Mathematical Programming 71 (1995) 207–219

A generalized subgradient method with relaxation step [1]

Ulf Brännlund *

Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology,
S-100 44 Stockholm, Sweden

Received 3 October 1991

Abstract

We study conditions for convergence of a generalized subgradient algorithm in which a relaxation step is taken in a direction, which is a convex combination of possibly all previously generated subgradients. A simple condition for convergence is given and conditions that guarantee a linear convergence rate are also presented. We show that choosing the steplength parameter and convex combination of subgradients in a certain sense optimally is equivalent to solving a minimum norm quadratic programming problem. It is also shown that if the direction is restricted to be a convex combination of the current subgradient and the previous direction, then an optimal choice of stepsize and direction is equivalent to the Camerini–Fratta–Maffioli modification of the subgradient method.

Keywords: Subgradient optimization; Relaxation methods; Projection methods

1. Introduction

This paper is concerned with minimizing a convex function $f: \mathbb{R}^n \to \mathbb{R}$, which is not necessarily differentiable. The well-known subgradient algorithm generates iterates x_k, $k \geqslant 1$, according to the recursion

$$x_{k+1} = x_k - h_k g_k,$$

where g_k is a subgradient of f at x_k and h_k is the steplength.

Popular steplength rules for which there exist global convergence proofs are:

* e-mail: uffe@math.kth.se.

[1] Research supported by the Swedish Research Council for Engineering Sciences (TFR).

SSDI 0025-5610(95)00014-3

Case 1. $h_k = \gamma_k(f(x_k) - f^*)/\|g_k\|^2$, where $0 < \delta \leqslant \gamma_k \leqslant 2 - \delta$, and f^* is the known optimal value;

Case 2. $h_k > 0$, $\lim_{k\to\infty} h_k\|g_k\| = 0$ and $\sum_{k=1}^{\infty} h_k\|g_k\| = \infty$;

Case 3. $h_k > 0$, $\lim_{k\to\infty} h_k = 0$ and $\sum_{k=1}^{\infty} h_k = \infty$ and $\{g_k\}_{k=1}^{\infty}$ is bounded.

Case 1 is called the *relaxation step* from its origin in solving systems of linear inequalities, see [1,5,13]. Some researchers call Case 1 the *Polyak step*, from its reinventor who generalized the relaxation method to infinitely many inequalities, i.e., convex optimization with a known optimal value, see [14].

A natural generalization of the subgradient method is to take a step in a direction that is a convex combination of previous subgradients, possibly all. Kim and Ahn [6] study Case 2 and Case 3 and give convergence proofs for these cases. In this paper Case 1 is analyzed.

To be specific, the algorithm analyzed in this paper is

$$x_{k+1} = x_k - h_k d_k, \tag{1}$$

where

$$d_k = \sum_{i=1}^{k} \alpha_i^k g_i, \quad \alpha_i^k \geqslant 0, \quad \sum_{i=1}^{k} \alpha_i^k = 1,$$

and

$$h_k = \gamma_k \frac{f(x_k) - f^*}{\| d_k \|^2}, \quad 0 < \delta \leqslant \gamma_k \leqslant 2 - \delta,$$

and f^* is the optimal value. At each iterate x_i a black box is called, which delivers the function value $f(x_i)$ and one subgradient g_i at x_i.

Section 2 contains some preliminaries and the basic convergence theorem and theorems on the rate of convergence. In Section 3 we point out relations to other methods. In particular we show that the natural choice of steplength and convex combination of subgradients leads to the popular Camerini–Fratta–Maffioli modification of the subgradient method and to so-called bundle level methods.

We use the notation X^* for the set of optimal solutions. A vector g_ϵ is said to be an *ϵ-subgradient* of f at x if $f(y) \geqslant f(x) + g_\epsilon^{\mathrm{T}}(y - x) - \epsilon$ for all y. A subgradient g of f at x is a 0-subgradient. The distance from a point x to a set X is denoted by $\operatorname{dist}(x, X) = \min_{y\in X} \|y - x\|$.

2. Convergence

The following two lemmas and the corollary concerning subgradients and ϵ-subgradients and their proofs can be found in [7].

