Mathematical Programming 39 (1987) 93-116
North-Holland

PROJECTED GRADIENT METHODS FOR LINEARLY CONSTRAINED PROBLEMS

Paul H. CALAMAI*
University of Waterloo, Ontario, Canada

Jorge J. MORÉ**
Argonne National Laboratory, Argonne, IL, USA

Received 9 June 1986
Revised manuscript received 16 March 1987

The aim of this paper is to study the convergence properties of the gradient projection method and to apply these results to algorithms for linearly constrained problems. The main convergence result is obtained by defining a projected gradient, and proving that the gradient projection method forces the sequence of projected gradients to zero. A consequence of this result is that if the gradient projection method converges to a nondegenerate point of a linearly constrained problem, then the active and binding constraints are identified in a finite number of iterations. As an application of our theory, we develop quadratic programming algorithms that iteratively explore a subspace defined by the active constraints. These algorithms are able to drop and add many constraints from the active set, and can either compute an accurate minimizer by a direct method, or an approximate minimizer by an iterative method of the conjugate gradient type. Thus, these algorithms are attractive for large scale problems. We show that it is possible to develop a finite terminating quadratic programming algorithm without non-degeneracy assumptions.

Key words: Linearly constrained problems, projected gradients, bound constrained problems, large scale problems, convergence theory.

1. Introduction

We are interested in the numerical solution of large scale minimization problems subject to linear constraints. Algorithms for solving these problems usually restrict the change in the dimension of the working subspace by only dropping or adding one constraint at each iteration. This implies, for example, that if there are $k \leq n$ constraints active at the solution but the starting point is in the interior of the feasible set, then the method will require at least k iterations to converge. This shortcoming does not apply to methods based on projected gradients, and for this reason there has been renewed interest in these methods.

* Work supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research of the U.S. Department of Energy under Contract W-31-109-Eng-38.

** Work supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research of the U.S. Department of Energy under Contract W-31-109-Eng-38.

Gradient projection methods for minimizing a continuously differentiable mapping $f: R^n \to R$ on a nonempty closed convex set $\Omega \subset R^n$ were originally proposed by Goldstein (1964) and Levitin and Polyak (1966). It is helpful to study the general problem

$$\min\{f(x): x \in \Omega\}, \tag{1.1}$$

because it clarifies the underlying structure of the algorithms. However, applications are usually concerned with special cases of (1.1). Most of the current interest in projected gradients has been concerned with the case where Ω is defined by the bound constraints

$$\Omega = \{x \in R^n: l \leqslant x \leqslant u\} \tag{1.2}$$

for some vectors l and u; we are interested in the general linearly constrained case where Ω is a polyhedral set.

Given an inner product norm $\|\cdot\|$ and a nonempty closed convex set Ω, the projection into Ω is the mapping $P: R^n \to \Omega$ defined by

$$P(x) = \operatorname{argmin}\{\|z - x\|: z \in \Omega\}. \tag{1.3}$$

The dependence of P on Ω is usually clear from the context, but if there is a possibility of confusion we shall use $P_\Omega(x)$ to denote the projection of x into Ω. Given the projection P into Ω, the gradient projection algorithm is defined by

$$x_{k+1} = P(x_k - \alpha_k \nabla f(x_k)), \tag{1.4}$$

where $\alpha_k > 0$ is the step, and ∇f is the gradient of f with respect to the inner product associated with the norm $\|\cdot\|$.

There are several schemes for selecting α_k which guarantee that any limit point x^* of $\{x_k\}$ satisfies the first order necessary conditions for optimality

$$(\nabla f(x^*), x - x^*) \geqslant 0, \quad x \in \Omega. \tag{1.5}$$

Any point x^* that satisfies condition (1.5) is a *stationary point* for problem (1.1). This terminology is appropriate because if the constraint functions that define Ω satisfy a constraint qualification then x^* is stationary if and only if x^* is a Kuhn-Tucker point.

The results of Goldstein (1964), and Levitin and Polyak (1966) show that if ∇f is Lipschitz continuous with Lipschitz constant κ, and if α_k satisfies

$$\varepsilon \leqslant \alpha_k \leqslant \frac{2}{\kappa}(1 - \varepsilon)$$

for some ε in $(0, 1)$, then any limit point of $\{x_k\}$ is a stationary point of problem (1.1). An obvious drawback of their results is the need for the Lipschitz constant κ. McCormick and Tapia [1972] essentially require the step α_k to be a global minimizer of

$$\min\{f(P(x_k - \alpha \nabla f(x_k)): \alpha > 0\}, \tag{1.6}$$

and prove that if f is continuously differentiable then limit points of (1.4) must be stationary. They also establish this result under the assumption that Ω is polyhedral with orthogonal constraint normals and α_k is a local minimizer of (1.6). Unfortunately, computing a minimizer of (1.6) for a polyhedral Ω is a difficult task because it requires the minimization of a piecewise continuously differentiable function. For additional results with this choice of step, see Phelps (1986).

Bertsekas (1976) was the first to propose a practical finite procedure to determine the step. Given β and μ in (0, 1) and $\gamma > 0$, Bertsekas proposed an Armijo procedure where

$$\alpha_k = \beta^{m_k}\gamma \tag{1.7}$$

and m_k is the smallest nonnegative integer such that

$$f(x_{k+1}) \leq f(x_k) + \mu(\nabla f(x_k), x_{k+1} - x_k). \tag{1.8}$$

Assuming that f is continuously differentiable and Ω is the convex set (1.2), Bertsekas (1976) shows that limit points of $\{x_k\}$ are stationary points of (1.1), and that if $\{x_k\}$ converges to a local minimizer which satisfies the strict complementarity and second order sufficiency conditions, then the set of active constraints is identified in a finite number of steps. Dunn (1981) analyzes the gradient projection algorithm for general convex sets Ω and f continuously differentiable. He shows that if α_k satisfies (1.8) and $\alpha_k \geq \varepsilon$ for some $\varepsilon > 0$, then any limit point of $\{x_k\}$ is stationary. He also notes that if ∇f is Lipschitz continuous with Lipschitz constant κ, and α_k is chosen by the Armijo procedure (1.7, 1.8) then

$$\alpha_k \geq \min\left\{\gamma, \frac{2}{\kappa}\beta(1-\mu)\right\}.$$

This shows that any limit point of (1.4) is a stationary point provided ∇f is Lipschitz. An unpublished report of Gafni and Bertsekas (1982) was brought to our attention after completing this work. In their report the gradient projection method with the Armijo procedure (1.7, 1.8) is considered, and it is shown that if f is continuously differentiable then any limit point of (1.4) is a stationary point of problem (1.1) for a general convex set Ω. For a compact Ω this result can also be obtained as a special case of the work of Gafni and Bertsekas (1984). In this paper we generalize these results by introducing the notion of a projected gradient and showing that the gradient projection algorithm drives the projected gradient to zero.

We are also interested in conditions which guarantee that the gradient projection method identifies the optimal active set in a finite number of steps. The first result in this direction was obtained by Bertsekas (1976) for the bound constrained problem (1.2). Related results have been obtained by Bertsekas (1982) for a projected Newton method, and by Gafni and Bertsekas (1984) for a two-metric projection method. We consider the gradient projection method and generalize the result of Bertsekas (1976) by showing that if the projected gradients converge to zero and if the iterates converge to a nondegenerate point, then the optimal active and binding constraints of a general linearly constrained problem are identified in a finite number of iterations. This result is independent of the method used to generate the iterates and can be applied to other linearly constrained algorithms.

The above results on the identification of the optimal active constraints have been generalized in recent work. Dunn (1986) proved that if a strict complementarity condition holds, then the gradient projection method identifies the optimal active constraints in a finite number of iterations. This result has been extended by Burke and Moré (1986). They proved, in particular, that under Dunn's nondegeneracy assumption the optimal active constraints are eventually identified if and only if the projected gradient converges to zero.

The convergence properties of the projected gradient method that we obtain revolve around the notion of the projected gradient. This concept is used frequently in the optimization literature, but it is invariably associated with affine subspaces. The projected gradient for convex sets has had limited use. For bound constrained problems the projected gradient is used by Dembo and Tulowitzki (1983) as a search direction and to determine stopping criteria for the solution of quadratic programming problems. For linearly constrained problems, Dembo and Tulowitzki (1984, 1985) use the projected gradient to develop rate of convergence results for sequential quadratic programming algorithms. Finally, for general convex sets, McCormick and Tapia (1972) use the projected gradient to define a steepest descent direction.

