Simplified Versions of the Conditional Gradient Method

I.V. Konnov[1]

Department of System Analysis and Information Technologies, Kazan Federal University, ul. Kremlevskaya, 18, Kazan 420008, Russia.

Abstract

We suggest simple modifications of the conditional gradient method for smooth optimization problems, which maintain the basic convergence properties, but reduce the implementation cost of each iteration essentially. Namely, we propose an adaptive step-size procedure without any line-search and inexact solution of the direction finding subproblem. Preliminary results of computational tests confirm efficiency of the proposed modifications.

Key words: Optimization problems; pseudo-convex function; conditional gradient method; simple adaptive step-size; inexact direction finding subproblem; convergence properties.

MSC codes: 90C25, 90C30

[1]E-mail: konn-igor@ya.ru

1 Introduction

Let $f : D \to \mathbb{R}$ be a function defined on some set D in the real n-dimensional space $\mathbb{R}^n$. Then one can define the usual optimization problem of finding the minimal value of the function f over the feasible set D. For brevity, we write this problem as

$$\min_{x \in D} \to f(x), \tag{1}$$

its solution set is denoted by D^* and the optimal value of the function by f^*, i.e.

$$f^* = \inf_{x \in D} f(x).$$

Let us first consider the well known class of smooth convex optimization problems, where the set D is supposed to be convex and closed and the function f is supposed to be convex and smooth. This class of problems is one of the most investigated and many iterative methods were proposed for their solution. During rather long time, the efforts were concentrated on developing more powerful and rapidly convergent methods, such as Newton and interior point type ones, which admit complex transformations at each iteration, but attain high accuracy of approximations. However, new significant areas of applications related to data mining and processing as well as allocation decisions in information and telecommunication networks and related systems, where large dimensionality and inexact data together with congestion effects and scattered necessary information are typical, force us to utilize methods whose iteration computation expenses and accuracy requirements are rather low, i.e., they do not utilize matrix transformations at all. Therefore, the well known first or even zero order methods with comparatively slow convergence may appear very useful here.

Let us turn to the conditional gradient method (CGM for short), which is one of the oldest methods applied to the above problem. It was first suggested in [1] for the case when the goal function is quadratic and further was developed by many authors; see e.g. [2, 3, 4, 5]. We recall that the main idea of this method consists in linearization of the goal function. That is, given the current iterate $x^k \in D$, one finds some solution y^k of the problem

$$\min_{y \in D} \to \langle f'(x^k), y \rangle \tag{2}$$

and defines $p^k = y^k - x^k$ as a descent direction at x^k. Taking a suitable step-size $\lambda_k \in (0, 1]$, one sets $x^{k+1} = x^k + \lambda_k p^k$ and so on.

During rather long time, this method was not considered as very efficient due to its relatively slow convergence, but it also became very popular recently. In fact, its auxiliary linearized problems of form (2) appear simpler essentially than the quadratic ones of the most other methods. Next, it usually yields so-called sparse approximations of a solution with few non-zero components; see e.g. [6, 7]. These properties are very significant for the new applications indicated above. We observe that many efforts were directed to enhance the usual (CGM); see e.g. [8, 9, 7, 10, 11] and the references therein. The most popular way consists in developing versions that attain more rapid convergence. At the same time, we can create more efficient methods via reduction of the implementation costs at each iteration with preserving all the useful properties of the initial method.

In this paper, we will follow the second way. First of all, being based on the approach in [12], we suggest a new adaptive step-size procedure in (CGM) without any line-search. Our new step-size procedure admits different changes of the step-size and wide variety of implementation rules, not only decrease as in [12]. It does not utilize a priori information such as Lipschitz constants and does not insist on the strict descent at each iteration, but takes into account behavior of the iteration sequence, unlike the well known divergent series rule. Moreover, the Lipschitz continuity of the gradient of the goal function is not necessary for its convergence. Afterwards, we introduce special threshold control and tolerances in order to avoid exact solution of the direction finding subproblem (2) at each iteration. We establish a complexity estimate for this method, which appears equivalent to the convergence rate of the custom (CGM). Furthermore, we propose a version that combines both the modifications and show that it possesses strengthened convergence properties with respect to the first modification of (CGM) without line-search. Preliminary results of computational experiments confirm efficiency of all the proposed modifications.

2 Properties of the usual conditional gradient method

We will take the following set of basic assumptions for problem (1).

(H) *D is a nonempty, convex, closed, and bounded subset of $\mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function on the set D.*

Together with problem (1), we will consider the following variational inequality (VI for short): Find a point $x^* \in D$ such that

$$\langle f'(x^*), x - x^* \rangle \geq 0 \quad \forall x \in D. \tag{3}$$

We denote by D^0 the solution set of VI (3).

We recall that a differentiable function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is called *pseudo-convex* on a set $D \subseteq \mathbb{R}^n$ if for each pair of points $x, y \in D$ we have

$$\langle \varphi'(x), y - x \rangle \geq 0 \implies \varphi(y) \geq \varphi(x).$$

It is well known that that the class of convex functions is strictly contained in the class of pseudo-convex functions. For instance, the function $\ln t$ is concave and pseudo-convex on $\mathbb{R}_> = \{t \mid t > 0\}$, but it is clearly non-convex. VI (3) can be used as an optimality condition for problem (1) so that solutions of VI (3) are called *stationary points* of (1).

Lemma 1 *[13, Theorems 5.5 and 9.12] Let* **(H)** *hold.*

(i) Each solution of problem (1) solves VI (3).
(ii) If f is pseudo-convex, then each solution of VI (3) solves problem (1).

