SIAM J. OPTIMIZATION
Vol. 3, No. 1, pp. 43–59, February 1993

ERROR BOUND AND REDUCED-GRADIENT PROJECTION ALGORITHMS FOR CONVEX MINIMIZATION OVER A POLYHEDRAL SET*

ZHI-QUAN LUO† AND PAUL TSENG‡

Abstract. Consider the problem of minimizing, over a polyhedral set, the composition of an affine mapping with a strongly convex differentiable function. The polyhedral set is expressed as the intersection of an affine set with a (simpler) polyhedral set and a new local error bound for this problem, based on projecting the reduced gradient associated with the affine set onto the simpler polyhedral set, is studied. A class of reduced-gradient projection algorithms for solving the case where the simpler polyhedral set is a box is proposed and this bound is used to show that algorithms in this class attain a linear rate of convergence. Included in this class are the gradient projection algorithm of Goldstein and Levitin and Poljak, and an algorithm of Bertsekas. A new algorithm in this class, reminiscent of active set algorithms, is also proposed. Some of the results presented here extend to problems where the objective function is extended real valued and to variational inequality problems.

Key words. local error bound, convex minimization, linear convergence, reduced-gradient projection algorithms

AMS(MOS) subject classifications. 49, 90

1. Introduction.

We consider the convex program

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && x \in \mathcal{X}, \end{aligned} \tag{1.1}$$

where $\mathcal{X}$ is a polyhedral set in the n-dimensional Euclidean space $\Re^n$ and f is a real-valued function defined on $\Re^n$. We assume that f is of the special form

$$f(x) = g(Ex) + \langle q, x \rangle, \tag{1.2}$$

where E is some $m \times n$ matrix, q is some vector in $\Re^n$, and g is a continuously differentiable function in $\Re^m$ with ∇g Lipschitz continuous and strongly monotone in the sense that there exist positive scalars $\rho > 0$ and $\sigma > 0$ such that

$$\|\nabla g(z) - \nabla g(w)\| \leq \rho \|z - w\| \quad \forall z, \quad \forall w, \tag{1.3}$$

and

$$\langle \nabla g(z) - \nabla g(w), z - w \rangle \geq \sigma \|z - w\|^2 \quad \forall z, \quad \forall w. \tag{1.4}$$

We also assume that the optimal solution set for (1.1), denoted by $\mathcal{X}^*$, is nonempty and denote by v^* the value of f on $\mathcal{X}^*$. In our notation, all vectors are column vectors, superscript T denotes matrix transpose, $\langle \cdot, \cdot \rangle$ denotes the usual Euclidean inner product, and $\| \cdot \|$ denotes the Euclidean norm induced by $\langle \cdot, \cdot \rangle$.

* Received by the editors January 3, 1991; accepted for publication (in revised form) November 18, 1991.

† Communications Research Laboratory, Room 225, McMaster University, Hamilton, Ontario, L8S 4K1, Canada (luozq@sscvax.cis.mcmaster.ca). The research of this author is supported by Natural Sciences and Engineering Research Council of Canada grant OPG0090391.

‡ Department of Mathematics, GN-50, University of Washington, Seattle, Washington 98195 (tseng@math.washington.edu).

There are many optimization problems that satisfy the above assumptions, including convex quadratic programs and a certain routing problem in data networks (see [BeG87]). We remark that the assumption that g be real valued is made only to simplify the analysis and can be relaxed so as to allow, for example, certain entropy optimization problems and their dual to be captured by the problem framework. (See §6 for detailed discussions.)

A classical method for solving (1.1) is the *gradient projection* algorithm of Goldstein [Gol64] and Levitin and Poljak [LeP65], which follows each gradient step by a projection onto the feasible set $\mathcal{X}$:

$$x := [x - \alpha \nabla f(x)]^+_{\mathcal{X}},$$

where $[\cdot]^+_{\mathcal{X}}$ denotes the orthogonal projection onto $\mathcal{X}$ and α is some suitably chosen positive stepsize. This method has been well studied and, when combined with second-order scaling, has been successful in solving large quadratic programs with box constraints (see, e.g., [Ber76], [Ber82], [GaB84], and [Mor89]). However, when $\mathcal{X}$ is not a box, the projection $[\cdot]^+_{\mathcal{X}}$ cannot be easily computed and this method can suffer from poor performance.

For the special case where $\mathcal{X}$ is the Cartesian product of simplices, Bertsekas proposed a modification of the gradient projection algorithm which avoids the relatively expensive operation of projecting onto the simplices (see [Ber80], [Ber82], [BeG83], and [BeG87]). (A simplex in $\Re^n$ is a set of the form $\{x \in \Re^n \mid \sum_i x_i = c,\ x \geq 0\}$ for some $c > 0$.) Instead, the algorithm of Bertsekas moves an iterate opposite the direction of a certain *reduced* gradient associated with the knapsack constraints and follows this step with a projection onto the nonnegative orthant. This algorithm has been successfully applied to solving a certain routing problem in data networks (see [BeG83], [BeG87], and [BeT89]) and can even be implemented in a distributed asynchronous manner (see [Tsa89] and [TsB86]).

A key question concerns the convergence and the rate of convergence of the above algorithms. For the gradient projection algorithm this question is largely resolved. It was shown by Bertsekas and Gafni [BeG82], in the more general context of variational inequality problems, and rediscovered by Luo and Tseng [LuT92b], that the gradient projection algorithm for solving (1.1) attains a linear rate of convergence, provided that the stepsize α is suitably chosen. Similar results were obtained by Dunn [Dun81], [Dun87] and Gawande and Dunn [GaD88] for the general problem of minimizing a differentiable function over a closed convex set, but under an additional assumption that all local minimizers are isolated and that the objective function satisfies a certain local growth condition. Central to their analysis is a certain local *error bound* for estimating the distance from a point $x \in \mathcal{X}$ to $\mathcal{X}^*$, defined as

$$\phi(x) = \min_{x^* \in \mathcal{X}^*} \|x - x^*\|. \tag{1.5}$$

In particular, it was shown in [LuT92b] that $\phi(x)$ can be bounded above by some constant times

$$\|x - [x - \nabla f(x)]^+_{\mathcal{X}}\|,$$

the norm of the "natural residual" at x, provided that the latter quantity is small. The same local error bound also extends to affine variational inequality problems (see [Rob81] and [LuT92c]) and holds globally if f is strongly convex [Pan87]. For the Bertsekas algorithm, however, no comparable result was known. We remark that

bounds for ϕ have been studied quite extensively, although the focus has been on global bounds and on using the bounds to terminate iterative algorithms and to extract sensitivity/stability information near the optimal solution set (see [MaS87], [MaD88], [Pan87], and [Rob82]).