The slow convergence rate of the gradient projection method is an obvious drawback, and thus current research has been aimed at developing a superlinearly convergent version of the gradient projection algorithm. Many of these algorithms, however, do not have the simplicity and elegance of the gradient projection method. For example, the projected Newton method developed by Bertsekas (1982) uses an anti-zigzagging strategy and the acceptance criterion in the search procedure does not have the appeal of (1.7, 1.8). Similar remarks apply to the projection methods of Gafni and Bertsekas (1984). In this paper we follow a different approach which uses a standard unconstrained minimization algorithm to explore the subspace defined by the current active set, and the gradient projection method to choose the next active set. A similar approach is used by Bertsekas (1976) for optimization problems subject to bound constraints, and by Dembo and Tulowitzki (1983) for strictly convex quadratic programming problems subject to bound constraints, but our approach appears to be simpler and more general.

The aim of this paper is to study the convergence properties of the gradient projection method and to apply these results to algorithms for linearly constrained problems. As an application of our theory, we develop quadratic programming algorithms that iteratively explore a subspace defined by the active constraints. These algorithms are able to drop and add many constraints from the active set, and can either compute an accurate minimizer by a direct method, or an approximate minimizer by an iterative method of the conjugate gradient type. Thus, these algorithms are attractive for large scale problems.

We start the convergence analysis of the gradient projection method in Section 2. Our procedure for choosing α_k generalizes the Armijo procedure (1.7, 1.8); in particular, we do not require that $\{\alpha_k\}$ be bounded. We show that if f is continuously differentiable then limit points of the gradient projection method are stationary points of problem (1.1). We also show that it is possible to obtain a convergence result without imposing boundedness conditions on $\{x_k\}$.

In Section 3 we define the notion of a projected gradient and prove that the gradient projection method forces the sequence of projected gradients to zero. This is a new and important property of the gradient projection method. We also show that this property implies that limit points of the gradient projection method are stationary points.

We consider the case of a polyhedral Ω in Section 4, and prove that if the projected gradients converge to zero and if $\{x_k\}$ converges to a nondegenerate point, then the active and binding constraints are identified in a finite number of iterations. An interesting aspect of this result is that it is independent of the method used to generate $\{x_k\}$ and can thus be applied to other linearly constrained algorithms.

Sections 5 and 6 examine algorithms which only use the gradient projection method on selected iterations, with Section 6 restricted to algorithms for the general quadratic programming problem. The results of Section 6 show, in particular, that it is possible to develop a finitely terminating quadratic programming algorithm without non-degeneracy assumptions.

2. Convergence analysis

In our analysis of the gradient projection method we assume that Ω is a nonempty closed convex set and that the mapping $f: R^n \to R$ is continuously differentiable on Ω. Given an iterate x_k in Ω, the step α_k is obtained by searching the path

$$x_k(\alpha) \equiv P(x_k - \alpha \nabla f(x_k)),$$

where P is the projection into Ω defined by (1.3). Given a step $\alpha_k > 0$, the next iterate is defined by $x_{k+1} = x_k(\alpha_k)$. The step α_k is chosen so that, for positive constants γ_1 and γ_2, and constants μ_1 and μ_2 in (0, 1),

$$f(x_{k+1}) \leqslant f(x_k) + \mu_1 (\nabla f(x_k), x_{k+1} - x_k), \tag{2.1}$$

and

$$\alpha_k \geqslant \gamma_1 \quad \text{or} \quad \alpha_k \geqslant \gamma_2 \bar{\alpha}_k > 0 \tag{2.2}$$

where $\bar{\alpha}_k$ satisfies

$$f(x_k(\bar{\alpha}_k)) > f(x_k) + \mu_2 (\nabla f(x_k), x_k(\bar{\alpha}_k) - x_k). \tag{2.3}$$

Condition (2.1) on α_k forces a sufficient decrease of the function while condition (2.2) guarantees that α_k is not too small. Since Dunn (1981) has shown that if x_k

is not stationary then (2.3) fails for all $\bar{\alpha}_k > 0$ sufficiently small, it follows that there is an $\alpha_k > 0$ which satisfies (2.1) and (2.2) provided $\mu_1 \le \mu_2$. Note, in particular, that the Armijo procedure (1.7, 1.8) satisfies these conditions with $\gamma_1 = \gamma$, $\gamma_2 = \beta$, and $\mu_1 = \mu_2 = \mu$.

The analysis of the gradient projection method defined by (2.1) and (2.2) requires some basic properties of the projection operator.

Lemma 2.1. *Let P be the projection into Ω.*

(a) *If $z \in \Omega$ then $(P(x) - x, z - P(x)) \ge 0$ for all $x \in R^n$.*

(b) *P is a monotone operator, that is, $(P(y) - P(x), y - x) \ge 0$ for x, $y \in R^n$. If $P(y) \ne P(x)$ then strict inequality holds.*

(c) *P is a nonexpansive operator, that is, $\|P(y) - P(x)\| \le \|y - x\|$ for x, $y \in R^n$.*

Lemma 2.1 is well known and easy to prove. For example, Zarantonello (1971) proves this result (Lemmas 1.1 and 1.2) and also provides much additional information on projections. An immediate consequence of part (a) is that

$$(\nabla f(x_k), x_k - x_k(\alpha)) \ge \frac{\|x_k(\alpha) - x_k\|^2}{\alpha}, \quad \alpha > 0. \tag{2.4}$$

In particular, this implies that

$$(\nabla f(x_k), x_k - x_{k+1}) \ge \frac{\|x_{k+1} - x_k\|^2}{\alpha_k}. \tag{2.5}$$

Inequalities (2.4) and (2.5) were used by Dunn (1981) in his analysis of the gradient projection method. We also need to note that part (b) of Lemma 2.1 implies that

$$(\nabla f(x_k), x_k - x_k(\alpha)) \ge (\nabla f(x_k), x_k - x_k(\beta)) \quad \text{if } \alpha \ge \beta. \tag{2.6}$$

The next result is due to Gafni and Bertsekas [1984], but our proof is simpler.

Lemma 2.2. *Let P be the projection into Ω. Given $x \in R^n$ and $d \in R^n$, the function ψ defined by*

$$\psi(\alpha) = \frac{\|P(x + \alpha d) - x\|}{\alpha}, \quad \alpha > 0,$$

is antitone (nonincreasing).

Proof. Let $\alpha > \beta > 0$ be given. If $P(x + \alpha d) = P(x + \beta d)$ then $\psi(\alpha) \le \psi(\beta)$, so we only consider the case $P(x + \alpha d) \ne P(x + \beta d)$. We now show that the result follows by noting that if $(v, u - v) > 0$ then

$$\frac{\|u\|}{\|v\|} \le \frac{(u, u - v)}{(v, u - v)}.$$

The proof of this geometrical result only requires a short computation. If we set

$$u = P(x+\alpha d) - x, \qquad v = P(x+\beta d) - x,$$

then part (a) of Lemma 2.1 implies that

$$(u, u-v) \leqslant \alpha(d, P(x+\alpha d) - P(x+\beta d)),$$

and that

$$(v, u-v) \geqslant \beta(d, P(x+\alpha d) - P(x+\beta d)).$$

Moreover, since $\alpha > \beta$ and $P(x+\alpha d) \neq P(x+\beta d)$, part (b) of Lemma 2.1 shows that

$$(d, P(x+\alpha d) - P(x+\beta d)) > 0.$$

Hence $(v, u-v) > 0$, and thus the result follows from the above inequalities. □

We now proceed to the convergence analysis of the gradient projection method. In the following result we do not assume that $\{x_k\}$ is bounded but assume that ∇f is uniformly continuous; later on we show that the uniform continuity assumption can be dropped if some subsequence of $\{x_k\}$ is bounded.