The boundedness of D guarantees that problem (1) has a solution, moreover, D^* is a compact set, which will be also convex if f is pseudo-convex. Given a point $x \in D$ we define the auxiliary problem

$$\min_{y \in D} \to \langle f'(x), y \rangle. \tag{4}$$

We denote by $Z(x)$ the solution set of problem (4), thus defining the set-valued mapping $x \mapsto Z(x)$. Observe that the set $Z(x)$ is always non-empty, convex, and compact. Also, let

$$\mu(x) = \max_{y \in D} \langle f'(x), x - y \rangle.$$

Given a set $V \subseteq \mathbb{R}^n$, a set-valued mapping $u \mapsto Q(u)$ is said to be *closed* on a set V, if for each pair of sequences $\{u^k\} \to u$, $\{q^k\} \to q$ such that $u^k \in V$ and $q^k \in Q(u^k)$, we have $q \in Q(u)$.

Lemma 2 *[13, Lemma 6.3] Let* **(H)** *hold. Then the mapping $x \mapsto Z(x)$ is closed on D.*

Lemma 3 *[13, Lemma 6.4] Let* **(H)** *hold. Then the following assertions are equivalent:*

(i) $x^* \in D^0$;

(ii) $x^* \in Z(x^*)$;

(iii) $\langle f'(x^*), x^* - z^* \rangle = 0$ *for* $z^* \in Z(x^*)$ *and* $x^* \in D$.

The above properties are very useful for substantiation of various (CGM) type methods. Following [4], we now describe the usual (CGM) with the Armijo step-size rule for a more clear comparison with the new methods. Here and below, $\mathbb{Z}_+$ denotes the set of non-negative integers.

Method (CGM).
Step 0: Choose a point $x^0 \in D$, numbers $\beta \in (0,1)$ and $\theta \in (0,1)$. Set $k = 0$.
Step 1: Find a point $y^k \in Z(x^k)$, set $d^k = y^k - x^k$. If $\langle f'(x^k), d^k \rangle = 0$, stop.
Step 2: Determine m as the smallest number in $\mathbb{Z}_+$ such that

$$f(x^k + \theta^m d^k) \le f(x^k) + \beta \theta^m \langle f'(x^k), d^k \rangle, \tag{5}$$

set $\lambda_k = \theta^m$, $x^{k+1} = x^k + \lambda_k d^k$, $k = k + 1$, and go to Step 1.

Clearly, termination of the method yields a point of D^0. For this reason, we will consider only the non-trivial case where the sequence $\{x^k\}$ is infinite. We give the basic convergence result of the above method.

Proposition 1 *(e.g. [13, Theorems 6.12 and 9.12]) Let* **(H)** *hold, the sequence* $\{x^k\}$ *be generated by (CGM). Then:*

(i) The linesearch procedure in Step 2 is always finite.

(ii) The sequence $\{x^k\}$ *has limit points, all these limit points belong to the set* D^0.

(iii) If f *is pseudo-convex, then all the limit points of the sequence* $\{x^k\}$ *belong to the set* D^*, *besides,*

$$\lim_{k\to\infty} f(x^k) = f^*. \tag{6}$$

We can in principle take the exact one-dimensional minimization rule instead of the current Armijo rule in (5), but it is not so suitable for implementation. Next, if the gradient of the function f is Lipschitz continuous on D with some constant $L > 0$, i.e., $\|f'(y) - f'(x)\| \leq L\|y - x\|$ for any vectors x and y, one can give bounds for the step-size and obtain the convergence rate.

Proposition 2 *([2, Theorem 6.1] and [3, Chapter III, Theorem 1.7]) Suppose that the assumptions in* **(H)** *are fulfilled, the function* f *is convex, the gradient of the function* f *is Lipschitz continuous on* D *with some constant* $L > 0$, *a sequence* $\{x^k\}$ *is generated by (CGM) where the step-size* λ_k *is chosen by the formula*

$$\lambda_k = \min\{1, \theta_k \sigma_k\}, \sigma_k = -\langle f'(x^k), d^k\rangle / \|d^k\|^2,\ \theta_k \in [\theta', \theta''],\ \theta' > 0, \theta'' < 2/L.$$

Then these exists some constant $C < +\infty$ *such that*

$$f(x^k) - f^* \leq C/k \quad \textit{for } k = 0, 1, \ldots \tag{7}$$

This version reduces the computational expenses essentially due to the absence of the line-search, but requires the evaluation of the Lipschitz constant. However, utilization of its inexact estimates usually leads to slow convergence. This is also the case for the known divergent series rule (see e.g. [14])

$$\sum_{k=0}^{\infty} \lambda_k = \infty,\ \sum_{k=0}^{\infty} \lambda_k^2 < \infty,\ \lambda_k \in (0, 1),\ k = 0, 1, 2, \ldots,$$

and for similar rules, which do not evaluate the information about the problem along the current iterates.

3 A simple adaptive step-size without line-search

We now describe a modification of the (CGM), which involves a simple adaptive step-size procedure without line-search. Moreover, it does not require any a priori information about the problem.

Method (CGMS).

Step 0: Choose a point $x^0 \in D$, numbers $\beta \in (0,1)$ and a sequence $\{\tau_l\} \to 0$, $\tau_0 \in (0,1)$. Set $k = 0$, $l = 0$, choose a number $\lambda_0 \in (0, \tau_0]$.

Step 1: Find a point $y^k \in Z(x^k)$, set $d^k = y^k - x^k$. If $\langle f'(x^k), d^k \rangle = 0$, stop.

Step 2: Set $x^{k+1} = x^k + \lambda_k d^k$. If

$$f(x^{k+1}) \leq f(x^k) + \beta \lambda_k \langle f'(x^k), d^k \rangle, \tag{8}$$

take $\lambda_{k+1} \in [\lambda_k, \tau_l]$. Otherwise set $\lambda'_{k+1} = \min\{\lambda_k, \tau_{l+1}\}$, $l = l + 1$ and take $\lambda_{k+1} \in (0, \lambda'_{k+1}]$. Set $k = k + 1$ and go to Step 1.

Again, termination of the method yields a point of D^0 due to Lemma 3. Hence, we will consider only the case where the sequence $\{x^k\}$ is infinite.

Theorem 1 *Let the assumptions in* **(H)** *be fulfilled. Then:*

(i) The sequence $\{x^k\}$ has a limit point, which belongs to the set D^0.