The goals of this paper are twofold. First, we propose a generalization of the above error bound based on a certain decomposition of the polyhedral set $\mathcal{X}$. More specifically, let us express $\mathcal{X}$ as

$$\mathcal{X} = \mathcal{C} \cap \{x \in \Re^n \mid Bx = c\}, \tag{1.6}$$

for some (simpler) polyhedral set $\mathcal{C} \in \Re^n$, some $l \times n$ matrix B, and some vector c in $\Re^l$. We will show that $\phi(x)$ can be bounded above by some constant times

$$\|x - [x - \nabla f(x) + B^T p]^+_{\mathcal{C}}\| + \|Bx - c\|, \tag{1.7}$$

for any $x \in \mathcal{C}$ and any $p \in \Re^l$ for which the above quantity is "sufficiently" small. Here $[\cdot]^+_{\mathcal{C}}$ denotes the orthogonal projection onto $\mathcal{C}$. Some obvious advantages of this new local error bound, relative to the earlier one, are (i) x is only required to be in $\mathcal{C}$, not $\mathcal{X}$, and (ii) instead of projecting onto $\mathcal{X}$, we project onto the simpler set $\mathcal{C}$.

Second, we propose a class of feasible descent algorithms for solving the special case of (1.1) where $\mathcal{C}$ is a box. At each iteration of these algorithms, we compute a z according to the projection step

$$z := [x - \alpha(\nabla f(x) - B^T p)]^+_{\mathcal{C}},$$

for some stepsize $\alpha > 0$ and some multiplier vector p, and then adjust a subset of the coordinates of z to obtain a new iterate in $\mathcal{X}$. Both the gradient projection algorithm and the algorithm of Bertsekas described earlier can be shown to belong to this class. By using the new local error bound, we show that the iterates generated by any algorithm in this class converge at least linearly to an optimal solution. (Here and throughout, by linear convergence we mean R-linear convergence in the sense of Ortega and Rheinboldt [OrR70].) We also propose a new algorithm in this class reminiscent of active set algorithms.

The remainder of this paper is organized as follows. In §2 we prove some technical facts concerning the problem (1.1); in §3 we use these facts to establish the new local error bound. In §4, we describe the class of feasible descent algorithms mentioned above and relate them to the gradient projection algorithm and to the algorithm of Bertsekas. In §5, we use the error bound of §3 to show that any algorithm in this class which uses an Armijo-like stepsize rule is linearly convergent. In §6, we give our conclusion and discuss extensions.

Throughout this paper, we adhere to the following notations. For any vector x in $\Re^k$, we denote by x_j the jth component of x and, for any subset $J \subseteq \{1, \ldots, k\}$, we denote by x_J the vector with components x_j, $j \in J$. For any matrix A, we denote by $\|A\|$ the matrix norm of A induced by the vector Euclidean norm $\|\cdot\|$, i.e., $\|A\| = \max_{\|x\|=1} \|Ax\|$.

2. Technical preliminaries. In this section we will prove a number of interesting facts concerning the solution set $\mathcal{X}^*$ and the level sets of f over certain subsets of $\mathcal{C}$. These facts will be used in the analysis of subsequent sections.

First, by using the strict convexity of g (cf. (1.4)) and the special structure of f (cf. (1.2)), we have the following simple lemma which says that the linear mapping $x \mapsto Ex$ is invariant over the solution set $\mathcal{X}^*$ (also see [LuT92a] and [Tse91]).

LEMMA 2.1. *There exists a* $t^* \in \Re^m$ *such that* $Ex^* = t^*$ *for all* $x^* \in \mathcal{X}^*$.

From (1.2) and the chain rule for differentiation, we have

$$\nabla f(x) = E^T \nabla g(Ex) + q, \quad \forall x. \tag{2.1}$$

Then, (1.3) yields that ∇f is Lipschitz continuous with Lipschitz constant $\rho\|E^T\|\|E\|$, that is,

$$\|\nabla f(x) - \nabla f(y)\| \leq \rho\|E^T\|\|E\|\|x - y\|, \quad \forall x, \quad \forall y, \tag{2.2}$$

and Lemma 2.1 yields that ∇f is invariant over $\mathcal{X}^*$ or, more precisely,

$$\nabla f(x^*) = d^*, \quad \forall x^* \in \mathcal{X}^*, \tag{2.3}$$

where we let $d^* = E^T \nabla g(t^*) + q$.

The optimality conditions for (1.1), together with (2.3), imply that $\mathcal{X}^*$ is equivalently the solution set of the linear program $\min_{x \in \mathcal{X}} \langle d^*, x \rangle$. Then, as we shall see in the next section, the question of finding a local error bound for (1.1) translates into a perturbation analysis on the solution set to this linear program. To perform this analysis, we will need the following result, due originally to Hoffman [Hof52] (see also [Rob73] and [MaS87]), on the Lipschitzian continuity of the solution set to a linear system as a multifunction of the right-hand side. This result will be used in the proofs of Lemma 3.1 and Theorem 3.2 which follow.

LEMMA 2.2. *Let* C *and* D *be any* $r \times k$ *and* $s \times k$ *matrices. Then, there exists a constant* $\theta > 0$ *depending on* C *and* D *only such that, for any* $\bar{x} \in \Re^k$ *and any* $(d, e) \in \Re^r \times \Re^s$ *such that the linear system* $Cy = d$, $Dy \geq e$ *is consistent, there is a point* $\bar{y}$ *satisfying* $C\bar{y} = d$, $D\bar{y} \geq e$ *with*

$$\|\bar{x} - \bar{y}\| \ \leq \ \theta(\|C\bar{x} - d\| + \|D\bar{x} - e\|).$$

For each $v \geq v^*$ and $\delta \geq 0$, define the level set

$$\mathcal{F}_\delta^v = \{x \in \mathcal{C} \mid \|Bx - c\| \leq \delta, \ f(x) \leq v\}.$$

(Note that $\mathcal{F}_0^{v^*} = \mathcal{X}^*$ and $\mathcal{F}_{\delta'}^{v'} \subseteq \mathcal{F}_\delta^v$ whenever $v' \leq v, \delta' \leq \delta$.) By using the polyhedral structure of $\mathcal{X}$ (cf. (1.6)) together with the strict convexity of g (cf. (1.4)), we can show the following boundedness property of $E\mathcal{F}_\delta^v$. This property will be used in the proofs of Lemma 3.1 and Theorem 5.3. Its proof is patterned after that of Fact 4.1 in [Tse91] and is based on the observation that a strictly convex function has bounded level sets whenever its infimum is attained at some point.

