algorithm for nonsmooth minimization", in: D.C. Sorensen and R.J.-B. Wets, eds., Nondifferential and Variational Techniques in Optimization, Mathematical Programming Study 17, North Holland, Amsterdam, 1982, pp.77-90.

[MS1] B.A. Murtagh and M.A. Saunders, "Large-scale linearly constrained optimization", Mathematical Programming 14 (1978), 41-72.

[NW1] L.Nazareth and R.J.-B. Wets, "Algorithms for stochastic programs: the case of nonstochastic tenders", WP-83-5, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1983.

[PD1] B.N. Pshenichny and Yu.M. Danilin , Numerical Methods in Extremal Problems, Mir, Moscow, 1980.

[PMW1] E. Polak, D.Q. Mayne and Y. Wardi, "On the extension of constrained optimization algorithms from differentiable to nondifferentiable problems", SIAM Journal on Control and Optimization 21 (1983) 179-203.

[SNH1] J.-J. Strodiot, V. Hien Nguyen and N. Heukemes," ε-Optimal solutions in nondifferentiable convex programming and related questions", Mathematical Programming 25 (1982) 307-323.

[W1] R.J.-B. Wets, "Stochastic programming: solution techniques and approximation schemes", in: A.Bachem, M. Groetschel and B. Korte, eds., Mathematical Programming: The State-of-the-Art, Springer, Berlin, 1983, pp.566-603..

[W2] A.P. Wierzbicki, "Lagrangian functions in nondifferentiable optimization, in: E.A. Nurminski, ed., Progress in Nondifferentiable Optimization, CP-82-S8, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1982, pp. 173-213.