(ii) If f is pseudo-convex, then all the limit points of the sequence $\{x^k\}$ belong to the set D^*, besides, (6) holds.

Proof. First we note that both the sequences $\{x^k\}$ and $\{y^k\}$ belong to the bounded set D and hence have limit points. Let us consider two possible cases.

Case 1: The number of changes of the index l is finite.
Then we have $\lambda_k \geq \bar{\lambda} > 0$ for k large enough, hence (8) gives

$$f(x^{k+1}) \leq f(x^k) + \beta \bar{\lambda} \langle f'(x^k), d^k \rangle$$

for k large enough. Since $f(x^k) \geq f^* > -\infty$, we must have

$$\lim_{k \to \infty} f(x^k) = \mu \tag{9}$$

and

$$\lim_{k \to \infty} \langle f'(x^k), d^k \rangle = 0. \tag{10}$$

Let x' be an arbitrary limit point of the sequence $\{x^k\}$. Taking a subsequence if necessary we have the corresponding limit point y' of the sequence $\{y^k\}$, i. e.

$$\lim_{s \to \infty} x^{k_s} = x' \text{ and } \lim_{s \to \infty} y^{k_s} = y'.$$

From (10) we now have

$$\langle f'(x'), y' - x' \rangle = 0,$$

but the mapping $x \mapsto Z(x)$ is closed due to Lemma 2, hence $y' \in Z(x')$. From Lemma 3 it follows that $x' \in D^0$. Hence, in this case all the limit points of the sequence $\{x^k\}$ belong to the set D^0. Therefore, assertion (i) is true. If f is pseudo-convex, then $D^0 = D^*$ due to Lemma 1, which gives $\mu = f^*$ in (9) and (6). We conclude that assertion (ii) is also true.

Case 2: The number of changes of the index l is infinite.
Then there exists an infinite subsequence of indices $\{k_l\}$ such that

$$f(x^{k_l} + \lambda_{k_l} d^{k_l}) - f(x^{k_l}) = f(x^{k_l+1}) - f(x^{k_l}) > \beta\lambda_{k_l}\langle f'(x^{k_l}), d^{k_l}\rangle,$$

or equivalently,

$$\frac{f(x^{k_l} + \lambda_{k_l} d^{k_l}) - f(x^{k_l})}{\lambda_{k_l}} > \beta\langle f'(x^{k_l}), d^{k_l}\rangle; \tag{11}$$

besides,

$$\lambda_{k_l} \in (0, \tau_l], \ \ \lambda_{k_l+1} \in (0, \tau_{l+1}],$$

and

$$\lim_{s\to\infty} \tau_l = 0.$$

Let $\bar{x}$ be an arbitrary limit point of this subsequence $\{x^{k_l}\}$. Taking a subsequence if necessary we can choose the corresponding limit point $\bar{y}$ of the subsequence $\{y^{k_l}\}$. Without loss of generality we can suppose that

$$\lim_{l\to\infty} x^{k_l} = \bar{x} \text{ and } \lim_{l\to\infty} y^{k_l} = \bar{y}.$$

Since $\lambda_{k_l} \to 0$ as $l \to +\infty$, taking the limit $l \to +\infty$ in relation (11) we obtain

$$\langle f'(\bar{x}), \bar{y} - \bar{x}\rangle \geq \beta\langle f'(\bar{x}), \bar{y} - \bar{x}\rangle,$$

i.e.

$$\langle f'(\bar{x}), \bar{y} - \bar{x}\rangle \geq 0.$$

By Lemma 2 we have $\bar{y} \in Z(\bar{x})$, but from Lemma 3 it now follows that $\bar{x} \in D^0$. Therefore, assertion (i) is true. If f is pseudo-convex, then $D^0 = D^*$ due to Lemma 1. It follows that all these limit points of the subsequence $\{x^{k_l}\}$ belong to the set D^*. Since $x^{k_l+1} = x^{k_l} + \lambda_{k_l} d^{k_l}$, $\lambda_{k_l} \to 0$, and the sequence $\{d^{k_l}\}$ is bounded, the limit points of the subsequences $\{x^{k_l}\}$ and $\{x^{k_l+1}\}$ coincide and all they belong to the set D^*.

For any index k we define the index $m(k)$ as follows:

$$m(k) = \max\{j \mid j \leq k, \ f(x^j) - f(x^{j-1}) > \beta\lambda_{j-1}\langle f'(x^{j-1}), d^{j-1}\rangle\},$$

i. e. j is the closest to k but not greater index from the subsequence $\{x^{k_l+1}\}$. This means that $j = k$ if $f(x^k) - f(x^{k-1}) > \beta\lambda_{k-1}\langle f'(x^{k-1}), d^{k-1}\rangle$. By definition, we have

$$f(x^k) \leq f(x^{m(k)}). \tag{12}$$

Let now x' be an arbitrary limit point of the sequence $\{x^k\}$, i.e. $\lim_{s\to\infty} x^{t_s} = x'$. Create the corresponding infinite subsequence $\{x^{m(t_s)}\}$. From (12) we have $f^* \leq f(x^{t_s}) \leq f(x^{m(t_s)})$, but all the limit points of the sequence $\{x^{m(t_s)}\}$ belong to the set D^* since it is contained in the sequence $\{x^{k_l+1}\}$. Choose any limit point x'' of $\{x^{m(t_s)}\}$. Then, taking a subsequence if necessary we obtain

$$f^* \leq f(x') \leq f(x'') = f^*.$$

therefore $x' \in D^*$. This means that all the limit points of the sequence $\{x^k\}$ belong to the set D^* and that (6) holds true. We conclude that assertion (ii) is also true. □

It should be observed that (CGMS) follows the approach in [12], but the step-size procedure in (CGMS) admits wide variety of implementation rules in comparison with those in [12], where only the strict decrease is indicated for possible changes of the step-size. Even the simplest implementation rule of (CGMS), where $\lambda_{k+1} = \max\{\lambda_k, \tau_l\}$ if (8) holds and $\lambda_{k+1} = \min\{\lambda_k, \tau_{l+1}\}$ otherwise, admits the increase of λ_{k+1}, which prevents from the too small step-size. Such a modification seems especially significant for the case where the computation of the goal function value is rather expensive.