Theorem 2.3. *Let $f: R^n \to R$ be continuously differentiable on Ω, and let $\{x_k\}$ be the sequence generated by the gradient projection method defined by* (2.1) *and* (2.2). *If f is bounded below on Ω and ∇f is uniformly continuous on Ω then*

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x_k\|}{\alpha_k} = 0.$$

Proof. Assume that there is an infinite subsequence K_0 such that

$$\frac{\|x_{k+1} - x_k\|}{\alpha_k} \geqslant \varepsilon > 0, \quad k \in K_0.$$

We will prove that this assumption leads to a contradiction. First note that if $k \in K_0$ then

$$\frac{\|x_{k+1} - x_k\|^2}{\alpha_k} \geqslant \varepsilon \max\{\varepsilon\alpha_k, \|x_{k+1} - x_k\|\},$$

and that since $\{f(x_k)\}$ converges, condition (2.1) implies that $\{(\nabla f(x_k), x_k - x_{k+1})\}$ converges to zero. Hence, (2.5) and the above inequality show that

$$\lim_{k\in K_0, k\to\infty} \alpha_k = 0 \quad \text{and} \quad \lim_{k\in K_0, k\to\infty} \|x_{k+1} - x_k\| = 0.$$

In particular, we have shown that eventually $\alpha_k < \gamma_1$ and hence, $\alpha_k \geqslant \gamma_2 \bar{\alpha}_k$ where $\bar{\alpha}_k$ satisfies (2.3). Now set $\bar{x}_{k+1} \equiv x_k(\bar{\alpha}_k)$ and $\beta_k = \min(\alpha_k, \bar{\alpha}_k)$. Lemma 2.2 implies that

$$\frac{\|x_k(\beta_k) - x_k\|^2}{\beta_k} \geqslant \beta_k \left(\frac{\|x_{k+1} - x_k\|}{\alpha_k}\right)\left(\frac{\|\bar{x}_{k+1} - x_k\|}{\bar{\alpha}_k}\right),$$

and since $\|x_{k+1}-x_k\| \geqslant \varepsilon\alpha_k$ and $\alpha_k \geqslant \gamma_2\bar{\alpha}_k$ for $k \in K_0$, we obtain that

$$\frac{\|x_k(\beta_k)-x_k\|^2}{\beta_k} \geqslant \varepsilon \min\{1, \gamma_2\}\|\bar{x}_{k+1}-x_k\|.$$

This inequality, together with (2.4) and (2.6), imply that for $k \in K_0$,

$$\begin{aligned}&\min\{(\nabla f(x_k), x_k - x_{k+1}), (\nabla f(x_k), x_k - \bar{x}_{k+1})\}\\ &\quad\geqslant \varepsilon \min\{1, \gamma_2\}\|\bar{x}_{k+1}-x_k\|.\end{aligned} \tag{2.7}$$

We now use (2.7) to obtain the desired contradiction. Since $\{(\nabla f(x_k), x_k - x_{k+1})\}$ converges to zero, (2.7) implies that $\{\|\bar{x}_{k+1}-x_k\|: k \in K_0\}$ converges to zero. Thus the uniform continuity of ∇f shows that if

$$\rho_k(\alpha) \equiv \frac{f(x_k)-f(x_k(\alpha))}{(\nabla f(x_k), x_k - x_k(\alpha))}$$

then

$$|\rho_k(\bar{\alpha}_k)-1| \leqslant \frac{o(\|\bar{x}_{k+1}-x_k\|)}{(\nabla f(x_k), x_k - \bar{x}_{k+1})}.$$

Hence (2.7) establishes that $\rho_k(\bar{\alpha}_k) > \mu_2$ for all $k \in K_0$ sufficiently large. This is the desired contradiction because (2.3) guarantees that $\rho_k(\bar{\alpha}_k) < \mu_2$. □

Note that in Theorem 2.3 the assumption that f is bounded below on Ω is only needed to conclude that $\{f(x_k)\}$ converges, and hence, that $\{(\nabla f(x_k), x_k - x_{k+1})\}$ converges to zero. The assumption that ∇f is uniformly continuous is used to conclude that

$$|f(\bar{x}_{k+1})-f(x_k)-(\nabla f(x_k), \bar{x}_{k+1}-x_k)| = o(\|\bar{x}_{k+1}-x_k\|).$$

These observations lead to the following result.

Theorem 2.4. *Let $f: R^n \to R$ be continuously differentiable on Ω, and let $\{x_k\}$ be the sequence generated by the gradient projection method defined by* (2.1) *and* (2.2). *If some subsequence $\{x_k: k \in K\}$ is bounded then*

$$\lim_{k \in K, k\to\infty} \frac{\|x_{k+1}-x_k\|}{\alpha_k} = 0.$$

Moreover, any limit point of $\{x_k\}$ is a stationary point of problem (1.1).

Proof. We assume that there is an infinite subsequence $K_0 \subset K$ such that

$$\frac{\|x_{k+1}-x_k\|}{\alpha_k} \geqslant \varepsilon > 0, \quad k \in K_0,$$

and reach a contradiction as in the proof of Theorem 2.3. The only difference is that since $\{x_k: k \in K\}$ is bounded, $\{f(x_k)\}$ converges and K_0 can be chosen so that $\{x_k: k \in K_0\}$ converges. Hence, continuity of ∇f is sufficient to show that $\rho_k(\bar{\alpha}_k) > \mu_2$ for all $k \in K_0$ sufficiently large.

Assume now that $\{x_k\}$ has a limit point x^*. A short computation shows that part (a) of Lemma 2.1 implies that for any $z \in \Omega$

$$\begin{aligned}\alpha_k(\nabla f(x_k), x_{k+1}-z) &\leq (x_{k+1}-x_k, z-x_{k+1})\\ &\leq (x_{k+1}-x_k, z-x_k) \leq \|x_{k+1}-x_k\| \|x_k-z\|.\end{aligned}$$

Hence,

$$(\nabla f(x_k), x_k-z) \leq (\nabla f(x_k), x_k-x_{k+1}) + \frac{\|x_{k+1}-x_k\|}{\alpha_k}\|x_k-z\|.$$

Since $\{f(x_k)\}$ converges, condition (2.1) implies that $\{(\nabla f(x_k), x_k-x_{k+1})\}$ converges to zero, and thus the above inequality shows that $(\nabla f(x^*), x^*-z) \leq 0$. This establishes that x^* is a stationary point of problem (1.1). □

Theorem 2.4 generalizes results of Bertsekas (1976) and Dunn (1981). Bertsekas shows that limit points of $\{x_k\}$ are stationary by assuming that Ω is the convex set (1.2) and that α_k is generated by the Armijo procedure (1.7, 1.8), while Dunn assumes that α_k satisfies (1.8) with $\alpha_k \geq \varepsilon$ for some $\varepsilon > 0$.

3. The projected gradient

We have shown that limit points of the gradient projection algorithm are stationary when α_k satisfies conditions (2.1) and (2.2). In this section we improve this result by showing that the sequence of projected gradients converges to zero provided the steps $\{\alpha_k\}$ are bounded.

The definition of the projected gradient requires some notions from convex analysis. As usual, we assume that Ω is a nonempty closed convex set in R^n and that $f: R^n \to R$ is continuously differentiable on Ω. A direction v is *feasible* at $x \in \Omega$ if $x + \tau v$ belongs to Ω for all $\tau > 0$ sufficiently small. The *tangent cone* $T(x)$ is defined as the closure of the cone of all feasible directions. The *projected gradient* $\nabla_\Omega f$ of f is defined by

$$\nabla_\Omega f(x) \equiv \operatorname{argmin}\{\|v+\nabla f(x)\|: v \in T(x)\}. \tag{3.1}$$

Since $T(x)$ is a nonempty closed convex set, this defines $\nabla_\Omega f(x)$ uniquely. Also note that for an arbitrary set Ω, the tangent cone at $x \in \Omega$ can also be defined as the set of all $v \in R^n$ such that

$$v = \lim_{k\to\infty} \frac{x_k - x}{\beta_k}$$

for some sequence $\{x_k\}$ in Ω converging to x, and some sequence of positive scalars $\{\beta_k\}$ converging to zero. It is not difficult to show that both definitions lead to the same tangent cone when Ω is convex.

McCormick and Tapia (1972) were led to the notion of a projected gradient by showing (also see Lemma 4.6 of Zarantonello (1971)) that if Ω is a closed convex set and $x \in \Omega$ then

$$\lim_{\alpha \to 0^+} \frac{P(x+\alpha d) - P(x)}{\alpha} = P_{T(x)}(d).$$

The projected gradient (3.1) is obtained if $d = -\nabla f(x)$. In the optimization literature the notion of a projected gradient is usually associated with affine subspaces. If for some matrix $C \in R^{n \times m}$ and $d \in R^m$

$$\Omega = \{x \in R^n \colon C^{\mathrm{T}} x = d\},$$

and if the columns of the matrix Z span the null space of C^{T}, then the tangent cone $T(x)$ is the range of Z, and the projected gradient for the l_2 norm is

$$\nabla_\Omega f(x) = -Z(Z^{\mathrm{T}} Z)^{-1} Z^{\mathrm{T}} g,$$

where g is the l_2 gradient of f. Some authors refer to $Z^{\mathrm{T}} g$ as a projected gradient, but the term *reduced gradient* is preferable because $Z^{\mathrm{T}} g$ is the gradient of f with respect to the reduced subspace *range*(Z), that is, $Z^{\mathrm{T}} g$ is the l_2 gradient of the mapping $f(x+Zv)$ at $v=0$. The projected gradient is also a reduced gradient because if Q is the orthogonal projection into *range*(Z) then $-\nabla_\Omega f(x)$ is the gradient of $f(x+Qv)$ at $v=0$.

