Math Meth Oper Res (2004) 59:235–247
DOI 10.1007/s001860300327

Mathematical Methods
of Operations Research

A conditional gradient method with linear rate of convergence for solving convex linear systems

Amir Beck, Marc Teboulle

School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel
(e-mail: becka@post.tau.ac.il, teboulle@post.tau.ac.il)

Manuscript received: June 2002/Final version received: October 2003

Abstract. We consider the problem of finding a point in the intersection of an affine set with a compact convex set, called a convex linear system (CLS). The conditional gradient method is known to exhibit a sublinear rate of convergence. Exploiting the special structure of (CLS), we prove that the conditional gradient method applied to the equivalent minimization formulation of (CLS), converges to a solution at a linear rate, under the sole assumption that Slater's condition holds for (CLS). The rate of convergence is measured explicitly in terms of the problem's data and a Slater point. Application to a class of conic linear systems is discussed.

Key words: Conic linear systems, Slater's condition, conditional gradient, efficiency and rate of convergence analysis

1991 AMS Classification numbers: 90C05, 90C20, 90C25, 90C60, 65K05

1 Introduction

Consider the convex feasibility problem

$$(I)\begin{cases} Mx = g \\ x \in S \end{cases}$$

where $M : \Re^n \to \Re^m$ is a linear map with full row rank, $g \in \Re^m$ is a given point and $S \subset \Re^n$ is a closed and bounded convex set, and its associated equivalent optimization formulation

$$(OP) \quad \min\{||Mx - g||^2 : x \in S\}.$$

The conditional gradient method is a feasible direction method and is applicable only when the feasible set S is compact. At each iteration of the

algorithm, a feasible direction (with respect to the linear approximation of the function) is chosen and then a line search is performed along that direction. The conditional gradient algorithm has been studied by several researchers, see for example, Bertsekas [2], Dunn [4], Levitin and Polyak [7] and references therein. The convergence of CGM can be established under relatively mild assumptions on the problem's data and is in fact an extension of the Frank and Wolfe algorithm [6] originally devised to minimize a quadratic function over a polyhedron. The advantage of the CGM is its simplicity, in particular when applied to problem of the form (OP), (at each iteration, it requires only simple matrix-vector multiplications), yet the efficiency of CGM is far less attractive. Sublinear rate of convergence of the function values was established by [7]. However, the improvement toward the derivation of a linear rate of convergence of the function values has been established only under very restrictive assumptions. Indeed, in [7] linear rate of convergence is proven under the assumptions that the feasible set is strongly (uniformly) convex and that $\|\nabla f(x)\|$ is bounded below by a positive number, which are severe and rarely met assumptions in most optimization models of interest. Other conditions ensuring linear rate of convergence can be found in [4]. Some basic and well known results on the CGM will be summarized in the next section.

Unfortunately, none of these general results are even applicable to the simple problem of minimizing the convex quadratic function over the compact convex set S described in (OP), which is the problem we intend to study. At this juncture, we mention that due to their simplicity, there has been a revived interest in studying gradient based methods for solving very large scale optimization problems, see e.g., the recent work of Ben-Tal et al. [1]. Indeed, more efficient algorithms (e.g. interior point methods) require heavy computational cost at each iteration and, due to the size of a given problem, often cannot even complete a single iteration. Thus, it is of interest to further study the possibility of improving efficiency of simpler gradient based algorithms when applied to specially structured problems. As recalled above, while the CGM exhibits only sublinear rate of convergence for a general problem, exploiting the special structure of problem (OP) allows for deriving an analysis with an improved rate. In Section 3, we prove that under the mild and standard Slater's condition on the system (I), the CGM converges to a solution of (I) at a linear rate. The rate of convergence depends on the problem's data, e.g., the matrix M, the vector g and on the radius of the largest ball contained in the feasible set of (I). In the course of our analysis, we also show that a recent algorithm studied in [5] for problems of the form (I), with S being a conic linear system in compact form, that is, with $S := C \cap \{x : u^T x = 1\}$ where C is a closed convex cone and $u \in \Re^n$ is a fixed given point such that S is compact, is nothing else but the CGM. We also compare our result with the one proven in [5], and show that a somewhat sophisticated condition imposed there on the problem's data, is in fact equivalent to the simple Slater's condition for the system (I). Our notations are mostly standard. The Euclidean space is denoted by $\Re^n$ with inner product $\langle \cdot, \cdot \rangle$ and the associated l_2 norm $\| \cdot \|$. For any matrix A, the norm of A is defined by $\|A\| = \max\{\|Ax\| : \|x\| \leq 1\}$. For any set $S \subset \Re^n$ we denote by $\mathrm{ri}(S)$, $\mathrm{int}(S)$, $\mathrm{cl}(S)$ respectively the relative interior, interior, and closure of S and by $\partial S = \mathrm{cl}(S) \setminus \mathrm{int}(S)$ the boundary of S. For a cone $K \subseteq \Re^n$ the polar cone is $K^* = \{x^* : \langle x, x^* \rangle \leq 0 \ \ \forall x \in K\}$.

2 The conditional gradient method and preliminary results

In this section we recall the basic steps and convergence results on the conditional gradient method, see e.g. [2] for details and references, as well as some other technical results that will be needed in the rest of this paper.