4 Inexact solution of the direction finding subproblem

It was noticed in Section 1 that the auxiliary direction finding subproblem (2) in (CGM) is simpler essentially than the quadratic ones in the projection based methods. Nevertheless, its exact solution may also be expensive. If the feasible set D is a general polyhedron with many vertices, one has to apply a special algorithm at each iteration. Then, the method with approximate solution of subproblem (2) may appear more efficient. There exist several versions of such methods; see e.g. [3, 14]. We observe that all these versions involve evaluation of the accuracy of a solution of subproblem (2), which must tend to zero. In this section, we intend to present some other version of this modification of (CGM), which is based on inserting tolerances and some threshold control of the descent property. We observe that this approach was first suggested for the bi-coordinate descent method in [15]. In [16], it was applied in a generalized conditional gradient method for optimization problems on Cartesian product sets, where the corresponding partial auxiliary problems in subspaces are still to be solved exactly. We now describe the general inexact (CGM) with the same Armijo step-size rule.

Method (CGMI).
Initialization: Choose a point $w^0 \in D$, numbers $\beta \in (0,1)$, $\theta \in (0,1)$, and a positive sequence $\{\delta_p\} \to 0$. Set $p = 1$.

Step 0: Set $k = 0$, $x^0 = w^{p-1}$.

Step 1: Find a point $z^k \in D$ such that

$$\langle f'(x^k), x^k - z^k \rangle \geq \delta_p. \tag{13}$$

If $\mu(x^k) < \delta_p$, set $w^p = x^k$, $p = p + 1$ and go to Step 0. *(Restart)*

Step 2: Set $d^k = z^k - x^k$, determine m as the smallest number in $\mathbb{Z}_+$ such that

$$f(x^k + \theta^m d^k) \le f(x^k) + \beta\theta^m \langle f'(x^k), d^k\rangle, \tag{14}$$

set $\lambda_k = \theta^m$, $x^{k+1} = x^k + \lambda_k d^k$, $k = k+1$ and go to Step 1.

Thus, the method has a two-level structure where each outer iteration (stage) p contains some number of inner iterations in k with the fixed tolerance δ_p. Completing each stage, that is marked as restart, leads to decrease of its value. Observe that only the restart situation requires the exact solution of the auxiliary subproblem (2). In all the other cases, we can take $z^k \in D$ as an arbitrary suitable point (say a vertex of D) within condition (13).

By (13), we have

$$\langle f'(x^k), d^k\rangle \le -\delta_p < 0$$

in (14). It follows that

$$f(x^{k+1}) - f(x^k) \le \beta\lambda_k \langle f'(x^k), d^k\rangle \le -\beta\lambda_k\delta_p. \tag{15}$$

We first justify the linesearch.

Lemma 4 *Let the assumptions in* **(H)** *be fulfilled. Then the linesearch procedure in Step 2 of (CGMI) is always finite.*

Proof. If we suppose that the linesearch procedure is infinite, then (14) does not hold and

$$\theta^{-m}(f(x^k + \theta^m d^k) - f(x^k)) > \beta\langle f'(x^k), d^k\rangle,$$

for $m \to \infty$. Hence, by taking the limit we have $\langle f'(x^k), d^k\rangle \ge \beta\langle f'(x^k), d^k\rangle$, hence $\langle f'(x^k), d^k\rangle \ge 0$, a contradiction with (13). □

We show that each stage is well defined.

Proposition 3 *Let the assumptions in* **(H)** *be fulfilled. Then the number of iterations at each stage p is finite.*

Proof. Fix any p and suppose that the sequence $\{x^k\}$ is infinite. By (15), we have $f^* \le f(x^k)$ and $f(x^{k+1}) \le f(x^k) - \beta\delta_p\lambda_k$, hence

$$\lim_{k\to\infty} \lambda_k = 0.$$

Both the sequences $\{x^k\}$ and $\{z^k\}$ belong to the bounded set D and hence have limit points. Without loss of generality, we can suppose that some subsequence $\{x^{k_s}\}$ converges to a point $\bar{x}$ and the corresponding subsequence $\{z^{k_s}\}$ converges to a point $\bar{z}$. Due to (13) we have

$$\langle f'(\bar{x}), \bar{y} - \bar{x}\rangle = \lim_{s\to\infty} \langle f'(x^{k_s}), y^{k_s} - x^{k_s}\rangle \le -\delta_p. \tag{16}$$

However, (14) does not hold for the step-size λ_k/θ. Setting $k = k_s$ gives

$$(\lambda_{k_s}/\theta)^{-1}(f(x^{k_s} + (\lambda_{k_s}/\theta)d^{k_s}) - f(x^{k_s})) > \beta\langle f'(x^{k_s}), d^{k_s}\rangle,$$

hence, by taking the limit $s \to \infty$ we obtain

$$\begin{aligned}\langle f'(\bar{x}), \bar{y} - \bar{x}\rangle &= \lim_{s\to\infty}\left\{(\lambda_{k_s}/\theta)^{-1}(f(x^{k_s} + (\lambda_{k_s}/\theta)d^{k_s}) - f(x^{k_s}))\right\}\\ &\geq \beta\langle f'(\bar{x}), \bar{y} - \bar{x}\rangle,\end{aligned}$$

i.e., $(1-\beta)\langle f'(\bar{x}), \bar{y} - \bar{x}\rangle \geq 0$, which is a contradiction with (16). □

We are ready to prove convergence of the whole method.

Theorem 2 *Let the assumptions in* **(H)** *be fulfilled. Then:*

(i) The number of changes of index k at each stage p is finite.

(ii) The sequence $\{w^p\}$ generated by method (CGMI) has limit points, all these limit points belong to D^0.