Lemma 3.1. *Let $\nabla_\Omega f(x)$ be the projected gradient of f at $x \in \Omega$.*

(a) $-(\nabla f(x), \nabla_\Omega f(x)) = \|\nabla_\Omega f(x)\|^2$.

(b) $\min\{(\nabla f(x), v) \colon v \in T(x),\ \|v\| \le 1\} = -\|\nabla_\Omega f(x)\|$.

(c) *The point $x \in \Omega$ is a stationary point of problem* (1.1) *if and only if* $\nabla_\Omega f(x) = 0$.

Proof. We first establish (a). Since $T(x)$ is a cone, and since $\nabla_\Omega f(x)$ belongs to $T(x)$, part (a) of Lemma 2.1 shows that if $\lambda \ge 0$ then

$$(\nabla_\Omega f(x) + \nabla f(x), (\lambda - 1)\nabla_\Omega f(x)) \ge 0.$$

Setting $\lambda = 0$ and $\lambda = 2$ in this inequality yields (a). We prove (b) by noting that if $v \in T(x)$ and $\|v\| \le \|\nabla_\Omega f(x)\|$ then

$$\|\nabla_\Omega f(x) + \nabla f(x)\|^2 \le \|v + \nabla f(x)\|^2 \le \|\nabla_\Omega f(x)\|^2 + 2(\nabla f(x), v) + \|\nabla f(x)\|^2.$$

This inequality and (a) yield (b). To prove (c) note that if the point x is a stationary point then

$$(\nabla f(x), z - x) \ge 0, \quad z \in \Omega.$$

If v is a feasible direction at x then $x + \tau v$ belongs to Ω for some $\tau > 0$ and hence setting $z = x + \tau v$ yields $(\nabla f(x), v) \ge 0$. Since $T(x)$ is the closure of the set of all feasible directions, $(\nabla f(x), v) \ge 0$ for all $v \in T(x)$, and thus (b) implies that $\nabla_\Omega f(x) = 0$. Conversely, if $\nabla_\Omega f(x) = 0$ then (b) implies that $(\nabla f(x), v) \ge 0$ for all $v \in T(x)$, and since $z - x \in T(x)$ for any $z \in \Omega$, this shows that x is a stationary point of problem (1.1). □

Part (b) of Lemma 3.1 justifies the definition of the projected gradient by showing that $\nabla_\Omega f(x)$ is a steepest descent direction for f. This result will be used repeatedly in the convergence analysis of the gradient projection method.

At first glance part (c) of Lemma 3.1 suggests that an iterative scheme for problem (1.1) should drive $\nabla_\Omega f(x_k)$ to zero. However, this may not be possible because $\nabla_\Omega f$ can be bounded away from zero in a neighborhood of a stationary point. Thus, it is surprising that we can actually show that the sequence of projected gradients converges to zero.

Theorem 3.2. *Let $f: R^n \to R$ be continuously differentiable on Ω, and let $\{x_k\}$ be the sequence generated by the gradient projection method defined by* (2.1) *and* (2.2) *with*

$$\alpha_k \leq \gamma_3 \tag{3.2}$$

for some constant γ_3. If f is bounded below on Ω and ∇f is uniformly continuous on Ω then

$$\lim_{k\to\infty} \|\nabla_\Omega f(x_k)\| = 0.$$

Proof. Let $\varepsilon > 0$ be given and choose a feasible direction v_k with $\|v_k\| \leq 1$ such that

$$\|\nabla_\Omega f(x_k)\| \leq -(\nabla f(x_k), v_k) + \varepsilon.$$

Now note that part (a) of Lemma 2.1 shows that for any $z_{k+1} \in \Omega$

$$\alpha_k(\nabla f(x_k), x_{k+1} - z_{k+1}) \leq (x_{k+1} - x_k, z_{k+1} - x_{k+1}) \leq \|x_{k+1} - x_k\| \|x_{k+1} - z_{k+1}\|.$$

Since v_k is a feasible direction, $z_k = x_k + \tau_k v_k$ belongs to Ω for some $\tau_k > 0$. Thus the above inequality and Theorem 2.3 show that

$$\limsup_{k\to\infty} -(\nabla f(x_k), v_{k+1}) \leq 0.$$

Since α_k is bounded above, Theorem 2.3 shows that $\{\|x_{k+1} - x_k\|\}$ converges to zero, and thus we can use the uniform continuity of ∇f to conclude that

$$\limsup_{k\to\infty} -(\nabla f(x_{k+1}), v_{k+1}) \leq 0.$$

The choice of v_k guarantees that

$$\limsup_{k\to\infty} \|\nabla_\Omega f(x_{k+1})\| < \varepsilon,$$

and since $\varepsilon > 0$ is arbitrary, this proves our result. □

Theorem 2.4 states that any limit point of $\{x_k\}$ is a stationary point of problem (1.1). It is also possible to deduce this result from Theorem 3.2 by appealing to the result below.

Lemma 3.3. *If $f: R^n \to R$ is continuously differentiable on Ω then the mapping $\|\nabla_\Omega f(\cdot)\|$ is lower semicontinuous on Ω.*

Proof. We must show that if $\{x_k\}$ is an arbitrary sequence in Ω which converges to x then

$$\|\nabla_\Omega f(x)\| \leq \liminf_{k\to\infty} \|\nabla_\Omega f(x_k)\|.$$

Note that part (b) of Lemma 3.1 shows that for any $z \in \Omega$

$$(\nabla f(x_k), x_k - z) \leq \|\nabla_\Omega f(x_k)\| \|x_k - z\|.$$

Hence

$$(\nabla f(x), x - z) \leq \liminf_{k\to\infty} \|\nabla_\Omega f(x_k)\| \|x - z\|.$$

If v is a feasible direction at x with $\|v\| \leq 1$, then $x + \mu v$ belongs to Ω for some $\mu > 0$. The above inequality with $z = x + \mu v$ implies that

$$-(\nabla f(x), v) \leq \liminf_{k\to\infty} \|\nabla_\Omega f(x_k)\|.$$

Since $T(x)$ is the closure of all feasible directions, part (b) of Lemma 3.1 implies that

$$\|\nabla_\Omega f(x)\| \leq \liminf_{k\to\infty} \|\nabla_\Omega f(x_k)\|,$$

and this shows that $\|\nabla_\Omega f(\cdot)\|$ is lower semicontinuous on Ω. □

We have used Theorem 2.3 to derive Theorem 3.2. We can also derive Theorem 2.3 from Theorem 3.2 by first noting that part (b) of Lemma 3.1 implies that

$$(\nabla f(x_k), x_k - x_{k+1}) \leq \|\nabla_\Omega f(x_k)\| \|x_{k+1} - x_k\|.$$

Hence, inequality (2.5) yields

$$\frac{\|x_{k+1} - x_k\|}{\alpha_k} \leq \|\nabla_\Omega f(x_k)\|.$$

This estimate clearly shows that Theorem 3.2 implies Theorem 2.3.

We conclude this section by establishing a variation on Theorem 3.2. In this result the assumptions that f is bounded below on Ω and that ∇f is uniformly continuous on Ω are replaced by the assumption that some subsequence of $\{x_k\}$ is bounded.

Theorem 3.4. *Let $f: R^n \to R$ be continuously differentiable on Ω, and let $\{x_k\}$ be the sequence generated by the gradient projection method defined by* (2.1), (2.2), *and* (3.2). *If some subsequence $\{x_k: k \in K\}$ is bounded then*

$$\lim_{k\in K, k\to\infty} \|\nabla_\Omega f(x_{k+1})\| = 0.$$

Proof. Assume that there is an infinite subsequence $K_0 \subset K$ and an $\varepsilon_0 > 0$ such that

$$\|\nabla_\Omega f(x_{k+1})\| \geq \varepsilon_0, \quad k \in K_0. \tag{3.3}$$

Since $\{x_k : k \in K\}$ is bounded, we can choose K_0 so that $\{x_k : k \in K_0\}$ converges. The proof proceeds as in Theorem 3.2 except that we now appeal to Theorem 2.4 instead of Theorem 2.3. We thus obtain

$$\limsup_{k \in K_0, k \to \infty} \|\nabla_\Omega f(x_{k+1})\| \leq \varepsilon$$

for any $\varepsilon > 0$. This contradicts (3.3) for $\varepsilon < \varepsilon_0$, and thus establishes the result. □

4. Linear constraints—active and binding sets

We now turn to the application of the gradient projection method to linearly constrained problems, and prove in particular, that the gradient projection method identifies the active constraints in a finite number of iterations provided the algorithm converges to a nondegenerate stationary point. This result was established by Bertsekas (1976) for the Armijo procedure (1.7, 1.8) and for the convex set defined by the bound constraints (1.2), but we shall show that this result can be generalized considerably.