Consider the convex optimization problem:

$(P) \quad \min_{x \in S} f(x)$

Unless otherwise specified, throughout this section we assume that f is a convex continuously differentiable function on the closed and bounded convex set $S \subset \Re^n$, with Lipschitz gradient ∇f on S, i.e.,

$$\exists L > 0 \text{ such that } \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \ \ \forall x, y \in S,$$

and we set $f^* := \min_{x \in S} f(x)$.

Conditional Gradient Method-CGM: Start with $x^0 \in S$. Generate the sequence $\{x^k\}$, $\forall k = 1, 2, \ldots$ via the following steps:

1. Compute $p^{k-1} = \operatorname{argmin}\{\langle p - x^{k-1}, \nabla f(x^{k-1})\rangle : p \in S\}$.
2. Stopping Criteria: Let $S(x) := \min_{p \in S}\langle p - x, \nabla f(x)\rangle$. If $S(x^{k-1}) = \langle p^{k-1} - x^{k-1}, \nabla f(x^{k-1})\rangle = 0$ STOP. Else, goto step 3.
3. Line search: Compute $\lambda^{k-1} = \operatorname*{argmin}_{\lambda \in [0,1]} f(x^{k-1} + \lambda(p^{k-1} - x^{k-1}))$.

   Update $x^k = x^{k-1} + \lambda^{k-1}(p^{k-1} - x^{k-1})$.
4. Set $k \leftarrow k + 1$. Goto step 1.

It is easy to verify that

$$f(x^{k-1}) \geq f^* \geq f(x^{k-1}) + S(x^{k-1}), \tag{1}$$

and thus one always has $S(x^{k-1}) \leq 0$ and $S(x^{k-1}) = 0$ if and only if x^{k-1} is an optimal solution of problem (P), justifying the stopping criteria given in step 2.

The bulk of computation in the CGM are in Step 1 and Step 3. The latter requires to find a step size λ^{k-1} by solving the following one dimensional problem. Given x, p in S find λ^* solution of

$$\min_{\lambda \in [0,1]} f(x + \lambda(p - x)).$$

This step can in fact be computed analytically by using an appropriate quadratic approximation of the function f. Such approximation exists since we assumed here that ∇f is Lipschitz continuous. Indeed, the quadratic approximation follows by using the so-called descent lemma [2, Proposition A.24] (see Appendix). Thus the only remaining computational step in CGM is step 1 which in many applications might be very easy to solve. For example, whenever S is a simplex, in which case the solution is immediate or whenever the constraint set is a polyhedron, namely we have to solve a linear programming problem. Thus, CGM is an attractive simple algorithm whenever step 1 can be performed efficiently. The main known results on the conditional gradient method without any more assumptions (except for the ones we have already assumed) are summarized in the following proposition.[1]

[1] Since many of these results have been scattered in several references in the literature (see e.g., [2, 3, 4, 7]), for convenience and the interested reader on general results for CGM, we have given in an appendix compact proofs.

Proposition 2.1 *Let $f \in C^1(\Re^n)$ be a convex function with Lipschitz continuous gradient and Lipschitz constant $L > 0$. Let $\{x^k\}$ be a sequence generated by the conditional gradient method. Then,*

(i) $x^k \in S$, the *sequence* $\{f(x^k)\}$ *is monotone decreasing and every limit point of the sequence* $\{x^k\}$ *solves* $\min_{x\in S} f(x)$.
(ii) $\lim_{n\to\infty} f(x^n) = f^* = \min_{x\in S} f(x)$.
(iii) There exists a positive constant c, which depends on L and the diameter $\delta_S := \sup_{x,y\in S} \|x - y\|$ *such that* $f(x^n) - f^* \leq \frac{c}{n}$.

Note that convexity is not needed to derive the first statement of the proposition. In that case of course, the statement on the sequence $\{x^k\}$ is that every limit point is a stationary point, i.e., it satisfies the necessary local optimality conditions for problem (P).

The sublinear rate of convergence for function values cannot be improved, see e.g., Cannon and Cullum [3], unless, as we already mentioned in the introduction, we make some further stronger assumptions on the feasible set S, and this even if we assume that the objective function is convex quadratic which is our problem of interest in this paper. Exploiting the special structure of the objective, the next section develops the required analysis to achieve a linear rate of convergence of CGM for such class of problems.

3 Linear rate of convergence analysis of CGM

We consider the problem of finding a point satisfying:

$$(I)\begin{cases} Mx = g \\ x \in S \end{cases}$$

where S is a closed convex and bounded set. To solve this problem we consider the equivalent optimization problem:

$$(OP) \quad v^* := \min_{x\in S} \frac{1}{2}\|Mx - g\|^2.$$

Clearly, if (I) is feasible the optimal function value of (OP) is $v^* = 0$, otherwise one has $v^* > 0$.