Lemma 2.1. *Let g_1 be an ϵ_1-subgradient of f at x_1. Then g_1 is an ϵ_2-subgradient of f at x_2, with $\epsilon_2 = f(x_2) - f(x_1) - g_1^{\mathrm{T}}(x_2 - x_1) + \epsilon_1$.*

Lemma 2.2. *Let g_1 and g_2 be, respectively, ϵ_1-subgradient and ϵ_2-subgradient of f at x. Then, a convex combination $\alpha g_1 + (1-\alpha)g_2$, $\alpha \in [0,1]$, of g_1 and g_2 is an $(\alpha\epsilon_1 + (1-\alpha)\epsilon_2)$-subgradient of f at x.*

Corollary 2.3. *A direction d_k, obtained as a convex combination of previously generated subgradients g_i at x_i, $d_k = \sum_{i=1}^k \alpha_i^k g_i$, $\alpha_i^k \geqslant 0$, $\sum_{i=1}^k \alpha_i^k = 1$, is an ϵ_k-subgradient at x_k, with $\epsilon_k = \sum_{i=1}^k \alpha_i^k \epsilon_k^i$, where $\epsilon_k^i = f(x_k) - f(x_i) - g_i^{\mathrm{T}}(x_k - x_i)$.*

The numbers ϵ_k^i can be calculated recursively by the formula

$$\begin{aligned} &\epsilon_k^i = \epsilon_{k-1}^i + f(x_k) - f(x_{k-1}) - g_i^{\mathrm{T}}(x_k - x_{k-1}), \quad \text{for } i = 1,\dots,k-1, \\ &\epsilon_k^k = 0. \end{aligned} \tag{2}$$

Thus, it is only necessary to store previous subgradients, not previous iterates nor function values, to update the numbers ϵ_k^i from one iteration to another.

To show convergence to an optimal point of our generalized subgradient algorithm, we have to add a restriction on to what extent the direction can deviate from being a subgradient. To be more specific, we require the direction to be an ϵ_k-subgradient, where ϵ_k is bounded away from $\frac{1}{2}(2-\gamma_k)(f(x_k) - f^*)$.

Theorem 2.4. *Let f be a convex function defined on $\mathbb{R}^n$ which has a nonempty set of minimum points X^*. For any x_1 consider the sequence of points generated according to algorithm (1). Furthermore, suppose that there exists $\xi \in (0,1]$ such that*

$$\epsilon_k = \sum_{i=1}^k \alpha_i^k \epsilon_k^i \leqslant \tfrac{1}{2}(1-\xi)(2-\gamma_k)(f(x_k) - f^*), \tag{3}$$

where $\epsilon_k^i = f(x_k) - f(x_i) - g_i^{\mathrm{T}}(x_k - x_i)$. Then, $\lim_{k\to\infty} x_k \in X^*$.

Proof. Let $x_k \notin X^*$ and $y \in X^*$. Then,

$$\|x_{k+1} - y\|^2 = \|x_k - y\|^2 - 2\gamma_k \frac{f(x_k) - f^*}{\|d_k\|^2} d_k^{\mathrm{T}}(x_k - y) + \gamma_k^2 \frac{(f(x_k) - f^*)^2}{\|d_k\|^2}. \tag{4}$$

Now observe that d_k is an ϵ_k-subgradient and thus

$$0 \leqslant f(x_k) - f^* = f(x_k) - f(y) \leqslant d_k^{\mathrm{T}}(x_k - y) + \epsilon_k,$$

which, used to bound $d_k^{\mathrm{T}}(x_k - y)$ in (4), gives

$$\|x_{k+1} - y\|^2 \leqslant \|x_k - y\|^2 - \gamma_k(2-\gamma_k)\frac{(f(x_k) - f^*)^2}{\|d_k\|^2} + 2\gamma_k \frac{f(x_k) - f^*}{\|d_k\|^2}\epsilon_k. \tag{5}$$

With (3), we get

$$\|x_{k+1} - y\|^2 \leqslant \|x_k - y\|^2 - \xi\gamma_k(2-\gamma_k)\frac{(f(x_k) - f^*)^2}{\|d_k\|^2} < \|x_k - y\|^2. \tag{6}$$

Due to the monotone convergence of the distance $\|x_k - y\|$, the sequences $\{x_k\}_{k=1}^{\infty}$ and $\{g_k\}_{k=1}^{\infty}$ are bounded, since f is defined on $\mathbb{R}^n$ and convex. Thus, the sequence $\{d_k\}_{k=1}^{\infty}$ is also bounded. Let x^* be an accumulation point of $\{x_k\}_{k=1}^{\infty}$ and suppose that $f(x^*) > f^*$; then there exist a $\delta > 0$ and an infinite sequence of indices $k_1 < k_2 < \cdots$ such that $f(x_{k_i}) - f^* > \delta$, but with the boundedness of $\{d_k\}_{k=1}^{\infty}$ and $0 < \delta \leqslant \gamma_k \leqslant 2-\delta$ this implies from (6) that

$$\|x_{k_i+1} - y\|^2 \leqslant \|x_{k_i} - y\|^2 - \eta,$$

for some $\eta > 0$. Hence $\|x_{k_i+1} - y\|^2 \leqslant \|x_{k_1} - y\|^2 - i\eta$, which is a contradiction. Therefore, $x^* \in X^*$.