We assume that the convex set Ω is polyhedral, and that the mapping $f: R^n \to R$ is continuously differentiable on Ω. A polyhedral Ω can be defined in terms of a general inner product by setting

$$\Omega = \{x \in R^n : (c_j, x) \geq \delta_j, j = 1, \ldots, m\}, \tag{4.1}$$

for some vectors $c_j \in R^n$ and scalars δ_j. The set of *active constraints* is then

$$A(x) \equiv \{j : (c_j, x) = \delta_j\}. \tag{4.2}$$

An immediate application of these concepts is that the tangent cone of Ω is

$$T(x) = \{v \in R^n : (c_j, v) \geq 0, j \in A(x)\}.$$

We can also obtain an expression for the projected gradient in the polyhedral case by appealing to the classical result (see, for example, Lemma 2.2 of Zarantonello (1971)) that if K is a closed convex cone then every $x \in R^n$ can be expressed as

$$x = P_K(x) + P_{K^\circ}(x),$$

where the polar K° of K is the set of all $u \in R^n$ such that $(u, v) \leq 0$ for all $v \in K$. If we apply this result with K as the tangent cone of the polyhedral set (4.1) then the Farkas lemma shows that the polar of $T(x)$ is

$$T(x)^\circ = \left\{ v \in R^n : v = -\sum_{j \in A(x)} \lambda_j c_j, \lambda_j \geq 0 \right\}.$$

Hence, the decomposition of $-\nabla f(x)$ yields

$$\nabla_\Omega f(x) = -\nabla f(x) + \sum_{j \in A(x)} \lambda_j c_j,$$

where λ is a solution to the problem

$$\min\left\{\left\|\nabla f(x)-\sum_{j\in A(x)}\lambda_j c_j\right\|: \lambda_j\geq 0\right\}.$$

The polar of the tangent cone is also of importance in connection with optimality conditions. Note that $x^*\in\Omega$ is a stationary point of problem (1.1) if and only if

$$-\nabla f(x^*)\in T(x^*)^\circ.$$

If Ω is polyhedral, then the above expression for the polar of the tangent cone shows that $x^*\in\Omega$ is a stationary point of problem (1.1) if and only if x^* is a Kuhn-Tucker point, that is,

$$\nabla f(x^*)=\sum_{j\in A(x^*)}\lambda_j^* c_j,\quad \lambda_j^*\geq 0. \tag{4.3}$$

We are now ready to show that under suitable conditions the active set of a stationary point is identified in a finite number of iterations. We only consider stationary points that are well-defined by (4.3).

Definition. *A stationary point $x^*\in\Omega$ of problem* (1.1) *is nondegenerate if the active constraint normals $\{c_j: j\in A(x^*)\}$ are linearly independent and $\lambda_j^*>0$ for $j\in A(x^*)$.*

An important aspect of our next result is that it is not dependent on the method used to generate the sequence $\{x_k\}$ and only assumes that the sequence of projected gradient $\{\|\nabla_\Omega f(x_k)\|\}$ converges to zero.

Theorem 4.1. *Let $f: R^n\to R$ be continuously differentiable on Ω, and let $\{x_k\}$ be an arbitrary sequence in Ω which converges to x^*. If $\{\|\nabla_\Omega f(x_k)\|\}$ converges to zero and x^* is nondegenerate then $A(x_k)=A(x^*)$ for all k sufficiently large.*

Proof. First note that Lemma 3.3 and the above discussion implies that x^* is a Kuhn-Tucker point. Let us now prove that the active set settles down. Since $\{x_k\}$ converges to x^*, it is clear that $A(x_k)\subset A(x^*)$ for all k sufficiently large. Assume, however, that there is an infinite subsequence K and an index l such that $l\in A(x^*)$ but $l\notin A(x_k)$ for all $k\in K$. Let P be the (linear) projection into the space

$$\{v: (c_j, v)=0, j\in A(x^*), j\neq l\}.$$

The linear independence of the active constraint normals guarantees that $Pc_l\neq 0$. Moreover, $-Pc_l\in T(x_k)$ for all $k\in K$ because $l\notin A(x_k)$. Hence, part (b) of Lemma 3.1 implies that

$$(\nabla f(x_k), Pc_l)\leq\|\nabla_\Omega f(x_k)\|\,\|Pc_l\|,$$

and since $\{x_k\}$ converges to x^*,

$$(\nabla f(x^*), Pc_l)\leq 0.$$

On the other hand, since x^* is a Kuhn–Tucker point, (4.3) and the nondegeneracy assumption imply that

$$(\nabla f(x^*), Pc_l) = \lambda_l^* \|Pc_l\|^2 > 0.$$

This contradiction proves that $A(x_k) = A(x^*)$ for all k sufficiently large. □

Theorem 4.1 generalizes a result of Bertsekas (1976) in several ways. Bertsekas proves that if Ω is the convex set (1.2), then the set of active constraints is identified in a finite number of steps provided the gradient projection iterates $\{x_k\}$ converge to a local minimizer which satisfies the strict complementarity and second order sufficiency conditions. On the other hand, in Theorem 4.1 there is no need to assume the second order sufficiency conditions, and Ω is a general polyhedral set. We also note that Gafni and Bertsekas (1984) have obtained a result similar to Theorem 4.1 for a projection method of the form $x_{k+1} = P(x_k - \alpha_k g_k)$ for some vector g_k. However, assumption C of their result rules out the choice of $g_k = \nabla f(x_k)$ in their algorithm.

If the algorithm computes an estimate $\lambda(x)$ of the Lagrange multipliers, then it is also of interest to show that the set of *binding constraints*

$$B(x) \equiv \{j : j \in A(x), \lambda_j(x) \geq 0\} \tag{4.4}$$

is identified in a finite number of iterations. This result requires a mild restriction on the Lagrange multiplier estimates.

Definition. *A Lagrange multiplier estimate $\lambda(\cdot)$ is consistent if whenever $\{x_k\}$ converges to a nondegenerate Kuhn–Tucker point x^* with $A(x_k) \equiv A(x^*)$, then $\{\lambda(x_k)\}$ converges to $\lambda(x^*)$.*

Note that most practical multiplier estimates are consistent. For example, if $\lambda(x)$ is a solution of the problem

$$\min\left\{\left\|\nabla f(x) - \sum_{j \in A(x)} \lambda_j c_j\right\| : \lambda_j \in R\right\}, \tag{4.5}$$

then the linear independence assumption and the continuity of ∇f imply that $\lambda(\cdot)$ is consistent. Also note the following easy consequence of Theorem 4.1.

Theorem 4.2. *Let $f: R^n \to R$ be continuously differentiable on Ω, and let $\{x_k\}$ be an arbitrary sequence in Ω which converges to x^*. If the binding sets (4.4) are defined by a consistent multiplier estimate, if $\{\|\nabla_\Omega f(x_k)\|\}$ converges to zero, and if x^* is nondegenerate, then $B(x_k) = B(x^*)$ for all k sufficiently large.*

5. Extensions

The slow convergence rate of the gradient projection method is an obvious drawback of this method. As a first step in the development of an algorithm with a superlinear convergence rate, we extend the results of the previous sections to an algorithm which only uses the gradient projection method on selected iterations. We first assume that Ω is an arbitrary nonempty closed convex set, and later on we consider the linearly constrained case.

Algorithm 5.1. Let $x_0 \in \Omega$ be given. For $k \geq 0$ choose x_{k+1} by either (a) or (b):

(a) Let $x_{k+1} = P(x_k - \alpha_k \nabla f(x_k))$ where α_k satisfies (2.1), (2.2) and (3.2).

(b) Determine $x_{k+1} \in \Omega$ such that $f(x_{k+1}) \leq f(x_k)$.

A key observation is that the results obtained for the gradient projection algorithm can be generalized provided the iterates belong to the set K_{PG} of iterates k which generate x_{k+1} by the gradient projection algorithm. For example, if $\{x_k\}$ is generated by Algorithm 5.1, if f satisfies the assumptions of Theorem 2.4, and if K is an infinite subset of K_{PG}, then

$$\lim_{k \in K, k \to \infty} \frac{\|x_{k+1} - x_k\|}{\alpha_k} = 0. \tag{5.1}$$

The argument used to establish this result is the same as that used in the proof of Theorem 2.4; the only difference is that now $\{f(x_k)\}$ converges because we have required that $f(x_{k+1}) \leq f(x_k)$ for $k \notin K_{\mathrm{PG}}$. Many of the results of Sections 2 and 3 can be generalized by similar arguments, but for our purposes we only need the following extension of Theorem 3.4.