We will apply the conditional gradient method CGM to (OP). The line search applied to the case of a convex quadratic objective is simple as it has an analytic expression (as it obviously does not require the use of a quadratic approximation of the objective). Indeed, with $f(x) = \frac{1}{2}\|Mx - g\|^2$ one has $\nabla f(x) = M^T(Mx - g)$ and we immediately obtain the following identity: for any $x, p \in \Re^n$ and any $\lambda \in \Re$:

$$g(\lambda) := f(x + \lambda(p - x)) = f(x) + \lambda\langle p - x, \nabla f(x)\rangle + \frac{1}{2}\lambda^2\|M(x - p)\|^2. \quad (2)$$

In order to simplify the expressions we use the following notations:

$$v_{k-1} = g - Mx^{k-1}, \quad (3)$$

$$w_{k-1} = g - Mp^{k-1}. \quad (4)$$

Using the identity (2) at the points $x = x^{k-1}, p = p^{k-1}$ and denoting by $g_k(\lambda)$ the resulting function, the step size computation in the line search of Step 3 of CGM, consists of finding $\lambda^* = \operatorname{argmin}_{\lambda\in[0,1]} g_k(\lambda)$.

This is a simple one dimensional convex quadratic minimization problem over the interval $[0, 1]$. A direct computation shows that one has $g'_k(\lambda) = 0$ if and only if:

$$\lambda = -\frac{\langle p^{k-1} - x^{k-1}, \nabla f(x^{k-1})\rangle}{\|M(p^{k-1} - x^{k-1})\|^2} = \frac{\langle v_{k-1}, v_{k-1} - w_{k-1}\rangle}{\|v_{k-1} - w_{k-1}\|^2}. \tag{5}$$

Thus,

$$\lambda^* = \underset{\lambda\in[0,1]}{\operatorname{argmin}} f(x^{k-1} + \lambda(p^{k-1} - x^{k-1})) \tag{6}$$

$$= \begin{cases} \frac{\langle v_{k-1}, v_{k-1} - w_{k-1}\rangle}{\|v_{k-1} - w_{k-1}\|^2} & \text{if } \frac{\langle v_{k-1}, v_{k-1} - w_{k-1}\rangle}{\|v_{k-1} - w_{k-1}\|^2} < 1 \\ 1 & \text{if } \frac{\langle v_{k-1}, v_{k-1} - w_{k-1}\rangle}{\|v_{k-1} - w_{k-1}\|^2} \geq 1 \end{cases} \tag{7}$$

Now, the main computational step of the conditional gradient method given in CGM-Step 1 is $p^{k-1} = \underset{p\in S}{\operatorname{argmin}}\{\langle p - x^{k-1}, \nabla f(x^{k-1})\rangle\}$. Substituting the expression of the gradient of f: $\nabla f(x) = M^T(Mx - g)$, and using the definition of v_k (cf. (3)), we obtain,

$$p^{k-1} = \underset{p\in S}{\operatorname{argmin}}\langle p - x^{k-1}, M^T(Mx^{k-1} - g)\rangle = \underset{p\in S}{\operatorname{argmin}}\langle v_{k-1}, g - Mp\rangle - \|v_{k-1}\|^2. \tag{8}$$

To summarize, the basic steps of the conditional gradient method for the quadratic problem (OP) has the following form:

The conditional gradient method applied to (OP): CGM-OP

Initialization step: Start with an arbitrary $x^0 \in S$

General step: Solve: $p^{k-1} = \underset{p\in S}{\operatorname{argmin}}\langle v_{k-1}, g - Mp\rangle \quad k = 1, 2, \ldots$

and compute: $\lambda^{k-1} = \begin{cases} \frac{\langle v_{k-1}, v_{k-1} - w_{k-1}\rangle}{\|v_{k-1} - w_{k-1}\|^2} & \text{if } \frac{\langle v_{k-1}, v_{k-1} - w_{k-1}\rangle}{\|v_{k-1} - w_{k-1}\|^2} < 1 \\ 1 & \text{if } \frac{\langle v_{k-1}, v_{k-1} - w_{k-1}\rangle}{\|v_{k-1} - w_{k-1}\|^2} \geq 1 \end{cases}$.

Update: $x^k = x^{k-1} + \lambda^{k-1}(p^{k-1} - x^{k-1})$

The stopping function $S(\cdot)$ defined in step 2 of CGM can be expressed as follows:

$$\begin{aligned} S(x^{k-1}) &= \langle p^{k-1} - x^{k-1}, \nabla f(x^{k-1})\rangle \\ &= \langle g - Mx^{k-1}, Mx^{k-1} - g + g - Mp^{k-1}\rangle = \langle v_{k-1}, w_{k-1}\rangle - \|v_{k-1}\|^2. \end{aligned}$$

The algorithm CGM-OP will produce an optimal solution at iteration k whenever $\langle v_{k-1}, w_{k-1}\rangle = \|v_{k-1}\|^2$. If in addition $v_{k-1} \neq 0$, i.e., $\langle v_{k-1}, w_{k-1}\rangle > 0$, then CGM-OP will stop with an infeasible solution of (OP) (which means that the original problem (I) is infeasible).

Applying CGM-OP to the special case when $S = \{x \in C : u^T x = 1\}$, where C is closed convex cone and $u \in \Re^n$ is a given fixed point such that S is

bounded, the above development shows that we have precisely recovered the algorithm GVNA proposed in [5, p.461–462].