LEMMA 2.3. *For any* $v \geq v^*$ *and any* $\delta \geq 0$, *the set* $E\mathcal{F}_\delta^v$ *is nonempty and bounded.*

Proof. Fix any $v \geq v^*$ and any $\delta \geq 0$. The set $E\mathcal{F}_\delta^v$ is clearly nonempty since $\mathcal{F}_\delta^v$ is nonempty. If $E\mathcal{F}_\delta^v$ were not bounded, then the closed convex set

$$\mathcal{L} = \{(t, x, \zeta) \in \Re^{m+n+1} \mid t = Ex, \ x \in \mathcal{C}, \ \|Bx - c\| \leq \delta, \ f(x) \leq \zeta\}$$

would have a direction of recession $(v, u, 0)$ with $v \neq 0$ (see [Roc70]). Let x^* be any element of $\mathcal{X}^*$. Then, by Lemma 2.1, (t^*, x^*, v^*) is a point in $\mathcal{L}$, so $(t^*, x^*, v^*) + \theta(v, u, 0)$ is also in $\mathcal{L}$ for all $\theta \geq 0$. This implies $x^* + \theta u \in \mathcal{C}$ and $f(x^* + \theta u) \leq v^*$ for all $\theta \geq 0$. Moreover, we see from the structure of $\mathcal{L}$ that $Bu = 0$ and $Eu = v$. The former implies $B(x^* + \theta u) = Bx^* = c$ for all $\theta \geq 0$, so $x^* + \theta u \in \mathcal{X}^*$ for all $\theta \geq 0$. On the other hand, the latter, together with $v \neq 0$, implies that $E(x^* + \theta u)$ is not constant for $\theta \geq 0$, a contradiction of Lemma 2.1. □

3. A new local error bound. In this section we show that the distance from a point x in $\mathcal{C}$ to $\mathcal{X}^*$ can be bounded above by the quantity (1.7) when the latter quantity is small and $f(x)$ is bounded. The proof of this is analogous to an argument used in [LuT92b] and is based on a certain property of (1.7) for identifying (locally) those constraints which are "active" at some optimal solution. By treating these active constraints as equalities, we then apply Hoffman's result (Lemma 2.2), together with the Lipschitz continuity and strong monotonicity properties of ∇g (cf. (1.3) and (1.4)), to establish the desired bound.

First, since $\mathcal{C}$ is a polyhedral set, we can express it as

$$\mathcal{C} = \{x \in \Re^n \mid Ax \geq b\}, \tag{3.1}$$

for some $k \times n$ matrix A and some $b \in \Re^k$. For convenience, we denote by A_i the ith row of A and, for any subset $I \subseteq \{1, \ldots, k\}$, by A_I the submatrix of A obtained by removing all rows i of A with $i \notin I$. Then, for any $(x, p) \in \mathcal{C} \times \Re^l$, the vector $z = [x - \nabla f(x) + B^T p]^+_{\mathcal{C}}$ satisfies, together with some multiplier vector $\lambda \in \Re^k$, the following Kuhn–Tucker conditions:

$$x - z + B^T p + A^T \lambda = \nabla f(x), \quad \lambda_i = 0, \quad \forall i \notin I, \quad A_i z = b_i, \quad \forall i \in I, \tag{3.2}$$

$$Az \geq b, \qquad \lambda \geq 0, \tag{3.3}$$

where I is some (possibly empty) subset of $\{1, \ldots, k\}$. We say that an $I \subseteq \{1, \ldots, k\}$ is *identifiably basic* at a vector $(x, p) \in \mathcal{C} \times \Re^l$ if (x, p), together with $z = [x - \nabla f(x) + B^T p]^+_{\mathcal{C}}$ and some $\lambda \in [0, \infty)^k$, satisfies (3.2).

By using Lemmas 2.1, 2.2, and 2.3, we show the following lemma which roughly says that if $x \in \mathcal{C}$ is sufficiently close to $\mathcal{X}^*$, then those indices which are identifiably basic at (x, p) for some p are also identifiably basic at some element of $\mathcal{X}^* \times \Re^l$.

LEMMA 3.1. *Fix any $\upsilon \geq \upsilon^*$. There exists an $\epsilon > 0$ such that, for any $(x, p) \in \mathcal{F}^{\upsilon}_{\epsilon} \times \Re^l$ with $\|x - [x - \nabla f(x) + B^T p]^+_{\mathcal{C}}\| \leq \epsilon$ and any $I \subseteq \{1, \ldots, k\}$ that is identifiably basic at (x, p), there is some $(x^*, p^*) \in \mathcal{X}^* \times \Re^l$ at which I is identifiably basic.*

Proof. We argue by contradiction. If the claim does not hold, then there would exist an $I \subseteq \{1, \ldots, k\}$ and a sequence of vectors $\{(x^r, p^r)\}_{r=1,2,\ldots}$ in $\mathcal{F}^{\upsilon}_1 \times \Re^l$ with I identifiably basic at (x^r, p^r) for all r and

$$x^r - z^r \to 0, \qquad Bx^r \to c, \tag{3.4}$$

where we let

$$z^r = [x^r - \nabla f(x^r) + B^T p^r]^+_{\mathcal{C}}, \quad \forall r, \tag{3.5}$$

and yet there is no $(x^*, p^*) \in \mathcal{X}^* \times \Re^l$ at which I is identifiably basic.

Since $x^r \in \mathcal{F}^{\upsilon}_1$ for all r, it follows from Lemma 2.3 that $\{Ex^r\}$ is bounded. Let t^∞ be any cluster point of $\{Ex^r\}$ and let R be a subsequence of $\{1, 2, \ldots\}$ such that

$$\{Ex^r\}_R \to t^\infty. \tag{3.6}$$

We show below that t^∞ is equal to t^*.