The uniqueness of x^* follows from the monotonic decrease of $\|x_k - y\|$. □

One should note that it is actually only required that in each iteration the number

$$\gamma_k(2-\gamma_k) - 2\gamma_k\eta_k, \tag{7}$$

where $\eta_k = \epsilon_k/(f(x_k) - f^*)$, is bounded away from zero. With this requirement added, the requirement $0 < \delta \leqslant \gamma_k \leqslant 2 - \delta$ can be omitted and the parameter ξ eliminated.

The following theorem states a result of linear convergence rate under a rather restrictive assumption, which is satisfied in the piecewise linear case, for example in the important case of Lagrangian relaxation of discrete optimization problems.

Theorem 2.5. *Let f be a convex function and let L be its Lipschitz constant in the area $\|x - x^*\| \leqslant \|x_1 - x^*\|$. Suppose that there exists an $m > 0$ such that for each $x \notin X^*$ the relation $f(x) - f^* \geqslant m\,\mathrm{dist}(x, X^*)$ holds. Then at each step of the generalized subgradient algorithm defined by* (1) *with $\gamma_k = \gamma$, the relation*

$$\mathrm{dist}(x_{k+1}, X^*) \leqslant q\,\mathrm{dist}(x_k, X^*)$$

holds, where

$$q = \left(1 - \frac{\gamma(2-\gamma)\xi m^2}{L^2}\right)^{1/2}.$$

Proof. Straightforward from (6) and the assumptions. □

[16, Theorem 2.12], which concerns the speed of convergence in the case of a strongly convex differentiable function, can also be translated to the general case. However, an extra condition on the direction d_k is needed, namely $\|d_k\| \leqslant \|g_k\|$. Such a requirement may be natural in this general setting, see Sections 3.1 and 3.2.

Theorem 2.6. *The generalized method defined by* (1), *with the extra restriction that $\|d_k\| \leqslant \|g_k\|$, converges at a linear rate if f satisfies $f(x) - f^* \geqslant m\|x - x^*\|^2$ for some $m > 0$, and the gradients of f are Lipschitz continuous with parameter L in the area $\|x - x^*\| \leqslant \|x_1 - x^*\|$.*

Proof. Straightforward using (6) and the fact that the assumptions imply $\|d_k\| \leqslant \|g_k\| = \|g_k - g(x^*)\| \leqslant L\|x_k - x^*\|$. □

3. Connections to other methods

From the proof of Theorem 2.4, we see that for the minimal rate of convergence to be as large as possible, γ_k and α_i^k should be chosen in each iteration so that

$$F(\gamma_k, \alpha^k) = \frac{\gamma_k(2-\gamma_k) - 2\gamma_k \sum_{i=1}^k \alpha_i^k \eta_k^i}{\| \sum_{i=1}^k \alpha_i^k g_i \|^2}, \tag{8}$$

where $\eta_k^i = \epsilon_k^i/(f(x_k) - f^*)$, is as large as possible. The reason for this being that in each iteration

$$\|x_{k+1} - y\|^2 \leqslant \|x_k - y\|^2 - F(\gamma_k, \alpha^k)(f(x_k) - f^*)^2, \quad \forall y \in X^*,$$

and hence if $F(\gamma_k, \alpha^k)$ is maximized, then the guaranteed decrease of the distance to the optimal set is maximized.

3.1. A popular modification

The problem of maximizing $F(\gamma_k, \alpha^k)$ is treated in the next section; but in order to get some insight, we consider a restricted version of the general subgradient method. In each step we take a convex combination of the new subgradient and the last direction to get our new direction:

$$d_k = \alpha_k g_k + (1-\alpha_k) d_{k-1}, \quad 0 \leqslant \alpha_k \leqslant 1.$$

For $k = 1$ we choose $\alpha_1 = 1$.

In this section it is shown that the very popular modified subgradient method proposed by Camerini, Fratta and Maffioli [4] can be viewed as choosing an "optimal" α as well as an "optimal" steplength parameter γ.

To simplify the notation, we drop the iteration index k, let g be a subgradient at the current iteration point x, let the previous direction be denoted by d, and assume that d is an ϵ-subgradient at x. Denote by d_+ the direction $d_+ = \alpha g + (1-\alpha)d$ and let the new iterate $x_+ = x - \gamma(f(x) - f^*)d_+/\|d_+\|^2$. The direction d_+ defined in this way is a $(1-\alpha)\epsilon$-subgradient at x.