As a byproduct of this equivalence between CGM and GVNA we can thus derive as an immediate consequence of Proposition 2.1

Corollary 3.1 *Suppose that system (I) is feasible and let* $\{x^k\}$ *be the sequence generated by CGM-OP, and let* $v_k = g - Mx^k$. *Then,* $\{\|v_k\|^2\}$ *converges to 0 with at least a sublinear rate, i.e.,* $\exists \eta > 0 : \|v_k\| \leq \frac{\eta}{\sqrt{k}}$.

Note that in the quadratic case, we don't need to use the quadratic approximation in the line search and thus we can write η explicitly in terms of the data (M, g, u) for the special case $S = C \cap \{x : u^T x = 1\}$, thus recovering the results of [5] in the feasible case.

We are now going to prove our main result concerning the efficiency of CGM-OP. Our approach is inspired from a proposition derived in [5, Proposition 6] for establishing linear convergence of GVNA. However, here we introduce a new idea that leads to a simple quantity for measuring the convergence rate, and which allows us to establish linear convergence under the sole and mild Slater's assumption on problem (I).

We denote the distance from a point $b \in \Re^n$ to the boundary ∂S of a closed convex set of $\Re^n$ by

$$d(b, \partial S) := \inf\{||z - b|| : z \in \partial S\}.$$

One thus has

$$d(b, \partial S) = \begin{cases} \min\{||z - b|| : z \in S\} & \text{if } b \notin S \\ \max\{r : B(b, r) \subset S\} & \text{if } b \in S, \end{cases}$$

where $B(b, r)$ is the ball centered at b with radius r.

We will make the following assumption throughout the rest of the paper:

Assumption. The row vectors of the matrix M are linearly independent.

This implies that the Gram matrix MM^T is positive definite and thus has an inverse. Note that this assumption is without loss of generality. It simply means that there are no redundant equations in the system $Mx = g$.

Proposition 3.1 *Let* $\{x^k\}$ *be the sequence generated by CGM-OP, let* p^k *be the direction computed in the general step at iteration* $k + 1$ *and let* $v_k = g - Mx^k$. *Suppose that the Slater condition for the convex linear system (I) is satisfied, i.e.,*

$$\exists\, \hat{x} \in \text{int}(S) \text{ such that } M\hat{x} = g.$$

Then,

$$\langle v_k, g - Mp^k \rangle + R_S(\hat{x}, M)\|v_k\| \leq 0, \tag{9}$$

where

$$R_S(\hat{x}, M) = \frac{d(\hat{x}, \partial S)}{\sqrt{\|(MM^T)^{-1}\|}}.$$

Proof. First, note that one has:

$$v_k = g - Mx^k = M\hat{x} - Mx^k = M(\hat{x} - x^k) := Md.$$

Thus, the system $Md = v_k$ has at least one solution. Among all possible solutions, we pick the one with minimum norm, that is we are interested in finding d^*, which solves the following optimization problem:

$$d^* = \min_{Md=v_k} \|d\|^2.$$

It is easy to see that the optimum of this minimization problem is attained at $d^* = M^T(MM^T)^{-1}v_k$ and $\|d^*\|^2 = v_k^T(MM^T)^{-1}v_k$. As a consequence,

$$\begin{aligned}\|d^*\| &= \sqrt{v_k^T(MM^T)^{-1}v_k} \\ &\leq \sqrt{\|(MM^T)^{-1}\| \cdot \|v_k\|^2} \\ &= \sqrt{\|(MM^T)^{-1}\|} \cdot \|v_k\|. \end{aligned} \tag{10}$$

Define $s := d(\hat{x}, \partial S)$. Since we assumed $\hat{x} \in \text{int}(S)$ one has $s > 0$. From the definition of s it follows that: $x = \hat{x} + s\frac{d^*}{\|d^*\|} \in S$, and hence,

$$\begin{aligned} Mx &= M\left(\hat{x} + s\frac{d^*}{\|d^*\|}\right) = M\hat{x} + s\frac{Md^*}{\|d^*\|} \\ &= g + s\frac{v_k}{\|d^*\|}. \end{aligned}$$

Thus, one has $g - Mx = -s\frac{v_k}{\|d^*\|}$ and therefore using (10) it follows that,

$$\langle v_k, g - Mp^k\rangle \overset{(8)}{\leq} \langle v_k, g - Mx\rangle = -\frac{s\|v_k\|^2}{\|d^*\|} \leq -\frac{s}{\sqrt{\|(MM^T)^{-1}\|}}\|v_k\|$$

proving the desired result. □

Remark 3.1 (a) Proposition 3.1 can be easily extended to bounded sets of the form $S = T \cap \{x : Ax = b\}$ where T is a closed convex set. Under the Slater condition (i.e., there exists $\hat{x} \in \text{int}(T)$ such that $M\hat{x} = g$ and $A\hat{x} = b$) (9) is satisfied, but here $R_S(\hat{x}, M)$ is defined by $R_S(\hat{x}, M) = \frac{d(\hat{x}, \partial S)}{\sqrt{\|(\tilde{M}\tilde{M}^T)^{-1}\|}}$ with $\tilde{M} = \begin{pmatrix} M \\ A \end{pmatrix}$, and the rows of $\tilde{M}$ are linearly independent.