Theorem 5.2. *Let $f: R^n \to R$ be continuously differentiable on Ω, and let $\{x_k\}$ be the sequence generated by Algorithm 5.1. Assume that K_{PG} is infinite. If some subsequence $\{x_k: k \in K\}$ with $K \subset K_{\mathrm{PG}}$ is bounded then*

$$\lim_{k \in K, k \to \infty} \|\nabla_\Omega f(x_{k+1})\| = 0. \tag{5.2}$$

Moreover, any limit point of $\{x_k: k \in K_{\mathrm{PG}}\}$ is a stationary point of problem (1.1).

Proof. We only outline the proof of (5.2). Since (5.1) holds, the arguments used in the proofs of Theorem 3.2 and 3.4 yield

$$\limsup_{k \in K, k \to \infty} -(\nabla f(x_{k+1}), v_{k+1}) \leq 0.$$

Since v_{k+1} is any feasible vector with $\|v_{k+1}\| \leq 1$, this implies that for any $\varepsilon > 0$

$$\limsup_{k \in K, k \to \infty} \|\nabla_\Omega f(x_{k+1})\| \leq \varepsilon.$$

Hence (5.2) holds as desired. For the reminder of the proof assume that $\{x_k: k \in K\}$ converges to x^* for some $K \subset K_{\mathrm{PG}}$. Since $\{\alpha_k: k \in K\}$ is bounded, (5.1) implies that $\{x_{k+1}: k \in K\}$ also converges to x^*. Hence, (5.2) and Lemma 3.3 imply that x^* is a stationary point. □

For linearly constrained problems the choice of x_{k+1} in step (b) of Algorithm 5.1 is usually such that $A(x_k) \subset A(x_{k+1})$, where the active set $A(\cdot)$ is defined by (4.2). In a typical case x_{k+1} is obtained by choosing a descent direction v_k orthogonal to the constraints in $A(x_k)$ and doing a line search along v_k. If the line search algorithm finds a sufficient reduction for some α_k in $(0, \sigma_k)$ where σ_k is the distance along v_k to the closest constraint not in $A(x_k)$, then $x_{k+1} = x_k + \alpha_k v_k$. In this case $A(x_{k+1}) = A(x_k)$. Otherwise $x_{k+1} = x_k + \sigma_k v_k$; in this case $A(x_k)$ is a subset of $A(x_{k+1})$. These considerations lead to the following modification to step (b) of Algorithm 5.1.

Algorithm 5.3. Let $x_0 \in \Omega$ be given. For $k \geq 0$ choose x_{k+1} by either (a) or (b):
(a) Let $x_{k+1} = P(x_k - \alpha_k \nabla f(x_k))$ where α_k satisfies (2.1), (2.2) and (3.2).
(b) Determine $x_{k+1} \in \Omega$ such that $f(x_{k+1}) \leq f(x_k)$ and $A(x_k) \subset A(x_{k+1})$.

The main reason for introducing Algorithm 5.3 is to motivate the study of algorithms for linearly constrained problems which use the gradient projection method on selected iterations. There are other possible variations on this algorithm, but this version is sufficient for our purposes.

Theorem 5.4. *Let $f: R^n \to R$ be continuously differentiable on a polyhedral Ω, and assume that the stationary points of f are nondegenerate. If the sequence $\{x_k\}$ generated by Algorithm* 5.3 *is bounded, then $A(x_k) = A$ for some active set A and all k sufficiently large.*

Proof. We show that $A(x_k)$ is a subset of $A(x_{k+1})$ for all k sufficiently large. If, on the contrary, there is an infinite sequence K such that $A(x_k)$ is not contained in $A(x_{k+1})$ then K must be a subset of K_{PG}. Choose an infinite subset K_0 of K such that $\{x_k: k \in K_0\}$ converges to some x^*. Theorem 5.2 shows that x^* is stationary, and by assumption, x^* is nondegenerate. Hence, the convergence of $\{x_k: k \in K_0\}$ together with (5.2) and Theorem 4.1 shows that

$$A(x_k) \subset A(x^*) = A(x_{k+1})$$

for all $k \in K_0$ sufficiently large. This contradiction establishes that $A(x_k)$ is a subset of $A(x_{k+1})$, and proves our result. □

Since Algorithm 5.3 allows the choice of $x_{k+1} = x_k$ for $k \notin K_{\text{PG}}$, interesting convergence results can only be obtained by making further assumptions on the algorithm used for $k \notin K_{\text{PG}}$. In the remainder of this paper we examine the case where f is a quadratic function.

6. Quadratic programming

The extensions of the gradient projection method that we have developed have applications in many areas. In this section we use these ideas to develop an algorithm for quadratic programming problems.

Given a quadratic function $f: R^n \to R$ defined on a polyhedral Ω, there are several ways to determine an x_{k+1} which satisfies the conditions of step (b) in Algorithm 5.3. Most of the approaches used in this calculation require a matrix Z_k whose columns form a basis for the subspace of vectors v with $(c_j, v) = 0$ for $j \in A(x_k)$, and define x_{k+1} in terms of the quadratic

$$q_k(w) = f(x_k + Z_k w). \tag{6.1}$$

Note that $\nabla q_k(0)$ and $\nabla^2 q_k(0)$ are, respectively, the reduced gradient and the reduced Hessian matrix of f at x_k.

We first consider the case where $\nabla q_k(0) = 0$ and the Hessian matrix $\nabla^2 q_k(0)$ is positive semidefinite. In this situation $w = 0$ is a global minimizer of the quadratic q_k and thus either x_k is stationary or a different active set must be considered in order to reduce f.

If the Hessian matrix $\nabla^2 q_k(0)$ is positive definite then

$$w_k = -\nabla^2 q_k(0)^{-1} \nabla q_k(0)$$

is a global minimizer of q_k. Moreover, if $v_k = Z_k w_k$ then v_k is a feasible descent direction, and thus we can compute the largest α_k in $(0, 1]$ such that $x_{k+1} = x_k + \alpha_k v_k$ satisfies the conditions of step (b) in Algorithm 5.3. Also note that if $\nabla q_k(0) \neq 0$ then $f(x_{k+1}) < f(x_k)$, and that if $\alpha_k < 1$ then at least one new constraint is added to the active set.

Consider now the case where the Hessian matrix $\nabla^2 q_k(0)$ is indefinite. In this case we choose w_k such that

$$(\nabla^2 q_k(0) w_k, w_k) < 0,$$

and pick the sign of w_k so that $(\nabla q_k(0), w_k) \leq 0$. If $v_k = Z_k w_k$ then v_k is a feasible descent direction and thus we can compute the largest positive α_k such that $x_{k+1} = x_k + \alpha_k v_k$ satisfies the conditions of step (b) in Algorithm 5.3. If f is bounded below on Ω then there is an α_k which satisfies these conditions.

The only remaining case occurs if $\nabla q_k(0) \neq 0$ and $\nabla^2 q_k(0)$ is positive semidefinite and singular. In this case we can choose any w_k such that $(\nabla q_k(0), w_k) < 0$, and choose v_k and α_k as in the case where the Hessian is indefinite.

Standard algorithms for the general quadratic programming problem (see, for example, Gill, Murray and Wright (1981), and Fletcher (1981)) use the ideas outlined above. At each iteration there is a *working* set W_k of constraints such that $W_k \subset A(x_k)$. If x_k is not a global minimizer of the problem

$$\min\{f(x): (c_j, x) = \delta_j, j \in W_k\}, \tag{6.2}$$

then it is possible to compute an $x_{k+1} \in \Omega$ such that $f(x_{k+1}) \leq f(x_k)$ and $W_k \subset W_{k+1}$. Moreover, if $W_{k+1} = W_k$ then x_{k+1} solves (6.2). If x_k is a global minimizer of problem (6.2) then either x_k is stationary or a different working set must be considered in order to decrease f. The new working set is obtained by computing Lagrange multipliers at x_k. If the Lagrange multipliers are nonnegative then x_k is stationary;

otherwise a new working set is obtained by dropping a constraint associated with a negative Lagrange multiplier. If $W_k = A(x_k)$ and the constraints in $A(x_k)$ are linearly independent, then dropping this constraint leads to a strictly lower function value.

The following algorithm is quite similar to standard quadratic programming algorithms, but uses the gradient projection method to predict the new working set. In this algorithm, and in the remainder of this paper, W_k is only required to be a subset of $A(x_k)$. Also note that although x_0 is required to be in Ω, it is always possible to use the gradient projection algorithm to produce a starting point in Ω.