In this setting we want to maximize

$$F(\gamma, \alpha) = \frac{\gamma(2-\gamma) - 2\gamma(1-\alpha)\eta}{\|\alpha g + (1-\alpha)d\|^2} = \frac{N(\gamma, \alpha)}{\|d_+\|^2}, \tag{9}$$

where $\eta = \epsilon/(f(x) - f^*)$ and $N(\gamma, \alpha)$ corresponds to the term (7). We start with maximizing F with respect to γ. The first-order necessary conditions yield

$$\hat{\gamma} = 1 - (1-\hat{\alpha})\eta. \tag{10}$$

Here is a rather surprising fact.

Proposition 3.1. *If γ_k and α_k are chosen according to* (10), *then $\eta_k = 1$ for all k.*

Proof. The update formula for ϵ-subgradients (2) yields

$$\begin{aligned}\epsilon_{k+1} &= f(x_{k+1}) - f(x_k) - d_k^{\mathrm{T}}(x_{k+1} - x_k) + (1-\alpha_k)\epsilon_k \\ &= f(x_{k+1}) - f(x_k) + d_k^{\mathrm{T}}\left(\gamma_k \frac{f(x_k) - f^*}{d_k^{\mathrm{T}} d_k}\right) d_k + (1-\alpha_k)\epsilon_k \\ &= f(x_{k+1}) - f(x_k) + \gamma_k(f(x_k) - f^*) + (1-\alpha_k)\epsilon_k.\end{aligned}$$

Eq. (10) can be rewritten as

$$\gamma_k(f(x_k) - f^*) + (1-\alpha_k)\epsilon_k = f(x_k) - f^*,$$

and thus

$$\epsilon_{k+1} = f(x_{k+1}) - f^*.$$

The proposition follows from the definition of η_k. □

This result can be explained by the nature of relaxation methods like this one: at each iteration, f is replaced by an affine function $l(x)$, where in this case $l(x) = f(x_k) + d_k^{\mathrm{T}}(x - x_k) - \epsilon_k$ and then the equality $l(x) = f^*$ is imposed. From the ϵ-subgradient update formula (2) we have $\epsilon_{k+1} = f(x_{k+1}) - l(x_{k+1}) = f(x_{k+1}) - f^*$.

Proposition 3.1 and condition (10) imply that $\hat{\gamma} = \hat{\alpha}$.

Substituting $\gamma = \alpha$ into (9) and inverting F, we get the minimization problem

$$\min_{\alpha\in[0,1]} \left\| g - d + \left(\frac{1}{\alpha}\right) d \right\|^2. \tag{11}$$

The objective of (11) is convex in the variable $\beta = 1/\alpha \geqslant 1$ and its optimal solution is

$$\hat{\alpha} = \hat{\gamma} = \begin{cases} d^{\mathrm{T}}d/(d^{\mathrm{T}}d - g^{\mathrm{T}}d), & \text{if } g^{\mathrm{T}}d < 0, \\ 1, & \text{otherwise.}\end{cases} \tag{12}$$

This optimal solution yields

$$N(\hat{\gamma}, \hat{\alpha}) = \begin{cases} (d^{\mathrm{T}}d)^2/(d^{\mathrm{T}}d - g^{\mathrm{T}}d)^2, & \text{if } g^{\mathrm{T}}d < 0, \\ 1, & \text{otherwise,}\end{cases}$$

and

$$F(\hat{\gamma}, \hat{\alpha}) = \begin{cases} d^{\mathrm{T}}d/(d^{\mathrm{T}}dg^{\mathrm{T}}g - (g^{\mathrm{T}}d)^2), & \text{if } g^{\mathrm{T}}d < 0, \\ 1/g^{\mathrm{T}}g, & \text{otherwise.}\end{cases}$$

It is easy to verify that the resulting algorithm is convergent since $F(\hat{\gamma}, \hat{\alpha}) \geqslant 1/g^{\mathrm{T}}g$, although $N(\hat{\gamma}, \hat{\alpha})$ may not be bounded from zero.

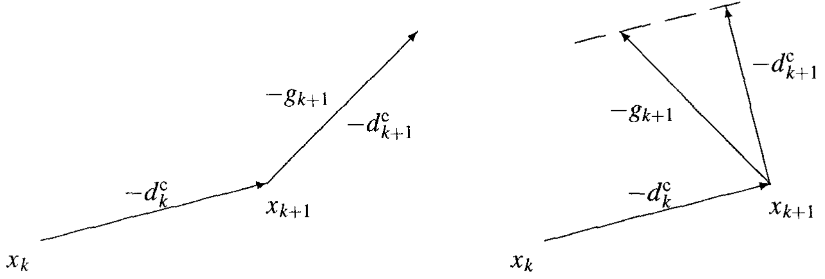

Fig. 1. The CFM-modification with $\lambda^{\mathrm{c}} = 1$; (left) not zigzagging; (right) zigzagging.