Since ∇g is continuous everywhere, then we obtain from (3.6) (and using the fact $\nabla f(x^r) = E^T \nabla g(Ex^r) + q$ for all r) that

$$\{\nabla f(x^r)\}_R \to E^T \nabla g(t^\infty) + q. \tag{3.7}$$

For each $r \in R$, consider the following linear system in x, p, z, and λ:

$$B^T p + A^T \lambda = \nabla f(x^r) + z^r - x^r, \quad Az \geq b, \quad \lambda \geq 0,$$
$$\lambda_i = 0, \quad \forall i \notin I, \quad A_i z = b_i, \quad \forall i \in I,$$
$$Ex = Ex^r, \quad z - x = z^r - x^r, \quad Bx = Bx^r.$$

The above system is consistent since, by I being identifiably basic at (x^r, p^r) and by (3.2)–(3.3), (x^r, p^r, z^r), together with some $\lambda^r \in \Re^k$, is a solution of it. Then, by Lemma 2.2, it has a solution $(\hat{x}^r, \hat{p}^r, \hat{z}^r, \hat{\lambda}^r)$ whose size is bounded by some constant (depending on A, B, and E only) times the size of the right-hand side. Since the right-hand side of the above system is clearly bounded as $r \to \infty$, $r \in R$ (cf. (3.4), (3.6), and (3.7)), we have that $\{(\hat{x}^r, \hat{p}^r, \hat{z}^r, \hat{\lambda}^r)\}_R$ is bounded. Moreover, every one of its cluster points, say $(x^\infty, p^\infty, z^\infty, \lambda^\infty)$, satisfies (cf. (3.4), (3.6), and (3.7))

$$B^T p^\infty + A^T \lambda^\infty = E^T \nabla g(t^\infty) + q, \quad Az^\infty \geq b, \quad \lambda^\infty \geq 0,$$
$$\lambda_i^\infty = 0, \quad \forall i \notin I, \quad A_i z^\infty = b_i, \quad \forall i \in I,$$
$$Ex^\infty = t^\infty, \quad z^\infty - x^\infty = 0, \quad Bx^\infty = c.$$

Upon using (cf. (2.1)) $E^T \nabla g(Ex^\infty) + q = \nabla f(x^\infty)$, we can simplify the above relations to

$$B^T p^\infty + A^T \lambda^\infty = \nabla f(x^\infty), \quad Ax^\infty \geq b, \quad \lambda^\infty \geq 0,$$
$$\lambda_i^\infty = 0, \quad \forall i \notin I, \quad A_i x^\infty = b_i, \quad \forall i \in I, \quad Bx^\infty = c.$$

This shows that $x^\infty \in \mathcal{X}$ and that $\langle \nabla f(x^\infty), x - x^\infty \rangle \geq 0$ for all $x \in \mathcal{X}$ (cf. (1.6) and (3.1)). Thus $x^\infty \in \mathcal{X}^*$ and, by Lemma 2.1, $t^\infty = t^*$. Moreover, I is identifiably basic at (x^∞, p^∞) (cf. (3.2)), so a contradiction is established. □

Lemmas 2.1, 2.2, and 3.1 together yield the main result of this section.

THEOREM 3.2 (local error bound). *Fix any $\upsilon \geq \upsilon^*$. There exist scalars $\epsilon > 0$ and $\kappa > 0$ (depending on υ and the problem data only) such that*

$$\phi(x) \leq \kappa(\|x - [x - \nabla f(x) + B^T p]_{\mathcal{C}}^+\| + \|Bx - c\|)$$

for any $(x, p) \in \mathcal{F}_\epsilon^\upsilon \times \Re^l$ with $\|x - [x - \nabla f(x) + B^T p]_{\mathcal{C}}^+\| \leq \epsilon$.

Proof. Let ϵ be the scalar in Lemma 3.1 corresponding to υ. Consider any $(x, p) \in \mathcal{F}_\epsilon^\upsilon \times \Re^l$ satisfying the hypothesis of the theorem and let I be any subset of $\{1, \ldots, k\}$ that is identifiably basic at (x, p) and let $z = [x - \nabla f(x) + B^T p]_{\mathcal{C}}^+$. By (3.2) and (3.3), there exists some $\lambda \in \Re^k$ satisfying, together with x, p, and z,

$$B^T p + A^T \lambda = z - x + \nabla f(x), \quad Ax \geq b + A(x - z), \quad \lambda \geq 0,$$
$$\lambda_i = 0, \quad \forall i \notin I, \quad A_i x = b_i + A_i(x - z), \quad \forall i \in I.$$

By Lemma 3.1, there exists an $(x^*, p^*) \in \mathcal{X}^* \times \Re^l$ such that I is identifiably basic at (x^*, p^*), so the following linear system in x^*, p^*, and λ^*:

$$B^T p^* + A^T \lambda^* = d^*, \quad Ax^* \geq b, \quad \lambda^* \geq 0,$$
$$\lambda_i^* = 0, \quad \forall i \notin I, \quad A_i x^* = b_i, \quad \forall i \in I, \quad Ex^* = t^*, \quad Bx^* = c$$

is consistent (cf. (2.3), (3.2)–(3.3), and Lemma 2.1). Conversely, it can be seen that every solution (x^*, p^*, λ^*) to this linear system satisfies $x^* \in \mathcal{X}^*$. Upon comparing

the above two systems, we see that, by Lemma 2.2, there exists a solution (x^*, p^*, λ^*) to the second system such that

$$\|(x,p,\lambda)-(x^*,p^*,\lambda^*)\| \le \theta(\|z-x+\nabla f(x)-d^*\|+\|A(x-z)\|+\|Ex-t^*\|+\|Bx-c\|),$$

where θ is some scalar constant depending on A, B, and E only. By (2.1), the definition of d^*, and the Lipschitz condition (1.3), we also have $\|\nabla f(x) - d^*\| = \|E^T\nabla g(Ex) - E^T\nabla g(t^*)\| \le \rho\|E^T\|\|Ex - t^*\|$, so the above relation yields

$$\|(x,p,\lambda) - (x^*,p^*,\lambda^*)\| \le \theta\big((\|A\|+1)\|x-z\| + (\rho\|E^T\|+1)\|Ex-t^*\| + \|Bx-c\|\big).$$

Upon rewriting some of the above relations and by using the fact $d^* = \nabla f(x^*)$ (cf. (2.3)), we have

$$x - z + B^T p + A_I^T\lambda_I = \nabla f(x), \quad B^T p^* + A_I^T\lambda_I^* = \nabla f(x^*), \tag{3.8}$$

$$A_I z = b_I, \quad A_I x^* = b_I, \quad Bx^* = c, \tag{3.9}$$

and

$$\|(x,p,\lambda) - (x^*,p^*,\lambda^*)\| \le O(\|Ex - t^*\| + \gamma), \tag{3.10}$$

where we let $\gamma = \|x - z\| + \|Bx - c\|$ and, for convenience, use the notation $\alpha \le O(\beta)$ to indicate that $\alpha \le \omega\beta$ for some scalar $\omega > 0$ depending on υ and the problem data only. In addition, I is identifiably basic at (x^*, p^*) and (cf. (1.4))

$$\sigma\|Ex - t^*\|^2 \le \langle Ex - t^*, \nabla g(Ex) - \nabla g(t^*)\rangle. \tag{3.11}$$