(b) The above analysis assumed that S is full dimensional. If this assumption fails, (e.g., like in problems given in Remark 3.1), it can be shown that the result holds by weakening the Slater condition of Proposition 3.1 to the more general one,

$$\exists\, \hat{x} \in ri(S) \;\; \textit{such that} \;\; M\hat{x} = g.$$

where $S = T \cap \{x : Ax = b\}$ and T is full dimensional. In this case, the formula for R_S is as in Remark 3.1(a), but with replacing ∂S by $\text{rbd}(S)$, the relative boundary of S.

Proposition 3.1 can now be used to prove the following linear convergence rate for the conditional gradient method.

Proposition 3.2 *Suppose that the Slater condition is satisfied at the point $\hat{x}$ for the system (I) and let ρ_S be the radius of a ball containing the compact set S. Then, the conditional gradient method has a linear rate of convergence:*

$$\|v_k\| \leq (1-q^2)^{\frac{1}{2}}\|v_{k-1}\| \quad \forall k = 1,2,\ldots.$$

where $q = \frac{R_S(\hat{x},M)}{\|g\|+\rho_S\|M\|}$. Equivalently, this means that

$$||v_k|| \leq ||v_0||e^{-\frac{kq}{2}}, \quad \forall k = 1,\ldots$$

Proof. First recall that from the CGM-OP one has $w_{k-1} = g - Mp^{k-1}, v_{k-1} = g - Mx^{k-1}$ and $x^k = x^{k-1} + \lambda^*(p^{k-1} - x^{k-1})$ where $\lambda^* = \min\left\{\frac{\langle v_{k-1}, v_{k-1}-w_{k-1}\rangle}{\|v_{k-1}-w_{k-1}\|^2}, 1\right\}$. A short computation shows that,

$$\|v_k\|^2 = \|g - Mx^k\|^2 = (\lambda^*)^2\|v_{k-1} - w_{k-1}\|^2 + 2\lambda^*\langle v_{k-1}, w_{k-1} - v_{k-1}\rangle + \|v_{k-1}\|^2 \tag{11}$$

By (9) we have that $\langle v_{k-1}, w_{k-1}\rangle \leq 0$. Therefore,

$$\begin{aligned}\langle v_{k-1}, v_{k-1} - w_{k-1}\rangle &= \|v_{k-1}\|^2 - \langle v_{k-1}, w_{k-1}\rangle \\ &\overset{\langle v_{k-1}, w_{k-1}\rangle \leq 0}{\leq} \|v_{k-1}\|^2 - \langle v_{k-1}, w_{k-1}\rangle + (\|w_{k-1}\|^2 - \langle v_{k-1}, w_{k-1}\rangle) \\ &= \|v_{k-1} - w_{k-1}\|^2,\end{aligned}$$

and hence $\frac{\langle v_{k-1}, v_{k-1}-w_{k-1}\rangle}{\|v_{k-1}-w_{k-1}\|^2} \leq 1$ which implies that $\lambda^* = \frac{\langle v_{k-1}, v_{k-1}-w_{k-1}\rangle}{\|v_{k-1}-w_{k-1}\|^2}$. Substituting this value of λ^* in (11) yields:

$$\|v_k\|^2 = \frac{\|v_{k-1}\|^2\|w_{k-1}\|^2 - \langle v_{k-1}, w_{k-1}\rangle^2}{\|v_{k-1} - w_{k-1}\|^2}. \tag{12}$$

Now, since S is a bounded set, it is contained in some ball $B(0,\rho_S)$ and thus one has $\|w_{k-1}\| = \|g - Mp^{k-1}\| \leq \|g\| + \|M\|\rho_S$. Moreover, note that $\|v_{k-1} - w_{k-1}\|^2 = \|v_{k-1}\|^2 - 2\langle v_{k-1}, w_{k-1}\rangle + \|w_{k-1}\|^2 \geq \|w_{k-1}\|^2$. Therefore we obtain from (12) (we set here $R := R_S(\hat{x}, M)$):

$$\begin{aligned}\|v_k\|^2 &= \frac{\|v_{k-1}\|^2\|w_{k-1}\|^2 - \langle v_{k-1}, w_{k-1}\rangle^2}{\|v_{k-1} - w_{k-1}\|^2} \\ &\overset{(9)}{\leq} \frac{\|v_{k-1}\|^2(\|w_{k-1}\|^2 - R^2)}{\|v_{k-1} - w_{k-1}\|^2} \\ &\overset{\|v_{k-1}-w_{k-1}\|^2 \geq \|w_{k-1}\|^2}{\leq} \frac{\|v_{k-1}\|^2(\|w_{k-1}\|^2 - R^2)}{\|w_{k-1}\|^2} \\ &= \left(1 - \frac{R^2}{\|w_{k-1}\|^2}\right)\|v_{k-1}\|^2 \\ &\leq \left(1 - \left(\frac{R}{\|g\| + \rho_S\|M\|}\right)^2\right)\|v_{k-1}\|^2,\end{aligned}$$

proving the first statement of the Proposition. From the last inequality it follows that $\|v_{k-1}\|^2 - \|v_k\|^2 \geq \frac{R}{\|g\|+\rho_S\|M\|}\|v_{k-1}\|^2$. Invoking Lemma A.1(ii) given in the appendix, to the nonnegative sequence $a_k := ||v_k||$ the equivalent part of the proposition is obtained. □

We can apply the above result to find an approximate solution of (I) with fixed accuracy. Given $\varepsilon > 0$, an ε-solution of (I), namely a point $x \in S$ such that $||Mx - g|| \leq \varepsilon$, is obtained in no more than

$$k = \left\lceil 2\frac{\|g\| + \rho_S\|M\|}{R_S(\hat{x}, M)} \ln\left(\frac{\|g - Mx^0\|}{\epsilon}\right)\right\rceil$$

iterations of CGM.