The following lemma characterizes the "optimal" choice of direction.

Lemma 3.2. *If $g^{\mathrm{T}}d < 0$, then the new direction $d_+ = \hat{\alpha}g + (1 - \hat{\alpha})d$ is orthogonal to d.*

Proof. Straightforward computations show that $d_+^{\mathrm{T}}d = 0$. □

Let us now review the modification proposed by Camerini et al. [4], which aims at trying to avoid zigzagging. They choose the direction

$$d^{\mathrm{c}}_k = g_k + \beta^{\mathrm{c}}_k d^{\mathrm{c}}_{k-1}, \tag{13}$$

where

$$\beta^{\mathrm{c}}_k = \begin{cases} -\lambda^{\mathrm{c}}_k g_k^{\mathrm{T}} d^{\mathrm{c}}_{k-1}/\|d^{\mathrm{c}}_{k-1}\|^2, & \text{if } g_k^{\mathrm{T}} d^{\mathrm{c}}_{k-1} < 0, \\ 0, & \text{otherwise,} \end{cases} \tag{14}$$

where $0 \leqslant \lambda^{\mathrm{c}}_k < 2$ and the steplength

$$h^{\mathrm{c}}_k = \gamma^{\mathrm{c}}_k \frac{f(x_k) - f^*}{\|d^{\mathrm{c}}_k\|^2}, \quad 0 < \gamma^{\mathrm{c}}_k \leqslant 1. \tag{15}$$

The choice $\lambda^{\mathrm{c}}_k = 1$ yields d^{c}_k orthogonal to d^{c}_{k-1} if $g_k^{\mathrm{T}} d^{\mathrm{c}}_{k-1} < 0$. This choice is illustrated in Fig. 1.

Theorem 3.3. *The iterates generated by the "optimal" choice of stepsize and direction, i.e.,*

$$x_{k+1} = x_k - \gamma_k \frac{f(x_k) - f^*}{\|d_k\|^2} d_k,$$

where

$$d_k = \alpha_k g_k + (1 - \alpha_k) d_{k-1}$$

and

$$\alpha_k = \gamma_k = \begin{cases} (d_{k-1}^{\mathrm{T}} d_{k-1})/(d_{k-1}^{\mathrm{T}} d_{k-1} - g_k^{\mathrm{T}} d_{k-1}), & \textit{if } g_k^{\mathrm{T}} d_{k-1} < 0, \\ 1, & \textit{otherwise,} \end{cases}$$

are exactly the same as the ones generated by the modification of Camerini et al. [4], *if* $\lambda_k^{\mathrm{c}} = \gamma_k^{\mathrm{c}} = 1$.

Proof. The theorem follows by induction. Straightforward computation shows that if $d_{k-1} = \beta d_{k-1}^{\mathrm{c}}$ for some $\beta > 0$ and $x_k = x_k^{\mathrm{c}}$, then the two algorithms will generate the same new iterate $x_{k+1} = x_{k+1}^{\mathrm{c}}$. Lemma 3.2 and Fig. 1 ensure that $d_k = \beta d_k^{\mathrm{c}}$ for some $\beta \in (0, 1]$ with $\beta = 1$ if $g_k^{\mathrm{T}} d_{k-1} \geqslant 0$. □

A comment is in order here. Camerini et al. [4] suggest from geometric interpretation of their modification that λ^{c} should be chosen larger than 1 and this choice is also supported by computational evidence. What we have shown is that in order to maximize the minimal guaranteed speed of convergence, it should be chosen as 1. It may of course *on average* be better to choose some other number, such as the one that Camerini et al. [4] suggest.

3.2. An optimal relaxation method

Let us go back to the problem of maximizing $F(\gamma_k, \alpha^k)$. Again, to simplify the notation, we drop the iteration index k. We also introduce vector notation wherever possible. In particular, α is the k-dimensional column vector of α_i's, η a column vector of η_i's, e a vector of k ones, and G an $n \times k$-matrix composed of the k subgradients. Using this notation, the problem of maximizing F takes the form

$$\begin{aligned} \sup_{\alpha,\gamma} \quad & \frac{\gamma(2-\gamma) - 2\gamma\alpha^{\mathrm{T}}\eta}{\alpha^{\mathrm{T}} G^{\mathrm{T}} G \alpha} \\ \text{s.t.} \quad & \gamma(2-\gamma) - 2\gamma\alpha^{\mathrm{T}}\eta > 0 \\ & e^{\mathrm{T}}\alpha = 1 \\ & \alpha \geqslant 0 \\ & \gamma \geqslant 0. \end{aligned} \tag{16}$$