We will use (3.8)–(3.11) to show that $\|x-x^*\| \le O(\gamma)$, which would then complete the proof. Since $Ex^* = t^*$ (cf. Lemma 2.1) and $\nabla f(x) - \nabla f(x^*) = E^T\nabla g(Ex) - E^T\nabla g(Ex^*)$ (cf. (2.1)), then (3.11), together with (3.8)–(3.9), yields

$$\begin{aligned}
\sigma\|Ex - t^*\|^2 &\le \langle Ex - Ex^*, \nabla g(Ex) - \nabla g(Ex^*)\rangle \\
&= \langle x - x^*, \nabla f(x) - \nabla f(x^*)\rangle \\
&= \langle x - x^*, B^T p + A_I^T\lambda_I + x - z - B^T p^* - A_I^T\lambda_I^*\rangle \\
&= \langle B(x - x^*), p - p^*\rangle + \langle A_I(x - x^*), \lambda_I - \lambda_I^*\rangle + \langle x - x^*, x - z\rangle \\
&= \langle Bx - c, p - p^*\rangle + \langle A_I(x - z), \lambda_I - \lambda_I^*\rangle + \langle x - x^*, x - z\rangle \\
&\le \|Bx - c\|\|p - p^*\| + \|x - z\|(\|A\|\|\lambda - \lambda^*\| + \|x - x^*\|) \\
&\le \|A\|(\|p - p^*\| + \|\lambda - \lambda^*\| + \|x - x^*\|)\gamma,
\end{aligned}$$

where the last inequality follows from the definition of γ. Applying the above relation once and (3.10) twice then gives

$$\begin{aligned}
\|x - x^*\|^2 &\le O((\|Ex - t^*\| + \gamma)^2) \\
&\le O(\|Ex - t^*\|^2 + \gamma^2) \\
&\le O((\|p - p^*\| + \|\lambda - \lambda^*\| + \|x - x^*\|)\gamma + \gamma^2) \\
&\le O((\|Ex - t^*\| + \gamma)\gamma + \gamma^2).
\end{aligned}$$

Since $\|Ex - t^*\| \le \|E\|\|x - x^*\|$, the above relation implies that there exists a scalar constant $\omega > 0$ (depending on υ and the problem data only) such that

$$\|Ex - t^*\|^2 \le \omega(\|Ex - t^*\|\gamma + \gamma^2).$$

This is a quadratic inequality of the form $a^2 \leq \omega(a\gamma + \gamma^2)$, which implies $a \leq \frac{1}{2}(\omega + \sqrt{\omega^2 + 4\omega})\gamma$ and therefore

$$\|Ex - t^*\| \leq \tfrac{1}{2}(\omega + \sqrt{\omega^2 + 4\omega})\gamma.$$

Combine this bound with (3.10) and we obtain $\|(x,p,\lambda) - (x^*,p^*,\lambda^*)\| \leq O(\gamma)$. □

We note that the proof of Theorem 3.2 in fact yields the stronger result that, for any $(x,p) \in \mathcal{F}_\epsilon^\upsilon \times \Re^l$ satisfying $\|x - [x - \nabla f(x) + B^T p]_{\mathcal{C}}^+\| \leq \epsilon$ and any $I \subseteq \{1,\ldots,k\}$ that is identifiably basic at (x,p), there exists an $(x^*,p^*) \in \mathcal{X}^* \times \Re^l$ such that I is identifiably basic at (x^*,p^*) and $\|x - x^*\| \leq \kappa(\|x - [x - \nabla f(x) + B^T p]_{\mathcal{C}}^+\| + \|Bx - c\|)$, for some scalar κ depending on υ and the problem data only. Roughly speaking, we can bound ϕ and identify the active constraints at the same time. Finally, we remark that, at the price of forgoing this stronger result, the proof of Theorem 3.2 can be simplified further by appealing to a result of Robinson [Rob81] on the local upper Lipschitzian nature of polyhedral multifunctions.

4. RGP algorithms. In this section, we introduce a general class of feasible descent algorithms for solving the special case of (1.1) where $\mathcal{C}$ is the nonnegative orthant in $\Re^n$, i.e.,

$$\mathcal{C} = [0,\infty)^n. \tag{4.1}$$

An algorithm in this class updates an iterate by first moving it opposite a certain reduced-gradient direction, then projecting it onto $\mathcal{C}$, and finally adjusting a subset of the coordinates with zero reduced gradient, so that the new iterate remains in $\mathcal{X}$. We will show that both the gradient projection algorithm and the algorithm of Bertsekas mentioned in §1 belong to this class. We also propose a new algorithm in this class reminiscent of active set algorithms and, in particular, of a projected Newton method of Bertsekas [Ber82]. Unlike most active set algorithms, this algorithm can add/drop many constraints from its active set at each iteration. We remark that the above class of algorithms readily extends to the case where $\mathcal{C}$ is a *box* in $\Re^n$, i.e., the Cartesian product of closed intervals, but, for simplicity, we will not consider this more general case here.

In what follows, we denote by B_j the jth column of B and, for each $J \subseteq \{1,\ldots,n\}$, by B_J the matrix obtained by removing all columns B_j, $j \notin J$, from B. We define $\nabla_j f$ and $\nabla_J f$ analogously. We also denote by $\bar{J}$ the complement of J with respect to $\{1,\ldots,n\}$.

To motivate our algorithms, consider an iteration of the gradient projection algorithm: $x' = [x - \alpha\nabla f(x)]_{\mathcal{X}}^+$, where x is the current iterate, α is the stepsize, and x' is the new iterate. Let $[\cdot]_+$ denote the orthogonal projection onto $[0,\infty)^n$. By using the structure of $\mathcal{X}$ given by (1.6) and (4.1), we can rewrite this iteration as $x' \in \mathcal{X}$ and, for some $p \in \Re^l$,

$$x' = [x - \alpha(\nabla f(x) - B^T p)]_+. \tag{4.2}$$

(It can be seen that p is in fact an optimal Lagrange multiplier vector associated with the constraints $Bx = c$ in the problem of projecting $x/\alpha - \nabla f(x)$ onto $\mathcal{X}$.) Thus, the above iteration is equivalent to the problem of finding a $p \in \Re^l$ so that x' given by (4.2) is in $\mathcal{X}$. Can the restriction (4.2) be relaxed so it would be relatively easy to find such a p?