It is interesting to compare the linear rate of convergence result derived in Proposition 3.1 with the one derived in [5]. The linear convergence rate derived in that paper, was obtained in terms of another quantity $r(M, g)$ defined by:

$$r(M, g) = \inf\{\|g - h\| : h \in \partial H\}, \tag{13}$$

where, $H = \{Mx : x \in S\} = M(S)$.

To derive linear convergence for a feasible problem (I), [5] need to impose the condition:

$$r(M, g) > 0. \tag{14}$$

Computing such a quantity does not appear to be an easy task even if we are given a feasible solution of (I) . However[2], it is easy to prove that for any feasible point $\hat{x}$ of (I) one has $r(M, g) \geq R_S(\hat{x}, M)$. In comparison with Proposition 3.1, $r(M, g)$ is needed not only to measure the rate, but also as a criteria (cf. (14)) to guarantee linear convergence. Interestingly enough, it turns out that the condition (14) imposed in the analysis of [5] is in fact just Slater's condition in disguise.

Proposition 3.3 *Suppose that the convex feasibility problem (I) is feasible. Then* $r(M, g) > 0$ *if and only if there exists* $\bar{x} \in \operatorname{int}(S)$ *such that* $M\bar{x} = g$.

Proof. Under the given feasibility assumption, problem (I) has a solution and thus we have that $g \in H$. Now, $r(M, g) > 0$ is equivalent to $g \notin \partial H$ and thus $g \in \operatorname{int} H$. Using relative interior calculus ([9, Proposition 6.6, p.48]) and the fact that the relative interior and the interior are the same in this case one has $\operatorname{int}(H) = \operatorname{int} M(S) = M(\operatorname{int}(S))$. Therefore, $g \in \operatorname{int} H$ translates to: there exists $\bar{x} \in \operatorname{int}(S)$ such that $g = M\bar{x}$. □

A Appendix

The following well known properties of nonnegative sequences, (for proofs see e.g., Polyak [8]), are used to derive the rate of convergence given in Corollary 3.1 and Propositions 2.1 and 3.2.

[2] We thank a referee for pointing out to us this fact.

Lemma A.1 *Let $\{a_k\}_{k=0}^m$ be a nonnegative sequence of real numbers.*

(i) Sublinear rate: If $\{a_k\}$ is such that $a_{k-1} - a_k \geq \gamma a_{k-1}^2$ for some $\gamma > 0$ and for any $k = 1, \ldots, m$, then

$$a_m \leq \frac{a_0}{1 + m\gamma a_0} < (\gamma m)^{-1}.$$

(ii) Linear Rate: If $\{a_k\}$ is such that $a_{k-1} - a_k \geq \gamma_k a_{k-1}$ for some $\gamma_k \geq 0$, $\forall\, k = 1, \ldots, m$, then

$$a_m \leq a_0 e^{-\sum_{k=1}^m \gamma_k}.$$

In the remaining of the appendix we outline compact proofs of the well known results summarized in Proposition 2.1, when minimizing a convex continuously differentiable function with Lipschitz gradient over a compact convex set S. In what follows $x^k \in S$ is the sequence produced by CGM as outlined in Section 2, and $S(x) = \min_{p \in S} \langle p - x, \nabla f(x) \rangle$.

Lemma A.2 *Let $f : \Re^n \to \Re$ be a continuously differentiable function with Lipschitz gradient and Lipschitz constant $L > 0$ over the compact set S. Let $\{x^k\}$ be the sequence generated by CGM. Then,*

$$f(x^{k-1}) - f(x^k) \geq \frac{1}{2} \frac{S^2(x^{k-1})}{\|x^{k-1} - p^{k-1}\|^2} \cdot \min\left\{\frac{1}{L}, \frac{\|x^{k-1} - p^{k-1}\|}{\|\nabla f(x^{k-1})\|}\right\}.$$

Proof. The proof follows by applying the descent Lemma [2, Proposition A.24] which gives $\forall \lambda \in [0, 1]$,

$$f(x^{k-1}) - f(x^{k-1} + \lambda(p^{k-1} - x^{k-1})) \geq \lambda \langle x^{k-1} - p^{k-1}, \times \nabla f(x^{k-1}) \rangle - \frac{L}{2} \lambda^2 \|x^{k-1} - p^{k-1}\|^2. \tag{15}$$

The later inequality is in particular true for

$$\lambda^* = \underset{0 \leq \lambda \leq 1}{\operatorname{argmax}} \left\{ \lambda\alpha - \frac{1}{2}\beta\lambda^2 \right\} = \begin{cases} 1 & \text{if } 1 \leq \frac{\alpha}{\beta} \\ \frac{\alpha}{\beta} & 1 > \frac{\alpha}{\beta} \end{cases},$$

where, $\alpha = \langle x^{k-1} - p^{k-1}, \nabla f(x^{k-1}) \rangle$, $\beta = L\|x^{k-1} - p^{k-1}\|^2$.