Note that the objective is concave with respect to γ and that the optimal γ is

$$\hat{\gamma} = 1 - \alpha^{\mathrm{T}}\eta. \tag{17}$$

Reducing problem (16) by insertion of (17), one gets after inversion and a strictly monotone transformation of the objective,

$$\begin{aligned} \inf_{\alpha} \quad & \frac{\sqrt{\alpha^{\mathrm{T}} G^{\mathrm{T}} G \alpha}}{1 - \alpha^{\mathrm{T}}\eta} \\ \text{s.t.} \quad & \alpha^{\mathrm{T}}\eta < 1 \\ & e^{\mathrm{T}}\alpha = 1 \\ & \alpha \geqslant 0. \end{aligned} \tag{18}$$

The numerator of the objective of problem (18) is convex since it is a norm and the denominator is affine. Problem (18) is thus a fractional program of the form

$$\begin{aligned} \inf \quad & \frac{f(x)}{g(x)} \\ \text{s.t.} \quad & x \in S, \end{aligned} \tag{19}$$

where $f(x)$ is convex, S is a convex set and g is affine and positive on S.

The following lemma, which is due to Schaible [15], provides a transformation which simplifies the solution of such a problem by converting it into a *convex* programming problem.

Lemma 3.4. *A program of the form* (19) *can be transformed into a convex program*

$$\begin{aligned} \inf \quad & tf(y/t) \\ \textit{s.t.} \quad & tg(y/t) = 1 \\ & y/t \in S \\ & t > 0, \end{aligned}$$

by the transformation

$$y = \frac{x}{g(x)}, \qquad t = \frac{1}{g(x)}. \tag{20}$$

The transformed problem has a solution if and only if (19) *has one and the solutions are connected by* (20).

Proof. See [15]. □

Transforming problem (18) according to Lemma 3.4 with

$$\beta = \frac{\alpha}{1 - \alpha^{\mathrm{T}}\eta} = t\alpha, \qquad t = \frac{1}{1 - \alpha^{\mathrm{T}}\eta}, \tag{21}$$

we get

$$\begin{aligned} \inf_{t,\beta} \quad & \sqrt{\beta^{\mathrm{T}} G^{\mathrm{T}} G \beta} \\ \text{s.t.} \quad & t - \beta^{\mathrm{T}}\eta = 1 \\ & e^{\mathrm{T}}\beta = t \\ & \beta^{\mathrm{T}}\eta < t \\ & \beta \geqslant 0 \\ & t > 0. \end{aligned} \tag{22}$$

The constraints $\beta^{\mathrm{T}}\eta < t$ and $t > 0$ are redundant, since β and η are nonnegative and $t - \beta^{\mathrm{T}}\eta = 1$ hold. Furthermore, t can be eliminated from the problem by combining $e^{\mathrm{T}}\beta = t$ and $t - \beta^{\mathrm{T}}\eta = 1$. Thus, we obtain the equivalent problem

$$\begin{aligned} \min_{\beta} \quad & \tfrac{1}{2}\beta^{\mathrm{T}}G^{\mathrm{T}}G\beta \\ \text{s.t.} \quad & (e^{\mathrm{T}} - \eta^{\mathrm{T}})\beta = 1 \\ & \beta \geqslant 0. \end{aligned} \tag{23}$$

In the variables β, the optimal steplength parameter $\hat{\gamma}$ can be expressed as $\hat{\gamma} = 1 - \hat{\alpha}^{\mathrm{T}}\eta = 1/e^{\mathrm{T}}\hat{\beta}$ using (21) and the new iteration point as

$$x_+ = x - \hat{\gamma}\frac{f(x) - f^*}{\hat{\alpha}^{\mathrm{T}}G^{\mathrm{T}}G\hat{\alpha}}G\hat{\alpha} = x - \frac{f(x) - f^*}{\hat{\beta}^{\mathrm{T}}G^{\mathrm{T}}G\hat{\beta}}G\hat{\beta}. \tag{24}$$

The Kuhn–Tucker conditions for problem (23) are as follows. A feasible solution $\hat{\beta}$ of (23) is optimal if and only if there exist multipliers μ and $\nu \geqslant 0$ such that

$$G^{\mathrm{T}}G\hat{\beta} - \mu e + \mu\eta - \nu = 0 \quad \text{and} \quad \hat{\beta}^{\mathrm{T}}\nu = 0. \tag{25}$$

These conditions (25) imply that $\mu = \hat{\beta}^{\mathrm{T}}G^{\mathrm{T}}G\hat{\beta}$.