To answer this question, suppose that, in addition to (4.1), we have $B = [1\ 1\ \cdots\ 1]$ and $c = 1$ (so $\mathcal{X}$ is the unit simplex). Consider the algorithm of Bertsekas mentioned

in §1 for solving this special case of (1.1), which operates as follows: Given an iterate $x \in \mathcal{X}$, it chooses an index $j \in \{1, \ldots, n\}$ for which

$$\nabla_j f(x) = \min_k \nabla_k f(x), \tag{4.3}$$

and computes a new iterate $x' \in \mathcal{X}$ according to

$$x'_k = [x_k - \alpha\,(\nabla_k f(x) - \nabla_j f(x))]_+ \quad \forall k \neq j, \tag{4.4}$$

$$x'_j = 1 - \sum_{k \neq j} x'_k, \tag{4.5}$$

where α is some positive stepsize. (The fact that $x'_j \geq 0$ follows from the observation that $x'_k \leq x_k$ for all $k \neq j$, so the fact $\sum_k x_k = 1 = \sum_k x'_k$ yields $x'_j \geq x_j$.) A moment of reflection shows that the iteration (4.4) is simply the following relaxed version of (4.2):

$$x'_k = [x_k - \alpha(\nabla_k f(x) - B_k^T p)]_+, \quad \forall k \neq j, \tag{4.6}$$

with $p = \nabla_j f(x)$. Moreover, by combining (4.4) with (4.5), we see that

$$\|x' - x\| \leq \sqrt{n}\|x - [x - \alpha(\nabla f(x) - B^T p)]_+\|. \tag{4.7}$$

We remark that, for simplicity, we considered only the unscaled version of the Bertsekas algorithm. See [BeG87, §5.7] for a description of the full algorithm; see [Ber82, §3] and [BeG83] for a related algorithm in which j is chosen by the maximum component rule: $j = \arg\max_k x_k$. This latter algorithm is closely linked to the active-set-type algorithm to be described below.

The formulas (4.6) and (4.7) suggest the following generalization of the gradient projection algorithm and the Bertsekas algorithm for solving (1.1) (under the condition (4.1)) whereby, given an iterate $x \in \mathcal{X}$, we choose a positive stepsize α and we compute a new iterate x' which, together with some $p \in \Re^l$, satisfies

$$x'_k = [x_k - \alpha(\nabla_k f(x) - B_k^T p)]_+, \quad \forall k \quad \text{with } \nabla_k f(x) \neq B_k^T p, \tag{4.8}$$

and

$$\|x' - x\| \leq \tau_1 \|x - [x - \alpha(\nabla f(x) - B^T p)]_+\|, \tag{4.9}$$

with τ_1 some scalar constant. In order to maintain feasibility, we assume that the new iterate x' has the property that

$$x' \in \mathcal{X} \quad \text{whenever } \alpha < \frac{\tau_2}{\|\nabla f(x)\|}, \tag{4.10}$$

with τ_2 some scalar constant (possibly $\tau_2 = \infty$). Thus x' is feasible whenever α is chosen to be sufficiently small.

We will call any iteration a *reduced-gradient* projection (RGP) iteration if it generates, for a given iterate $x \in \mathcal{X}$ and a stepsize $\alpha > 0$, a new iterate x' satisfying (together with some $p \in \Re^l$) the relations (4.8)–(4.10). Roughly speaking, at each RGP iteration we take a step opposite the *reduced-gradient* direction $\nabla f(x) - B^T p$, project onto $[0, \infty)^n$, and then adjust those coordinates with zero reduced gradient so as to remain in $\mathcal{X}$. Any algorithm that generates iterates in $\mathcal{X}$ by successive applications of RGP iterations will be called an RGP algorithm. We now describe three

example RGP algorithms, the first two of which we have encountered earlier. The issue of stepsize rules will be addressed in the next section.

Example 4.1. Gradient projection algorithm. By (4.2), the gradient projection algorithm is an RGP algorithm with $\tau_1 = 1$, $\tau_2 = \infty$, and p an optimal multiplier vector associated with $Bx = c$ in the problem of projecting $x/\alpha - \nabla f(x)$ onto $\mathcal{X}$.

Example 4.2. Bertsekas algorithm. By (4.6) and (4.7), the Bertsekas algorithm (4.3)–(4.5) is an RGP algorithm with $\tau_1 = \sqrt{n}$, $\tau_2 = \infty$, and $p = \min_k \nabla_k f(x)$.

Example 4.3. An active-set-type algorithm. Consider the following algorithm for solving (1.1), under the condition (4.1): Fix any $\gamma > 0$. Given an iterate $x \in \mathcal{X}$, we choose a positive stepsize α and a (possibly empty) subset $J \subseteq \{\, j \in \{1, \ldots, n\} \mid x_j \geq \gamma \,\}$ with B_J having full column rank, and we compute a new iterate x' as the (unique) solution of a convex quadratic program, given by

$$x' = \arg \min_{\substack{\xi \text{ with } B\xi = c \\ \xi_k \geq 0 \ \forall k \notin J}} \sum_{k \in J} \nabla_k f(x)(\xi_k - x_k) + \frac{1}{2\alpha} \sum_{k \notin J} |\xi_k - (x_k - \alpha \nabla_k f(x))|^2. \tag{4.11}$$

We will show that the iteration (4.11) is well defined and the x' thus generated, together with some p, satifies (4.8)–(4.10) for some scalar constants τ_1 and τ_2.

The above algorithm may be viewed as a generalization of the gradient projection algorithm in which projection is omitted for coordinates that are far from the boundary. In particular, if we take J to be the empty set, then we recover the gradient projection algorithm (see Example 4.1). A key advantage of the algorithm is its flexibility. For example, we can choose the set J so that the work in solving (4.11) is less than that for performing the full projection (see discussions to follow). The parameter γ, however, needs to be chosen with care. If γ is too large, the choices for J would be restricted; if γ is small, then, as we shall see, α may need to be small (cf. (4.14)), in which case the algorithm would take small steps. Finally, we note that γ need not be fixed but can be adjusted dynamically, provided that it remains bounded away from zero.