Thus, the step size in Step 3 of CGM can be taken as $\lambda^* = \min\left\{1, \frac{\langle x^{k-1} - p^{k-1}, \nabla f(x^{k-1}) \rangle}{L\|x^{k-1} - p^{k-1}\|^2}\right\}$. If $\lambda^* = 1$, then in this case,

$$1 \leq \frac{\langle x^{k-1} - p^{k-1}, \nabla f(x^{k-1}) \rangle}{L\|x^{k-1} - p^{k-1}\|^2}. \tag{16}$$

and therefore one obtains:

$$
\begin{aligned}
f(x^{k-1}) - f(x^k) \quad &\geq \langle x^{k-1} - p^{k-1}, \nabla f(x^{k-1})\rangle - \frac{L}{2}\|x^{k-1} - p^{k-1}\|^2 \\
&\overset{(16)}{\geq} \frac{1}{2}\langle x^{k-1} - p^{k-1}, \nabla f(x^{k-1})\rangle \\
&\overset{a^T b \geq \frac{(a^T b)^2}{\|a\|\cdot\|b\|}}{\geq} \frac{1}{2}\frac{\langle x^{k-1} - p^{k-1}, \nabla f(x^{k-1})\rangle^2}{\|x^{k-1} - p^{k-1}\| \cdot \|\nabla f(x^{k-1})\|} \\
&= \frac{1}{2}\frac{S^2(x^{k-1})}{\|x^{k-1} - p^{k-1}\| \cdot \|\nabla f(x^{k-1})\|}. \qquad (17)
\end{aligned}
$$

In a similar way, in the other case $\lambda^* = \frac{\langle x^{k-1} - p^{k-1}, \nabla f(x^{k-1})\rangle}{L\|x^{k-1} - p^{k-1}\|^2}$, and we obtain,

$$
\begin{aligned}
f(x^{k-1}) - f(x^k) &\geq \frac{\langle x^{k-1} - p^{k-1}, \nabla f(x^{k-1})\rangle^2}{L\|x^{k-1} - p^{k-1}\|^2} - \frac{L}{2}\frac{\langle x^{k-1} - p^{k-1}, \nabla f(x^{k-1})\rangle^2}{L^2\|x^{k-1} - p^{k-1}\|^2} \\
&= \frac{1}{2}\frac{\langle x^{k-1} - p^{k-1}, \nabla f(x^{k-1})\rangle^2}{L\|x^{k-1} - p^{k-1}\|^2} = \frac{1}{2}\frac{S^2(x^{k-1})}{L\|x^{k-1} - p^{k-1}\|^2}. \qquad (18)
\end{aligned}
$$

□

Lemma A.3 *For any $f : \Re^n \to \Re$ which is continuously differentiable with Lipschitz gradient and Lipschitz constant L over $S \subseteq \Re^n$ closed, bounded and convex one has*:

(i) $\sup_{x \in S} \|\nabla f(x)\| \leq c_2$ *for some constant* c_2.

(ii) $\forall x \in S \quad \|p(x) - x\| \leq c_1$ *for some* $c_1 > 0$ *where* $p(x) := \operatorname{argmin}_{p \in S}\langle p - x, \nabla f(x)\rangle$.

Proof. (i) Since the gradient of f is Lipschitz, we obtain for any $x, y \in S$: $\|\nabla f(x)\| = \|\nabla f(x) - \nabla f(y) + \nabla f(y)\| \leq L\|x - y\| + \|\nabla f(y)\| \leq L\delta_S + \|\nabla f(y)\|$, where, $\delta_S = \sup_{x,y \in S}\|x - y\|$ and (i) is proved with $c_2 := L\delta_S + \|\nabla f(y)\|$.

(ii) Since S is compact and for all $x \in S$ we have that $p(x) \in S$. Thus for $c_1 = \delta_S$ we have that $\|p(x) - x\| \leq c_1 \quad \forall x \in S$. □

Applying the results of the previous lemma to lemma A.2 we obtain:

Proposition A.1 *Let $f : \Re^n \to \Re$ be a continuously differentiable function with Lipschitz gradient and Lipschitz constant $L > 0$ over the compact set S. Let $\{x^k\}$ be the sequence generated by CGM. Define, $C := \min\left\{\frac{1}{2c_1c_2}, \frac{1}{2Lc_1^2}\right\} > 0$. Then,*

$$
f(x^{k-1}) - f(x^k) \geq CS^2(x^{k-1}), \quad \forall k = 1, \ldots \qquad (19)
$$

Proof. By lemma A.2 we have that:

$$f(x^{k-1}) - f(x^k) \geq \frac{1}{2}\frac{S^2(x^{k-1})}{\|x^{k-1}-p^{k-1}\|^2}\cdot\min\left\{\frac{1}{L},\frac{\|x^{k-1}-p^{k-1}\|}{\|\nabla f(x^{k-1})\|}\right\} \overset{\text{lemma A.3}}{\geq} \left\{\frac{1}{2c_1c_2},\frac{1}{2Lc_1^2}\right\}S^2(x^{k-1}).$$

□

Before proving Proposition 2.1, the next result shows that every limit point of CGM is a stationary point of $\min_{x\in S} f(x)$. No convexity assumption is needed.