Analogous to Proposition 3.1, we have that indices corresponding to the "active" subgradients in the solution of (23) will have $\eta_i = 1$ in the next iteration.

Proposition 3.5. *If $\hat{\beta}_i > 0$, then $\eta_i^+ = 1$. If $\hat{\beta}_i = 0$, then $\eta_i^+ \geqslant 1$.*

Proof. From the update formula of ϵ-subgradients (2), (24) and the Kuhn–Tucker conditions (25) we have

$$\begin{aligned} \epsilon_i^+ &= f(x_+) - f(x) - g_i^{\mathrm{T}}(x_+ - x) + \epsilon_i \\ &= f(x_+) - f(x) + g_i^{\mathrm{T}}\frac{f(x) - f^*}{\hat{\beta}^{\mathrm{T}}G^{\mathrm{T}}G\hat{\beta}}G\hat{\beta} + \epsilon_i \\ &= f(x_+) - f(x) + (\mu - \mu\eta_i + \nu_i)\frac{f(x) - f^*}{\hat{\beta}^{\mathrm{T}}G^{\mathrm{T}}G\hat{\beta}} + \epsilon_i \\ &= f(x_+) - f^* + \left(1 - \frac{\epsilon_i}{f(x) - f^*}\right)(f(x) - f^*) + \frac{f(x) - f^*}{\hat{\beta}^{\mathrm{T}}G^{\mathrm{T}}G\hat{\beta}}\nu_i + \epsilon_i \\ &= f(x_+) - f^* + \frac{f(x) - f^*}{\hat{\beta}^{\mathrm{T}}G^{\mathrm{T}}G\hat{\beta}}\nu_i. \end{aligned}$$

The proposition follows from the definition of η_i and the complementary slackness condition $\beta_i\nu_i = 0$. □

In analogy to the explanation of Proposition 3.1, we have for affine linearizations $l_i(x) = f(x_k) + g_i^{\mathrm{T}}(x - x_i) - \epsilon_k^i$, which are "active", that $l_i(x) = f^*$. This suggests the following theorem.

Theorem 3.6. *Choosing the parameters γ and α_i, $i = 1, \ldots, k$, in the generalized algorithm (1) optimally in each iteration, in the sense that the guaranteed decrease of the distance to the optimum is maximized, is equivalent to solving the problem*

$$\begin{array}{ll} \min\limits_{x} & \tfrac{1}{2}(x-x_k)^{\mathrm{T}}(x-x_k) \\ \textit{s.t.} & f^* \geqslant f(x_i) + g_i^{\mathrm{T}}(x-x_i), \quad \textit{for } i=1,\dots,k, \end{array} \tag{26}$$

and letting the solution be the new iteration point.

Proof. Rewrite problem (26) as

$$\begin{array}{ll} \min\limits_{d} & \tfrac{1}{2}d^{\mathrm{T}}d \\ \text{s.t.} & f^* \geqslant f(x_k) + g_i^{\mathrm{T}}d - \epsilon_i, \quad \text{for } i=1,\dots,k, \end{array} \tag{27}$$

using $d = x - x_k$. By the Kuhn-Tucker conditions, a feasible $\hat{d}$ is a solution to (27) if and only if there exist multipliers $u \geqslant 0$ such that $\hat{d} = -Gu$ and

$$(f(x_k) - f^*)u^{\mathrm{T}}e + u^{\mathrm{T}}G^{\mathrm{T}}\hat{d} - u^{\mathrm{T}}\epsilon = 0. \tag{28}$$

Let $\hat{\beta}$ solve (23) and $\hat{\mu}$ and $\hat{\nu}$ be the corresponding multipliers. It can easily be shown, using (25) and feasibility of $\hat{\beta}$, that $u = (f(x_k) - f^*)\hat{\beta}/\hat{\mu}$ satisfies (28). It then follows from (24) and (25) that

$$\hat{d} = -Gu = -\frac{f(x_k) - f^*}{\hat{\mu}}G\hat{\beta} = -\frac{f(x_k) - f^*}{\hat{\beta}^{\mathrm{T}}G^{\mathrm{T}}G\hat{\beta}}G\hat{\beta} = x_+ - x_k.$$

The proof is complete. □

We call the algorithm of Theorem 3.6 *the optimal relaxation method.*

Of course, from a computational point of view, one would prefer to solve the dual of (26), or to solve (23).

4. Some additional comments

The optimal relaxation algorithm is also presented in Polyak's paper [14], although its optimality was not shown. This part of Polyak's paper seems to be less known to the optimization community. (The present author did not know about it until after the main part of this work was completed.) The reason for this may be that solving the quadratic minimum norm problem (or its dual) may in those days have been considered too difficult and thus this procedure may have been viewed as not implementable. Today, however, there exist very efficient and reliable QP-solvers that are also easy to use.