(iii) If f is pseudo-convex, then all the limit points of the sequence $\{w^p\}$ belong to the set D^*, besides,

$$\lim_{p\to\infty} f(w^p) = f^*; \tag{17}$$

Proof. Assertion (i) has been obtained in Proposition 3. By construction, the sequence $\{w^p\}$ is bounded, hence it has limit points. Moreover, $f^* \leq f(w^{p+1}) \leq f(w^p)$, hence

$$\lim_{p\to\infty} f(w^p) = \mu. \tag{18}$$

For each p and any point $u^p \in Z(w^p)$ it holds that

$$\langle f'(w^p), w^p - u^p\rangle \leq \delta_p. \tag{19}$$

Fix this sequence $\{u^p\}$. It is also bounded and must have limit points. Take an arbitrary limit point $\bar{w}$ of $\{w^p\}$. Then, without loss of generality we can suppose that

$$\bar{u} = \lim_{t\to\infty} u^{p_t} \text{ and } \bar{w} = \lim_{t\to\infty} w^{p_t},$$

for some subsequences $\{u^{p_t}\}$ and $\{w^{p_t}\}$. Taking the limit $t \to \infty$ in (19) with $p = p_t$, we obtain

$$\langle f'(\bar{w}), \bar{w} - \bar{u}\rangle = \lim_{t\to\infty}\langle f'(w^{p_t}), w^{p_t} - u^{p_t}\rangle \leq 0.$$

By Lemma 2 we have $\bar{u} \in Z(\bar{w})$, hence $\langle f'(\bar{w}), \bar{w} - \bar{u}\rangle = 0$ and $\bar{w} \in D^0$ due to Lemma 3. This means that all the limit points of $\{u^p\}$ belong to D^0. This gives assertion (ii). If f is pseudo-convex, then $D^0 = D^*$ due to Lemma 1, which gives $\mu = f^*$ in (18) and (17). We conclude that assertion (iii) is true. □

It was observed in Section 2 that the usual (CGM) attains the convergence rate $O(1/k)$ under the additional assumptions that the function f is convex and its gradient is Lipschitz continuous; see Proposition 2 and formula (7). This means that the total number of iterations $N(\varepsilon)$ that is necessary for attaining some prescribed accuracy $\varepsilon > 0$ for the gap value $\Delta(x) = f(x) - f^*$ is estimated as follows:

$$N(\varepsilon) \leq C/\varepsilon, \text{ where } 0 < C < \infty. \tag{20}$$

We can try to obtain a similar estimate for (CGMI) with the proper specialization. In fact, if the gradient of the function f is Lipschitz continuous on D with some constant $L > 0$, we can take the well known property of such functions

$$f(y) \leq f(x) + \langle f'(x), y - x\rangle + 0.5L\|y - x\|^2;$$

see [3, Chapter III, Lemma 1.2]. Then, at Step 2 we have

$$f(x^k + \lambda d^k) - f(x^k) \leq \lambda[\langle f'(x^k), d^k\rangle + 0.5L\lambda\|d^k\|^2] \leq \beta\lambda\langle f'(x^k), d^k\rangle,$$

if $\lambda \leq -2(1 - \beta)\langle f'(x^k), d^k\rangle/(L\|d^k\|^2)$. Next, $\langle f'(x^k), d^k\rangle \leq -\delta_p$ at stage p, besides, $\|d^k\| \leq \rho \triangleq \mathrm{Diam}D < \infty$. If we simply take $\lambda_k = \lambda\delta_l$ with $\lambda \in (0, \bar{\lambda}]$ and

$$\bar{\lambda} = 2(1 - \beta)/(L\rho^2),$$

then

$$f(x^k + \lambda_k d^k) \leq f(x^k) + \beta\lambda_k\langle f'(x^k), d^k\rangle, \tag{21}$$

as desired; cf. (14). This means that we can drop the line-search procedure in Step 2. We call this modification (CGMIL). Obviously, the assertions of Proposition 3 and Theorem 2 remain true for this version.

As (CGMIL) has a two-level structure with each stage containing a finite number of inner iterations, it is more suitable to derive its complexity estimate, which gives the total amount of work of the method. Given a starting point w^0 and a number $\varepsilon > 0$, we define the complexity of the method, denoted by $N(\varepsilon)$, as the total number of inner iterations at $p(\varepsilon)$ stages such that $p(\varepsilon)$ is the maximal number p with $\Delta(w^p) \geq \varepsilon$, hence,

$$N(\varepsilon) \leq \sum_{p=1}^{p(\varepsilon)} N_p, \tag{22}$$

where N_p denotes the total number of iterations at stage p. We have to estimate the right-hand side of (22).

Theorem 3 *Let a sequence $\{w^l\}$ be generated by (CGMIL) with the rule:*

$$\delta_p = \nu^p\delta_0, \; p = 0, 1, \ldots; \quad \nu \in (0, 1), \delta_0 > 0. \tag{23}$$

Suppose that the assumptions in **(H)** *be fulfilled and also that the function f is convex and its gradient is Lipschitz continuous with constant L. Then the method has the complexity estimate*

$$N(\varepsilon) \leq C_1\nu((\delta_0/\varepsilon) - 1)/(1 - \nu),$$

where $C_1 = \rho^2 L/(2\beta(1 - \beta)\delta_0)$.