Proposition A.2 *Let $f:\Re^n\to\Re$ be a continuously differentiable function with Lipschitz gradient and Lipschitz constant $L>0$ over the compact set S. Let $\{x^k\}$ be the sequence generated by CGM. Then,*

(i) $x^k\in S$ and $\{f(x^k)\}$ *is a monotone decreasing sequence, and* $S(x^k)\to 0$ as $k\to\infty$.

(ii) Every limit point x^* of $\{x^k\}$ *is a stationary point, i.e., it satisfies the necessary conditions for local minimum:* $\langle\nabla f(x^*),x-x^*\rangle\geq 0\ \ \forall x\in S$.

Proof. (i) The first part of the statement follows immediately from (19), while the second part is a consequence of

$$\sum_{k=1}^{n} S^2(x^{k-1}) \overset{(19)}{\leq} C^{-1}(f(x^0)-f(x^n)) \leq C^{-1}(f(x^0)-f^*) < \infty,$$

where f^* is the global minimum of f over S. This implies that $S(x^k)\to 0$ as $k\to\infty$. To show (ii), suppose first that there is a k such that $S(x^{k-1})=0$. Then $\langle\nabla f(x^{k-1}),p-x^{k-1}\rangle\geq 0\ \forall p\in S$ thus x^{k-1} is a stationary point by definition and the proposition is proved. Otherwise, one has $f(x^k)<f(x^{k-1})\ \ \forall k=1,\ldots$. Let x^* be a limit point of $\{x^k\}$. Then there exists a subsequence $\{x^{n_k}\}$ that converges to x^* and we have:

$$\begin{cases}\langle p^{n_k},\nabla f(x^{n_k})\rangle\leq\langle p,\nabla f(x^{n_k})\rangle & \forall p\in S\\ \langle x^{n_k}-p^{n_k},\nabla f(x^{n_k})\rangle\to 0\end{cases}$$

$\{p^{n_k}\}\subseteq S$ and thus it is a bounded sequence and consequently has a limit point $\bar{p}$. Also, ∇f is continuous and we have:

$$\begin{cases}\langle\bar{p},\nabla f(x^*)\rangle\leq\langle p,\nabla f(x^*)\rangle & \forall p\in S\\ \langle x^*-\bar{p},\nabla f(x^*)\rangle=0\end{cases}$$

Therefore $\langle x^*,\nabla f(x^*)\rangle\leq\langle p,\nabla f(x^*)\rangle\ \ \forall p\in S$, which proves that x^* is a stationary point. □

Proof of Proposition 2.1 For convex functions the optimal points are exactly the stationary points and thus (i) has already been proved. By (1) we have that:

$$S(x^{k-1})\leq f^*-f(x^{k-1})\leq 0 \qquad \forall k=1,2,\ldots, \tag{20}$$

and since $S(x^{k-1}) \to 0$ we obtain that $\lim_{n\to\infty} f(x^n) = f^*$ which proves (ii). It remains to prove the sublinear rate in function values (iii). From (19) we have

$$(f(x^{k-1}) - f^*) - (f(x^k) - f^*) \geq CS^2(x^{k-1}),$$

but from (20) we have $S^2(x^{k-1}) \geq (f(x^{k-1}) - f^*)^2$. Defining $a_k = f(x^{k-1}) - f^*$, $\gamma := C$, the result follows from Lemma A.1(i). □

Acknowldegements. We thank two referees for their constructive comments which has led to improve the presentation.

References

[1] Ben-Tal A, Margalit T, Nemirovski, A (2001) The ordered subsets mirror descent optimization method with applications to tomography. SIAM J. Optimization 12:79–108
[2] Bertsekas D (1999) Nonlinear Programming. Athena Scientific, Belmont Massachussets, second edition
[3] Cannon MD, Cullum CD (1968) A tight upper bound on the rate of convergence of the Frank-Wolfe algorithm. SIAM J. Control Optimization 6:509–516
[4] Dunn JC (1979) Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. SIAM J. Control Optim 17:187–211
[5] Epelman M, Freund RM (2000) Condition number complexity of an elementry algorithm for computing a reliable solution of a conic linear system. Math. Program. 88:451–485
[6] Frank M, Wolfe P (1956) An algorithm for quadratic programming. Naval Research Logistics Quaterly 3:95–110
[7] Levitin ES, Polyak BT (1966) Minimization methods in the presence of constraints. USSR Computational Math. and Math. Phys. 6:787–823
[8] Polyak BT (1987) Introduction to Optimization. Optimization Software Inc., New York
[9] Rockafellar RT (1970) Convex Analysis. Princeton University Press, Princeton, NJ