Algorithm 6.1. Let $f: R^n \to R$ be a quadratic function and let $x_0 \in \Omega$ be given. For $k \geq 0$ choose x_{k+1} as follows:

(a) If x_k is a global minimizer of (6.2), let $x_{k+1} = P(x_k - \alpha_k \nabla f(x_k))$ where α_k satisfies (2.1), (2.2) and (3.2).

(b) Otherwise choose $x_{k+1} \in \Omega$ such that $f(x_{k+1}) \leq f(x_k)$ and $W_k \subset W_{k+1}$. If $W_{k+1} = W_k$ then x_{k+1} must be a global minimizer of (6.2).

An alternative and perhaps simpler view of Algorithm 6.1 is obtained by noting that step (b) is used until the algorithm computes a global minimizer of a problem

$$\min\{f(x): (c_j, x) = \delta_j, j \in S_k\} \tag{6.3}$$

with $W_k \subset S_k$. Thus, it is possible to renumber the iterates so that step (b) produces an $x_{k+1} \in \Omega$ such that $f(x_{k+1}) \leq f(x_k)$ and x_{k+1} is a global minimizer of (6.3).

We have discussed one possible implementation of Algorithm 6.1 when $W_k = A(x_k)$ and the iterates are obtained by the minimization of the quadratic q_k defined by (6.1). It is also possible to base the algorithm on the minimization of q_k when Z_k is a basis for the constraints in a working set W_k in $A(x_k)$. For example, at the start of step (b) we could choose $W_k = B(x_k)$ where the binding set $B(x_k)$ is defined by (4.4). The only difference is that now the directions v_k may not be feasible and thus a step $\alpha_k = 0$ would be required. This situation is handled by adding a constraint to W_k and repeating the process. Eventually a feasible descent direction is obtained.

Theorem 6.2. *If $f: R^n \to R$ is a quadratic function bounded below on a polyhedral Ω, then Algorithm* 6.1 *terminates at some iterate x_l which is stationary.*

Proof. Note that step (b) can only be executed a finite number of consecutive times because this step generates a nested sequence of working sets and thus a solution to problem (6.2) is eventually generated. Moreover, step (a) can only be executed a finite number of times because there are a finite number of working sets W_k and $f(x_{k+1}) < f(x_k)$ whenever x_{k+1} is produced by step (a). □

In contrast to standard quadratic programming algorithms, Algorithm 6.1 does not require a nondegeneracy assumption or an anti-cycling rule in order to establish finite termination of the algorithm. On the other hand, the computation of the projections required by step (a) can be quite costly unless the set Ω is relatively simple. We are particularly interested in sets Ω for which the projections can be computed with order n operations. This is the case, for example, if Ω is the bound constrained set (1.2), or more generally, if Ω is of the form

$$\Omega = \{x \in R^n \colon l_i \leqslant (c_i, x) \leqslant u_i,\ 1 \leqslant i \leqslant m\},$$

and the constraint normals are such that if $i \neq j$ then the vectors c_i and c_j do not have a nonzero in the same position.

Algorithm 6.1 allows the use of an iterative method to determine x_{k+1} in step (b) of Algorithm 6.1. For example, if f is a strictly convex quadratic then the conjugate gradient method on the quadratic q_k defined by (6.1) generates an iterate w_k such that $x_k + Z_k w_k$ belongs to Ω and either $x_k + Z_k w_k$ solves problem (6.2), or there is a direction d_k such that $q_k(w_k + \alpha d_k)$ is a strictly decreasing function of α in the set

$$\Lambda_k \equiv \{\alpha \geqslant 0 \colon (x_k + Z_k(w_k + \alpha d_k)) \in \Omega\}.$$

The latter case arises when the step of the conjugate gradient method does not belong to Λ_k. If α_k is the largest α in Λ_k then

$$x_{k+1} = x_k + Z_k(w_k + \alpha_k d_k) \tag{6.4}$$

satisfies the conditions of step (b) and $A(x_{k+1})$ has at least one more constraint than $A(x_k)$.

The use of the conjugate gradient method with Algorithm 6.1 is closely related to the algorithms of Polyak (1969), and O'Leary (1980) for strictly convex quadratic programming problems subject to bound constraints. These algorithms can be described in terms of Algorithm 6.1 but with a different choice of x_{k+1} in step (a). Let Ω be the bound constrained set (1.2) and consider the set

$$B(x) = \{i \colon x_i = l_i \text{ and } \partial_i f(x) \geqslant 0, \text{ or } x_i = u_i \text{ and } \partial_i f(x) \leqslant 0\}. \tag{6.5}$$

Note that this is the binding set defined by (4.4) when the multipliers are chosen by the first order estimate (4.5). The initial working set W_0 is $B(x_0)$. Now assume that for some iterate $x_k \in \Omega$ the working set W_k is $B(x_k)$. If x_k is not a global minimizer of (6.2) then the conjugate gradient method is used as outlined above. Thus constraints are added to the working set, usually one at a time, until a solution to problem (6.2) is generated. If x_k is a global minimizer of (6.2) then the algorithm sets $x_{k+1} = x_k$ and $W_{k+1} = B(x_{k+1})$. Although the choice of $x_{k+1} = x_k$ in step (a) is not necessarily the best choice, it does lead to a complete change in the working set W_k. Of course, the same effect is obtained by choosing x_{k+1} by the gradient projection method and then using (6.5) to define the next working set.

Algorithm 6.1 is also related to the projected conjugate gradient algorithms of Dembo and Tulowitzki (1983) for strictly convex quadratic programming problems subject to bound constraints. Dembo and Tulowitzki use a projected gradient method of the form

$$x_{k+1} = P(x_k + \alpha_k p_k) \tag{6.6}$$

where the ith component of p_k is set to zero if i is a binding constraint and otherwise set to $-\partial_i f(x_k)$. Dembo and Tulowitzki choose the step α_k by the Armijo procedure (1.7) but with the sufficient decrease condition (1.8) replaced by

$$f(x_{k+1}) \le f(x_k) + \mu\alpha_k(\nabla f(x_k), p_k). \tag{6.7}$$

For a given α_k the iterate x_{k+1} generated by the gradient projection method (1.4) agrees with (6.6). However, condition (6.7) differs considerably from (1.8), and does not seem to be fully supported by theory.

It is also possible to use the conjugate gradient method in conjunction with Algorithm 6.1 even if f is not a strictly convex quadratic, but this requires a modification to Algorithm 6.1. If the conjugate gradient method solves problem (6.2) or generates a step that does not belong to Λ_k, then x_{k+1} can be generated as in the strictly convex case. Otherwise the conjugate gradient method generates an iterate w_k such that $x_k + Z_k w_k$ belongs to Ω and either $\nabla q_k(w_k) = 0$, or

$$(\nabla q_k(w_k), d_k) < 0 \quad \text{and} \quad (\nabla^2 q_k(w_k) d_k, d_k) \le 0$$

for some direction d_k. In the latter case q_k is strictly decreasing and unbounded below on the ray $w_k + \alpha d_k$, and thus either f is unbounded below on Ω, or x_{k+1} can be defined by (6.4) where the largest α in Λ_k is a suitable value for α_k.

Now consider the case where $\nabla q_k(w_k) = 0$. In this case $x_k + Z_k w_k$ is a stationary point of q_k but not necessarily a global minimizer. This situation can be handled by executing step (a) whenever x_k is a stationary point of problem (6.2), and only requiring that x_{k+1} be a stationary point of problem (6.2) whenever $W_{k+1} = W_k$.

Algorithm 6.3. Let $f: R^n \to R$ be a quadratic function and let $x_0 \in \Omega$ be given. For $k \ge 0$ choose x_{k+1} as follows:

(a) If x_k is a stationary point of (6.2), let $x_{k+1} = P(x_k - \alpha_k \nabla f(x_k))$ where α_k satisfies (2.1), (2.2) and (3.2).

(b) Otherwise choose $x_{k+1} \in \Omega$ such that $f(x_{k+1}) \le f(x_k)$ and $W_k \subset W_{k+1}$. If $W_{k+1} = W_k$ then x_{k+1} must be a stationary point of (6.2).

If we accept the claim that all stationary points of a quadratic have the same function value, then the same argument used to prove Thoerem 6.2 shows that Algorithm 6.3 terminates in a finite number of iterations. It is not difficult to establish the claim that all stationary points of a quadratic have the same function value. Just note that if z_1 and z_2 are stationary points of a quadratic q, then the function

$$\phi(\alpha) = q(z_1 + \alpha(z_2 - z_1))$$

is a quadratic whose derivative vanishes at $\alpha = 0$ and at $\alpha = 1$. Thus ϕ is constant, and therefore $q(z_2) = q(z_1)$.