We now show that the iteration (4.11) is a well-defined RGP iteration. If J is the empty set, then (4.11) reduces to a gradient projection iteration, so it is well defined and the x' generated by it, together with some p, satisfies (4.8)–(4.10) with $\tau_1 = 1$ and $\tau_2 = \infty$ (cf. Example 4.1). Thus, it remains to prove the above assertion for the case where J is nonempty. First, notice that the feasible set for the minimization in (4.11) is nonempty (since it contains $\mathcal{X}$) and bounded (since the objective function is strongly convex in $\xi_{\bar{J}}$ and, by virtue of B_J having full column rank, ξ_J is determined uniquely by $\xi_{\bar{J}}$ on the feasible set). Thus, the minimization in (4.11) has an optimal solution. It is easily seen that this optimal solution is unique, so (4.11) is well defined. From the optimality conditions for the minimization in (4.11) we have that $Bx' = c$ and

$$x'_{\bar{J}} = [x_{\bar{J}} - \alpha(\nabla_{\bar{J}} f(x) - B_{\bar{J}}^T p)]_+, \qquad \nabla_J f(x) = B_J^T p, \tag{4.12}$$

where p is any optimal Lagrange multiplier vector associated with the constraints $B\xi = c$ in (4.11). The former, together with the fact $Bx = c$, implies $0 = B(x' - x) = B_J(x'_J - x_J) + B_{\bar{J}}(x'_{\bar{J}} - x_{\bar{J}})$ so, multiplying both sides by B_J^T and using the fact that B_J has full column rank, we can solve for $x'_J - x_J$ to obtain

$$x'_J - x_J = -(B_J^T B_J)^{-1} B_J^T B_{\bar{J}}(x'_{\bar{J}} - x_{\bar{J}}),$$

implying

$$\|x'_J - x_J\| \leq \|(B_J^T B_J)^{-1} B_J^T B_{\bar{J}}\| \|x'_{\bar{J}} - x_{\bar{J}}\|. \tag{4.13}$$

Relations (4.12) and (4.13) show that x', together with p, satisfies (4.8) and (4.9) with $\tau_1 = 1 + \|(B_J^T B_J)^{-1} B_J^T B_{\bar{J}}\|$.

It only remains to show that x' satisfies (4.10) for some scalar constant τ_2. For any subset I of $\{1, \ldots, m\}$ and any subset J of $\{1, \ldots, n\}$, let B_{IJ} denote the matrix obtained by removing from B_J all rows i with $i \notin I$. We show below that $x' \in \mathcal{X}$ whenever

$$\alpha \leq \frac{\min_{k \in J}\{x_k\}}{\|(B_J^T B_J)^{-1} B_J^T B_{\bar{J}}\| \|\nabla_{\bar{J}} f(x) - B_{I\bar{J}}^T (B_{IJ}^T)^{-1} \nabla_J f(x)\|}, \tag{4.14}$$

where I is any subset of $\{1, \ldots, m\}$ such that B_{IJ} is invertible. This, together with the fact $x_j \geq \gamma$ for all $j \in J$, would then complete the proof. First, we observe that the constraints $B\xi = c$ can be rewritten as $B_{IJ}\xi_J + B_{I\bar{J}}\xi_{\bar{J}} = c_I$ and $B_{\tilde{I}J}\xi_J + B_{\tilde{I}\bar{J}}\xi_{\bar{J}} = c_{\tilde{I}}$, where $\tilde{I}$ is the complement of I relative to $\{1, \ldots, m\}$. Using the first set of constraints to eliminate ξ_J from the second set and from the objective function in (4.11), we reduce the minimization in (4.11) to the following problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2\alpha} \sum_{k \notin J} |\xi_k - (x_k - \alpha \nabla_k f(x))|^2 - \langle \nabla_J f(x), (B_{IJ})^{-1} B_{I\bar{J}} \xi_{\bar{J}} \rangle \\ \text{subject to} \quad & \left(B_{\tilde{I}\bar{J}} - B_{\tilde{I}J}(B_{IJ})^{-1} B_{I\bar{J}}\right) \xi_{\bar{J}} = c_{\tilde{I}} - B_{\tilde{I}J}(B_{IJ})^{-1} c_I, \qquad \xi_{\bar{J}} \geq 0, \end{aligned}$$

to which $x'_{\bar{J}}$ is an optimal solution. Then, $x'_{\bar{J}}$ satisfies the optimality conditions:

$$x'_{\bar{J}} = [x_{\bar{J}} - \alpha(\nabla_{\bar{J}} f(x) - B_{I\bar{J}}^T (B_{IJ}^T)^{-1} \nabla_J f(x))]^+_{\mathcal{D}},$$

where $\mathcal{D}$ denotes the feasible set for the reduced problem. This, combined with the observation that $x_{\bar{J}} \in \mathcal{D}$ (cf. $x \in \mathcal{X}$), implies

$$\begin{aligned} \|x'_{\bar{J}} - x_{\bar{J}}\| &= \|[x_{\bar{J}} - \alpha(\nabla_{\bar{J}} f(x) - B_{I\bar{J}}^T (B_{IJ}^T)^{-1} \nabla_J f(x))]^+_{\mathcal{D}} - [x_{\bar{J}}]^+_{\mathcal{D}}\| \\ &\leq \alpha \|\nabla_{\bar{J}} f(x) - B_{I\bar{J}}^T (B_{IJ}^T)^{-1} \nabla_J f(x)\|, \end{aligned}$$

where the last inequality follows from the nonexpansive property of the projection mapping $[\cdot]^+_{\mathcal{D}}$. Combining this with (4.13) gives

$$\|x'_J - x_J\| \leq \alpha \|(B_J^T B_J)^{-1} B_J^T B_{\bar{J}}\| \|\nabla_{\bar{J}} f(x) - B_{I\bar{J}}^T (B_{IJ}^T)^{-1} \nabla_J f(x)\|,$$

and it follows that $x'_J \geq 0$ whenever α satisfies (4.14). Since $Bx' = c$ and (cf. (4.12)) $x'_{\bar{J}} \geq 0$, this shows that $x' \in \mathcal{X}$ (cf. (1.6) and (4.1)) whenever α satisfies (4.14).

The iteration (4.11) admits an interesting interpretation as an active-set-type iteration. To see this, let us assume for simplicity that the matrix B_J therein is invertible. Then, since $\nabla_J f(x) = B_J^T p$ (cf. (4.12)), we can eliminate p from the first expression in (4.12) to obtain

$$x'_{\bar{J}} = [x_{\bar{J}} - \alpha(\nabla_{\bar{J}} f(x) - B_{\bar{J}}^T (B_J^T)^{-1} \nabla_J f(x))]_+.$$

Also, since $Bx' = c$, we can solve for x'_J to obtain

$$x'_J = (B_J)^{-1}(c - B_{\bar{J}} x'_{\bar{J}}).$$

Thus we may interpret (4.11) as an iteration in which we first take a reduced-gradient projection step, and then we adjust those coordinates for which the reduced gradient is zero so that the new iterate x' satisfies $Bx' = c$. This philosophy of taking a descent

step with respect to those coordinates "active" at their respective bounds (i.e., $x_{\bar{J}}$) is reminiscent of active set schemes for solving problems with simple bounds. In fact, it can be seen that the above iteration is very similar to an unscaled version of a projected Newton method studied by Bertsekas [Ber82, §3] and Bertsekas and Gafni [BeG83]. In contrast to conventional active set schemes, the above scheme has the advantage that it can add and drop many elements from its currently active set $\bar{J}$ at each iteration.