Proof. From (21) and (13) we have

$$f(x^{k+1}) \leq f(x^k) - \beta\lambda_k\delta_p = f(x^k) - \beta\bar{\lambda}\delta_p^2,$$

at any fixed stage p. It follows from the definition of $\bar{\lambda}$ that

$$N_p \leq (f(w^{p-1}) - f^*)/(\beta\bar{\lambda}\delta_p^2) \leq \rho^2 L\Delta(w^{p-1})/(2\beta(1 - \beta)\delta_p^2). \tag{24}$$

By the convexity of f, for some $x^* \in D^*$ we have

$$\Delta(w^p) = f(w^p) - f(x^*) \leq \langle f'(w^p), w^p - x^*\rangle \leq \delta_p.$$

Using this estimate in (24) gives

$$N_p \leq \rho^2 L\delta_{p-1}/(2\beta(1 - \beta)\delta_p^2).$$

From (23) it now follows that

$$N_p \leq \rho^2 L\nu^{-p}/(2\beta(1 - \beta)\delta_0\nu) = C_1\nu^{-p-1}.$$

On the other side, since $\varepsilon \leq \Delta(w^p) \leq \delta_p = \delta_0\nu^p$, we have

$$\nu^{-p(\varepsilon)} \leq \delta_0/\varepsilon.$$

Combining both the inequalities in (22), we obtain

$$\begin{aligned} N(\varepsilon) \quad &\leq C_1\sum_{p=1}^{p(\varepsilon)}\nu^{-p-1} = C_1\nu(\nu^{-p(\varepsilon)} - 1)/(1 - \nu) \\ &\leq C_1\nu((\delta_0/\varepsilon) - 1)/(1 - \nu). \end{aligned}$$

□

We observe that the above estimate is the same as in (20), which corresponds to the usual (CGM) under the same assumptions.

5 A parametric inexact method without line-search

In this section, we describe the combined method, which involves both the inexact solution of the auxiliary direction finding subproblem (2) due to special parametric threshold control of the descent property and the simple adaptive step-size rule without line-search.

Method (CGMIS).
Initialization: Choose a point $w^0 \in D$, numbers $\beta \in (0,1)$, $\theta \in (0,1)$, and a positive sequence $\{\delta_p\} \to 0$. Set $p = 1$.

Step 0: Choose a sequence $\{\tau_{l,p}\} \to 0$, $\tau_{0,p} \in (0,1)$. Set $k = 0$, $l = 0$, $x^0 = w^{p-1}$, choose a number $\lambda_0 \in (0, \tau_{0,p}]$.

Step 1: Find a point $z^k \in D$ such that

$$\langle f'(x^k), x^k - z^k \rangle \geq \delta_p. \tag{25}$$

If $\mu(x^k) < \delta_p$, set $w^p = x^k$, $p = p+1$ and go to Step 0. *(Restart)*

Step 2: Set $d^k = z^k - x^k$, $x^{k+1} = x^k + \lambda_k d^k$. If

$$f(x^{k+1}) \leq f(x^k) + \beta\lambda_k \langle f'(x^k), d^k \rangle, \tag{26}$$

take $\lambda_{k+1} \in [\lambda_k, \tau_{l,p}]$. Otherwise set $\lambda'_{k+1} = \min\{\lambda_k, \tau_{l+1,p}\}$, $l = l+1$ and take $\lambda_{k+1} \in (0, \lambda'_{k+1}]$. Set $k = k+1$ and go to Step 1.

Again, each outer iteration (stage) p contains some number of inner iterations in k with the fixed tolerance δ_p. Completing each stage, that is marked as restart, leads to decrease of its value. Note that the choice of the parameters $\{\tau_{l,p}\}$ can be in principle independent for each stage p.

By (25), we again have

$$\langle f'(x^k), d^k \rangle \leq -\delta_p < 0$$

in (26). It follows that

$$f(x^{k+1}) - f(x^k) \leq \beta\lambda_k \langle f'(x^k), d^k \rangle \leq -\beta\lambda_k\delta_p. \tag{27}$$

We show that each stage is well defined.

Proposition 4 *Let the assumptions in* **(H)** *be fulfilled. Then the number of iterations at each stage p is finite.*

Proof. Fix any p and suppose that the sequence $\{x^k\}$ is infinite. Then the number of changes of index l is also infinite. In fact, otherwise we have $\lambda_k \geq \bar{\lambda} > 0$ for k large enough, hence (27) gives

$$f^* \leq f(x^{k+t}) \leq f(x^k) - t\beta\bar{\lambda}\delta_p \;\to -\infty \text{ as } t \to \infty,$$

for k large enough, which is a contradiction. Therefore, there exists an infinite subsequence of indices $\{k_l\}$ such that

$$f(x^{k_l} + \lambda_{k_l} d^{k_l}) - f(x^{k_l}) = f(x^{k_l+1}) - f(x^{k_l}) > \beta \lambda_{k_l} \langle f'(x^{k_l}), d^{k_l} \rangle,$$

or equivalently,

$$\frac{f(x^{k_l} + \lambda_{k_l} d^{k_l}) - f(x^{k_l})}{\lambda_{k_l}} > \beta \langle f'(x^{k_l}), d^{k_l} \rangle, \tag{28}$$

where $d^{k_l} = z^{k_l} - x^{k_l}$. Besides, it holds that

$$\lambda_{k_l} \in (0, \tau_{l,p}], \ \ \lambda_{k_l+1} \in (0, \tau_{l+1,p}],$$

where

$$\lim_{l \to \infty} \tau_{l,p} = 0.$$

Both the sequences $\{x^k\}$ and $\{z^k\}$ belong to the bounded set D and hence have limit points. Without loss of generality, we can suppose that the subsequence $\{x^{k_l}\}$ converges to a point $\bar{x}$ and the corresponding subsequence $\{z^{k_l}\}$ converges to a point $\bar{z}$. Due to (25) we have

$$\langle f'(\bar{x}), \bar{z} - \bar{x} \rangle = \lim_{l \to \infty} \langle f'(x^{k_l}), z^{k_l} - x^{k_l} \rangle \leq -\delta_p. \tag{29}$$

At the same time, taking the limit $l \to \infty$ in (28), we obtain

$$\begin{aligned} \langle f'(\bar{x}), \bar{z} - \bar{x} \rangle &= \lim_{l \to \infty} \left\{ \lambda_{k_l}^{-1} (f(x^{k_l} + (\lambda_{k_l}/\theta) d^{k_l}) - f(x^{k_l})) \right\} \\ &\geq \beta \langle f'(\bar{x}), \bar{z} - \bar{x} \rangle, \end{aligned}$$

i.e., $(1 - \beta) \langle f'(\bar{x}), \bar{z} - \bar{x} \rangle \geq 0$, which is a contradiction with (29). □

We are ready to prove convergence of the whole method. Although it is similar to Theorem 2, we give the full proof for more clarity.