We believe that this way of computing the next iterate should be used in Lagrangian relaxation when the subproblems are considerably harder than the subgradient iterations. It is often possible to find an underestimate $\bar{f}$ of the optimal function value, by a feasible solution to the primal problem. Of course, we have as in the regular relaxation step case the result that for any $\epsilon > 0$ there exists a finite $\bar{k}$ such that $f(x_{\bar{k}}) \leqslant 2f^* - \bar{f} + \epsilon$, which can be used in heuristic procedures, see, e.g., [12]. Furthermore, as the number of supporting hyperplanes increases, they eventually will provide a better lower bound than $\bar{f}$. This idea was explored recently in so-called level methods, see [11].

Another interesting connection is the one to bundle descent methods. We will refrain from giving a thorough description of these methods, but refer the interested reader to [8, 10] and the references therein. In these methods, QP-problems similar to (26) are solved. They maintain a control parameter which is adjusted at each iteration in order to guarantee that the step is a descent step. A different control parameter would be an estimate of the optimal function value. It is possible to design a descent algorithm based on this control, as is shown in [3].

Another algorithm that benefits from this approach is the algorithm presented in [2], a convergent algorithm based on lower and upper estimates of the optimal function value. It could easily be modified in order to use this elaborate way of computing the next iteration point.

Finally, we point out that Kiwiel [9] has established similar optimality properties of the Camerini–Fratta–Maffioli modification and of the generalization in Theorem 3.6. Kiwiel's derivation depends on projection properties and is not in the framework of taking steps in directions that are convex combinations of previously generated subgradients, as is the derivation of this paper.

Acknowledgements

The author is truly grateful for the advice received from P.O. Lindberg (LiTH), and also to C. Lemaréchal (INRIA) who suggested many improvements to earlier versions of this manuscript.

References

[1] S. Agmon, "The relaxation method for linear inequalities," *Canadian Journal of Mathematics* 6 (1954) 282-292.

[2] U. Brännlund, "A convergent subgradient method based on the relaxation step," in: U. Brännlund, "On relaxation methods for nonsmooth optimization," Ph.D. Thesis, Department of Mathematics, Kungliga Tekniska Högskolan (Stockholm, 1993).

[3] U. Brännlund, K.C. Kiwiel and P.O. Lindberg, "A descent proximal level bundle method for convex nondifferentiable optimization," *Operations Research Letters* 17 (3) (1995) 121-126.

[4] P.M. Camerini, L. Fratta and F. Maffioli, "On improving relaxation methods by modified gradient techniques," *Mathematical Programming Study* 3 (1975) 26-34.

[5] J.L. Goffin, "Nondifferentiable optimization and the relaxation method," in: C.L. Lemaréchal and R. Mifflin, eds., *Nonsmooth Optimization* (Pergamon, Oxford, 1977) pp. 31-49.

[6] S. Kim and H. Ahn, "Convergence of a generalized subgradient method for nondifferentiable convex optimization," *Mathematical Programming* 50 (1) (1991) 75-80.

[7] K.C. Kiwiel, "An aggregate subgradient method for nonsmooth convex minimization," *Mathematical Programming* 27 (3) (1983) 320-341.

[8] K.C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization* (Springer, Berlin, 1985).

[9] K.C. Kiwiel, "The efficiency of subgradient projection methods for convex nondifferentiable optimization, part II: implementations and extensions," *SIAM Journal on Control and Optimization*, to appear.

[10] C. Lemaréchal, "Nondifferentiable optimization," in: G.L. Nemhauser, A.H.G. Rinnooy Kan and M.J. Todd, eds., *Optimization*, Handbooks in Operations Research and Management Science, Vol. 1 (North-Holland, Amsterdam, 1989) pp. 529-572.

[11] C. Lemaréchal, A. Nemirovskii and Yu. Nesterov, "New variants of bundle methods," *Mathematical Programming* 69 (1) (1995) 111-147.

[12] M. Minoux, *Mathematical Programming, Theory and Algorithms* (Wiley, New York, 1986).

[13] T. Motzkin and I.J. Schoenberg, "The relaxation method for linear inequalities," *Canadian Journal of Mathematics* 6 (1954) 393–404.

[14] B.T. Polyak, "Minimization of unsmooth functionals," *USSR Computational Mathematics and Mathematical Physics* 9 (1969) 14–29.

[15] S. Schaible, "Fractional programming. I, duality," *Management Science* 22 (1976) 858–867.

[16] N.Z. Shor, *Minimization Methods for Non-Differentiable Functions* (Springer, Berlin, 1985).