A disadvantage of the strategy in Algorithm 6.3 is that step (b) continues to be executed until a stationary point of problem (6.3) is generated. An alternative strategy would allow for the use of the gradient projection method to predict a new working set. The following algorithm formalizes this strategy and specifies when the gradient projection method is used.

Algorithm 6.4. Let $f: R^n \to R$ be a quadratic function and let $x_0 \in \Omega$ be given. For $k \geq 0$ choose x_{k+1} as follows:

(a) If x_k is a stationary point of (6.2), or if $W_{k-1} \neq W_k$, let $x_{k+1} = P(x_k - \alpha_k \nabla f(x_k))$ where α_k satisfies (2.1), (2.2) and (3.2).

(b) Otherwise choose $x_{k+1} \in \Omega$ such that $f(x_{k+1}) \leq f(x_k)$ and $W_k \subset W_{k+1}$. If $W_{k+1} = W_k$ then x_{k+1} must be a stationary point of (6.2).

The motivation for this algorithm is that the gradient projection method is used to identify a working set that is worth exploring. If the gradient projection method produces an x_{k+1} with $W_{k+1} = W_k$, then this working set is explored by (b).

Theorem 6.5. *Let $f: R^n \to R$ be a quadratic function bounded below on a polyhedral Ω, and let $\{x_k\}$ be the sequence generated by Algorithm 6.4. Either the algorithm terminates at some iterate x_l which is stationary, or any limit point of $\{x_k\}$ is stationary.*

Proof. Assume that some subsequence $\{x_k: k \in K\}$ converges to x^*, and consider the method used to generate x_{k+1}. If there is an infinite sequence $K_0 \subset K$ with $K_0 \subset K_{PG}$ then Theorem 5.2 shows that x^* is stationary. The other possibility is that $W_{k-1} = W_k$ for all $k \in K$. Now consider the method used to generate x_k. If an infinite number of x_k are generated by the gradient projection then (5.2) and Lemma 3.3 imply that x^* is stationary. On the other hand, if x_k is generated by (b) in Algorithm 6.4 then x_k must be a stationary point of problem (6.2). However, since there is a finite number of working sets, and since the gradient projection method guarantees a strict decrease in function value, the algorithm must terminate after a finite number of iterations. □

The finite termination properties of Algorithm 6.4 require a restriction on the working sets. The simplest restriction is to assume that W_k is the active set at x_k.

Theorem 6.6 *Let $f: R^n \to R$ be a quadratic function bounded below on a polyhedral Ω, and assume that the stationary points of f are nondegenerate. If the sequence $\{x_k\}$ generated by Algorithm 6.4 with $W_k = A(x_k)$ is bounded, then the algorithm terminates at some iterate x_l which is stationary.*

Proof Theorem 5.4 shows that $W_{k+1} = W_k$ for all k sufficiently large. Hence, x_{k+1} is a stationary point of problem (6.2). However, this can only happen a finite number of times because $f(x_{k+1}) < f(x_k)$ whenever x_{k+1} is produced by the gradient projection method. □

The finite termination of Algorithm 6.4 can also be established when W_k is not defined by $A(x_k)$ provided one can show that $W_k \subset W_{k+1}$ for all k sufficiently large. For example, assume that $\lambda(\cdot)$ is a consistent Lagrange multiplier estimate, and that

$$B(x_k) \subset W_k \subset A(x_k). \tag{6.8}$$

We now show that eventually $W_k \subset W_{k+1}$. The argument used in the proof of Theorem 5.4 shows that if there is an infinite sequence K such that W_k is not contained in W_{k+1} then there is an infinite subset K_0 of K and a stationary point x^* such that $\{x_k \colon k \in K_0\}$ converges to x^* and $\{\|\nabla_\Omega f(x_{k+1})\|;\, k \in K_0\}$ converges to zero. Since $\{x_k \colon k \in K_0\}$ converges to x^*,

$$W_k \subset A(x_k) \subset A(x^*)$$

for all $k \in K_0$ sufficiently large. Moreover, since $\{\|\nabla_\Omega f(x_{k+1})\| \colon k \in K_0\}$ converges to zero, Theorem 4.2 shows that

$$B(x^*) = B(x_{k+1}) \subset W_{k+1}.$$

Since $A(x^*) = B(x^*)$ by definition, we have shown that $W_k \subset W_{k+1}$ for all $k \in K_0$ sufficiently large. This contradiction establishes that $W_k \subset W_{k+1}$ for all sufficiently large k, and proves that Algorithm 6.5 terminates in a finite number of steps when W_k satisfies (6.8).

7. Concluding remarks

We have presented algorithms for linearly constrained problems which use the gradient projection method to choose the active set. These algorithms are able to drop and add many constraints from the active set and thus are attractive for large scale problems. Note, however, that the computational efficiency of these algorithms depends on the cost of computing the projection into the feasible set. Future work will consider, in particular, the efficient computation of these projections, and the extension of the quadratic programming results to general linearly constrained problems.

Acknowledgment

We would like to thank Danny Sorensen for suggesting the notation for the projected gradient. We also acknowledge the helpful comments of Dimitri Bertsekas and Joe Dunn.

References

D.P. Bertsekas, "On the Goldstein-Levitin-Polyak gradient projection method," *IEEE Transactions on Automatic Control* 21 (1976) 174–184.

D.P. Bertsekas, "Projected Newton methods for optimization problems with simple constraints," *SIAM Journal on Control and Optimization* 20 (1982) 221-246.

J.V. Burke and J.J. Moré, "On the identification of active constraints," Argonne National Laboratory, Mathematics and Computer Science Division Report ANL/MCS-TM-82, (Argonne, IL, 1986).

R.S. Dembo and U. Tulowitzki, "On the minimization of quadratic functions subject to box constraints," Working Paper Series B #71, School of Organization and Management, Yale University (New Haven, CT, 1983).

R. S. Dembo and U. Tulowitzki "Local convergence analysis for successive inexact quadratic programming methods," Working Paper Series B #78, School of Organization and Management, Yale University, (New Haven, CT, 1984).

R.S. Dembo and U. Tulowitzki, "Sequential truncated quadratic programming methods," in: P.T. Boggs, R.H. Byrd and R.B. Schnabel, eds., *Numerical Optimization 1984* (Society of Industrial and Applied Mathematics, Philadelphia, 1985) pp. 83-101.

J.C. Dunn, "Global and asymptotic convergence rate estimates for a class of projected gradient processes," *SIAM Journal on Control and Optimization* 19 (1981) 368-400.

J.C. Dunn, "On the convergence of projected gradient processes to singular critical points," *Journal of Optimization Theory and Applications* (1986) to appear.

R. Fletcher, *Practical Methods of Optimization Volume 2: Constrained Optimization* (John Wiley & Sons, New York, 1981).

E.M. Gafni and D.P. Bertsekas, "Convergence of a gradient projection method," Massachusetts Institute of Technology, Laboratory for Information and Decision Systems Report LIDS-P-1201 (Cambridge, Massachusetts, 1982).

E.M. Gafni and D.P. Bertsekas, "Two-metric projection methods for constrained optimization," *SIAM Journal on Control and Optimization* 22 (1984) 936-964.

P.E. Gill, W. Murray and M.H. Wright, *Practical Optimization* (Academic Press, New York, 1981).

A.A. Goldstein, "Convex programming in Hilbert space," *Bulletin of the American Mathematical Society* 70 (1964) 709-710.

E.S. Levitin and B.T. Polyak, "Constrained minimization problems," *USSR Computational Mathematics and Mathematical Physics* 6 (1966) 1-50.

G.P. McCormick and R.A. Tapia, "The gradient projection method under mild differentiability conditions," *SIAM Journal on Control* 10 (1972) 93-98.

D.P. O'Leary, "A generalized conjugate gradient algorithm for solving a class of quadratic programming problems," *Linear Algebra and its Applications* 34 (1980) 371-399.

R.R. Phelps "The gradient projection method using Curry's steplength," *SIAM Journal on Control and Optimization* 24 (1986) 692-699.

B.T. Polyak, "The conjugate gradient method in extremal problems," *USSR Computational Mathematics and Mathematical Physics* 9 (1969) 94-112.

E.H. Zarantonello, "Projections on convex sets in Hilbert space and spectral theory," in: E.H. Zarantonello, ed., *Contributions to Nonlinear Functional Analysis* (Academic Press, New York, 1971).