Theorem 4 *Let the assumptions in* **(H)** *be fulfilled. Then:*

(i) The number of changes of index k at each stage p is finite.

(ii) The sequence $\{w^p\}$ generated by method (CGMIS) has limit points, all these limit points belong to D^0.

(iii) If f is pseudo-convex, then all the limit points of the sequence $\{w^p\}$ belong to the set D^*, besides, (17) holds.

Proof. Assertion (i) has been obtained in Proposition 4. By construction, the sequence $\{w^p\}$ is bounded, hence it has limit points. For each p and any point $u^p \in Z(w^p)$ it holds that

$$\langle f'(w^p), w^p - u^p \rangle \leq \delta_p. \tag{30}$$

Fix this sequence $\{u^p\}$. It is also bounded and must have limit points. Take an arbitrary limit point $\bar{w}$ of $\{w^p\}$. Then, without loss of generality we can suppose that

$$\bar{u} = \lim_{t\to\infty} u^{p_t} \text{ and } \bar{w} = \lim_{t\to\infty} w^{p_t},$$

for some subsequences $\{u^{p_t}\}$ and $\{w^{p_t}\}$. Taking the limit $t \to \infty$ in (30) with $p = p_t$, we obtain

$$\langle f'(\bar{w}), \bar{w} - \bar{u}\rangle = \lim_{t\to\infty} \langle f'(w^{p_t}), w^{p_t} - u^{p_t}\rangle \leq 0.$$

By Lemma 2 we have $\bar{u} \in Z(\bar{w})$, hence $\langle f'(\bar{w}), \bar{w} - \bar{u}\rangle = 0$ and $\bar{w} \in D^0$ due to Lemma 3. This means that all the limit points of $\{u^p\}$ belong to D^0. This gives assertion (ii). If f is pseudo-convex, then $D^0 = D^*$ due to Lemma 1, which gives (17). We conclude that assertion (iii) is true. □

Comparing Theorems 1 and 4, we observe that the joint modifications enable us to attain strengthened convergence properties for (CGMIS) with respect to (CGMS) in the non-convex case.

In this paper, we describe modifications for the basic conditional gradient method. Obviously, the same modifications can be applied to most of the gradient type smooth optimization methods.

6 Computational experiments

In order to check the performance of the proposed methods we carried out computational experiments. We compared (CGM), (CGMS), (CGMI), and (CGMIS) with respect to (1) for different dimensionality. They were implemented in Delphi with double precision arithmetic. The main goal was to compare the number of iterations (it), the total number of calculations of the goal function value (kf), and the total number of calculations of partial derivatives of f (kg) for attaining the same accuracy $\varepsilon = 0.1$ with respect to the gap function $\mu(x)$. We chose the rule $\delta_{p+1} = \nu\delta_p$ with $\nu = 0.5$ for (CGMI) and (CGMIS). For (CGMS) and (CGMIS), we simply set $\lambda_{k+1} = \lambda_k$ if (8) (respectively, (26)) holds, and $\lambda_{k+1} = \sigma\lambda_k$ with $\sigma = 0.9$ otherwise. Next, in the case of restart in (CGMIS) we took $\lambda_0 = \tau_{0,p} = \lambda_k/\sigma$, where λ_k was the current step-size from the previous stage. We set $\beta = \theta = 0.5$ for all the methods.

We took the simplex as the feasible set for all the test problems, i.e.,

$$D = \left\{ x \in \mathbb{R}^n_+ \;\middle|\; \sum_{i=1}^{n} x_i = b \right\}.$$

We set $b = 10$ and took the same starting point $x' = (b/n)e$ for all the methods. For (CGMI) and (CGMIS), we applied the cyclic selection of indices. In all the series, we took the convex cost functions.

Table 1: Quadratic cost function φ_1

	(CGM)			(CGMS)		
n	it	kf	kg	it	kf	kg
5	202	2098	1010	65	65	325
10	713	8256	7130	87	87	870
20	503	6500	10060	833	833	16660
50	1729	24624	86450	2155	2155	107750
100	2540	38454	254000	5430	5430	543000
	(CGMI)			(CGMIS)		
n	it	kf	kg	it	kf	kg
5	199	2072	472	71	71	192
10	743	8690	4098	743	743	4060
20	583	7687	5953	124	124	1296
50	2037	29558	50532	2214	2214	53868
100	2888	44296	14864	6417	6417	21764

In the first series, we chose $f(x) = \varphi_1(x)$ where

$$\varphi_1(x) = 0.5\langle Px, x\rangle, \tag{31}$$

the elements of the matrix P were defined by

$$p_{ij} = \begin{cases} \sin(i)\cos(j) & \text{if } i < j, \\ \sin(j)\cos(i) & \text{if } i > j, \\ \sum\limits_{s\neq i} |p_{is}| + 1 & \text{if } i = j. \end{cases} \tag{32}$$

The results are given in Table 1. In the second series, we took the cost function

$$f(x) = \varphi_1(x) + \varphi_2(x)$$

where the function φ_1 was defined as in (31)–(32) and the function φ_2 was defined by the formula

$$\varphi_2(x) = 1/(\langle c, x\rangle + d), \tag{33}$$

where the elements of the vector c were defined by

$$c_i = 2 + \sin(i) \quad \text{for} \quad i = 1, \dots, n,$$

and $d = 5$. The results are given in Table 2.

In the third series, we chose $f(x) = \varphi_3(x)$ where

$$\varphi_3(x) = 0.5\|Px - q\|^2, \tag{34}$$

Table 2: Convex cost function $\varphi_1 + \varphi_2$

	(CGM)			(CGMS)		
n	it	kf	kg	it	kf	kg
5	203	2116	1015	65	65	325
10	705	8155	7050	87	87	870
20	491	6329	9820	833	833	16660
50	1760	25105	88000	2155	2155	107750
100	2594	39338	259400	5496	5496	549600
	(CGMI)			(CGMIS)		
n	it	kf	kg	it	kf	kg
5	209	2192	517	71	71	192
10	731	8528	4008	706	706	3866
20	547	7150	5660	123	123	1291
50	2026	29359	50551	2070	2070	50506
100	2921	44797	16343	6321	6321	17517

the elements of the $m \times n$ matrix P were defined by

$$p_{ij} = \begin{cases} \tilde{p}_{ij} & \text{if } i \neq j, \\ \tilde{p}_{ij} + 2 & \text{if } i = j; \end{cases} \tag{35}$$

where

$$\tilde{p}_{ij} = \ln(1 + i/j) \sin(i/j)/(i + j), \; i = 1, \dots, m, j = 1, \dots, n; \tag{36}$$

and

$$q_i = b \sum_{j=1}^{n} p_{ij}, \; i = 1, \dots, m. \tag{37}$$

The results are given in Table 3. In the fourth series, we took the cost function

$$f(x) = \varphi_3(x) + \varphi_2(x)$$

where the function φ_3 was defined as in (34)–(37) and the function φ_2 was defined as in (33). The results are given in Table 4.

In almost all the cases, (CGMS) and (CGMIS), which do not use line-search, showed rather rapid convergence, they outperformed (CGM) and (CGMI), respectively, in the total number of goal function calculations. Similarly, the inexact versions (CGMI) and (CGMIS) showed essential reduction of the total number of partial derivatives calculations in comparison with (CGM) and (CGMS), respectively.

Table 3: Quadratic cost function φ_3

		(CGM)			(CGMS)		
m	n	it	kf	kg	it	kf	kg
2	5	2362	25584	11810	68	68	340
5	10	2998	33752	29980	88	88	880
10	20	4076	43381	81520	77	77	1540
25	50	219	2438	10950	497	497	24850
50	100	4197	48491	419700	93	93	9300
		(CGMI)			(CGMIS)		
m	n	it	kf	kg	it	kf	kg
2	5	2412	26273	3617	54	54	103
5	10	3328	38601	9379	107	107	255
10	20	4296	50567	18508	93	93	393
25	50	3924	45769	35894	837	837	7343
50	100	4176	48969	5385	116	116	1774

Table 4: Convex cost function $\varphi_3 + \varphi_2$

		(CGM)			(CGMS)		
m	n	it	kf	kg	it	kf	kg
2	5	2365	25633	11825	68	68	340
5	10	2926	32957	29260	93	93	930
10	20	3972	46047	79440	83	83	1660
25	50	229	2564	11450	497	497	24850
50	100	4192	48894	419200	92	92	9200
		(CGMI)			(CGMIS)		
m	n	it	kf	kg	it	kf	kg
2	5	2386	25956	3577	54	54	103
5	10	3276	37911	9271	107	107	255
10	20	4543	53856	19532	92	92	392
25	50	3806	44284	34835	867	867	7615
50	100	4304	50627	7408	115	115	1763

Acknowledgement

The results of this work were obtained within the state assignment of the Ministry of Science and Education of Russia, project No. 1.460.2016/1.4. In this work, the author was also supported by Russian Foundation for Basic Research, project No. 13-01-00368-a.

References

[1] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Nav. Res. Logist. Quart., vol. 3 (1956), pp.95–110.

[2] E.S. Levitin and B.T. Polyak, *Constrained minimization methods*, USSR Comp. Maths. Math. Phys., vol. 6 (1966), pp.1–50.

[3] V.F. Dem'yanov and A.M. Rubinov, *Approximate Methods for Solving Extremum Problems*, Leningrad Univ. Press, Leningrad, 1968. (Engl. transl. in Elsevier, Amsterdam, 1970)

[4] B.N. Pshenichnyi and Yu.M. Danilin, *Numerical Methods in Extremal Problems*, MIR, Moscow, 1978.

[5] J.C. Dunn, *Convergence rates for conditional gradient sequences generated by implicit step length rules*, SIAM J. Control Optim., vol. 18 (1980), pp.473–487.

[6] K.L. Clarkson, *Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm*, ACM Trans. on Algor., vol. 6 (2010), Art. No. 63, pp.1–30.

[7] M. Jaggi, *Revisiting Frank-Wolfe: Projection-free sparse convex optimization*, Proc. of the 30th International Conference on Machine Learning (ICML-13), 2013, pp.427–435.

[8] J. Guelat and P. Marcotte, *Some comments on Wolfe's "away step"*, Math. Progr., vol. 35 (1986), pp.110–119.

[9] A. Beck and M. Teboulle, *A conditional gradient method with linear rate of convergence for solving convex linear systems*, Math. Meth. Oper. Res., vol. 59 (2004), pp.235–247.

[10] R.M. Freund and P. Grigas, *New analysis and results for the Frank-Wolfe method*, Mathem. Progr., vol. 155 (2016), pp.199–230.

[11] I.V. Konnov, *The method of pairwise variations with tolerances for linearly constrained optimization problems*, J. Nonlin. Variat. Anal., vol. 1 (2017), pp.25–41.

[12] I.V. Konnov, *Conditional gradient method without line-search*, Russ. Mathem. (Izv. VUZ), vol. 62 (2018), pp.82–85.

[13] I.V. Konnov, *Nonlinear Optimization and Variational Inequalities*, Kazan Univ. Press, Kazan, 2013. [In Russian]

[14] J.C. Dunn and S. Harshbarger, *Conditional gradient algorithms with open loop step size rules*, J. Math. Anal. Appl., vol. 62 (1978), pp. 432–444.

[15] I.V. Konnov, *Selective bi-coordinate variations for resource allocation type problems*, Comp. Optim. Appl., vol. 64 (2016), pp.821–842.

[16] I.V. Konnov, *An adaptive partial linearization method for optimization problems on product sets*, J. of Optim. Theory and Appl., vol. 175 (2017), pp.478–501.