5. Convergence of RGP algorithms. In this section we show, by using the local error bound of §3, that every RGP algorithm with the stepsizes chosen according to an Armijo-like rule is linearly convergent. The proof of this is analogous to a proof given in [LuT92b].

First, we describe the rule for choosing the stepsizes α. This rule is based on the efficient Armijo-like rule proposed by Bertsekas for the gradient projection algorithm [Ber76]. Let τ_1 and τ_2 be the parameters of a given RGP iteration (cf. (4.9) and (4.10)). We fix two parameters $\beta \in (0,1)$ and $\tau_3 > 0$ and we let

$$\tau_4 = \tfrac{1}{2}\|E^T\|\|E\|\rho(\tau_1)^2 + \tau_3.$$

Given an iterate $x \in \mathcal{X}$, we choose a number α_0 with $\alpha_0 \geq \min\{1/\tau_4, \tau_2/\|\nabla f(x)\|\}$ and we set

$$\alpha = \alpha_0 \beta^k, \tag{5.1}$$

where k is the first nonnegative integer for which an x' and a p generated by the RGP iteration with α given as above (i.e., x' and p together satisfy (4.8)–(4.10)) satisfies $x' \in \mathcal{X}$ and the sufficient descent condition

$$f(x) - f(x') \geq \tau_3 \alpha \|x - [x - \nabla f(x) + B^T p]_+\|^2. \tag{5.2}$$

We remark that, instead of the Armijo-like rule given above, we can also use a stepsize rule analogous to one proposed by Goldstein [Gol74] and the analysis can be adapted accordingly.

We next show that the stepsize rule (5.1)–(5.2) is well defined and that the stepsize generated is sufficiently large.

LEMMA 5.1. *The stepsize rule* (5.1)–(5.2) *is well defined. Moreover, the stepsize* α *generated by this rule is bounded below by* $\beta \min\{1/\tau_4, \tau_2/\|\nabla f(x)\|\}$.

Proof. First, we show that, for a given $x \in \mathcal{X}$ and a positive number α strictly less than $\min\{1/\tau_4, \tau_2/\|\nabla f(x)\|\}$, any x' and any $p \in \Re^l$ that together satisfy (4.8)–(4.10) also satisfy $x' \in \mathcal{X}$ and (5.2). Since ∇f is Lipschitz continuous with Lipschitz constant $\|E^T\|\|E\|\rho$ (cf. (2.2)), we have

$$f(x) - f(x') \geq \langle \nabla f(x), x - x' \rangle - \frac{\|E^T\|\|E\|\rho}{2}\|x' - x\|^2. \tag{5.3}$$

Let $J = \{j \in \{1, \ldots, n\} \mid B_j^T p = \nabla_j f(x)\}$. Then, by (4.8), $x'_{\bar{J}}$ is the orthogonal projection of $x_{\bar{J}} - \alpha(\nabla_{\bar{J}} f(x) - B_{\bar{J}}^T p)$ onto the nonnegative orthant. Since $x \geq 0$, this implies

$$\langle x'_{\bar{J}} - x_{\bar{J}} + \alpha(\nabla_{\bar{J}} f(x) - B_{\bar{J}}^T p), x_{\bar{J}} - x'_{\bar{J}} \rangle \geq 0.$$

Since $Bx = Bx'$ (cf. $x \in \mathcal{X}$ and $x' \in \mathcal{X}$), we have from the definition of J and the above relation that

$$\begin{aligned} \langle \nabla f(x), x - x' \rangle &= \langle \nabla f(x) - B^T p, x - x' \rangle \\ &= \langle \nabla_{\bar{J}} f(x) - B_{\bar{J}}^T p, x_{\bar{J}} - x'_{\bar{J}} \rangle \\ &\geq \frac{1}{\alpha} \|x_{\bar{J}} - x'_{\bar{J}}\|^2. \end{aligned} \tag{5.4}$$

Upon combining (5.3) with (5.4), we obtain

$$f(x) - f(x') \geq \frac{1}{\alpha} \|x_{\bar{J}} - x'_{\bar{J}}\|^2 - \frac{\|E^T\| \|E\| \rho}{2} \|x - x'\|^2,$$

so (4.8), (4.9) together with the definitions of J and τ_4 yield

$$f(x) - f(x') \geq \left(\frac{1}{\alpha} - \tau_4 + \tau_3 \right) \|x - [x - \alpha(\nabla f(x) - B^T p)]_+\|^2.$$

Since $\|x - [x - \alpha d]_+\| \geq \alpha \|x - [x - d]_+\|$ for any $d \in \Re^n$ (see, for example, Lemma 1 in [GaB84]), this shows

$$f(x) - f(x') \geq (1 - \tau_4 \alpha + \tau_3 \alpha) \|x - [x - \nabla f(x) + B^T p]_+\|^2.$$

Thus x' together with p satisfies (5.2) whenever α is less than $1/\tau_4$. Since x' satisfies (4.10), we also have that $x' \in \mathcal{X}$ whenever α is less than $\tau_2 / \|\nabla f(x)\|$.

The above result implies that, for a given $x \in \mathcal{X}$, if the integer k is sufficiently large, then any x' and p satisfying (4.8)–(4.10), with α given by (5.1), also satisfies $x' \in \mathcal{X}$ and (5.2). There must be a first k for which this occurs, so the stepsize rule (5.1)–(5.2) is well defined. Now we prove the second claim. Let $\bar{\alpha}$ be the stepsize given by this rule. Then, either $\bar{\alpha} = \alpha_0$ or $\bar{\alpha} < \alpha_0$. In the former case the second claim holds trivially (by choice of α_0). In the latter case, there must exist some x' and p satisfying (4.8)–(4.10), with α set to $\bar{\alpha}/\beta$, such that either $x' \notin \mathcal{X}$ or (5.2) fails to hold. By the result proven above, this means that $\bar{\alpha}/\beta$ must be greater than or equal to $\min\{1/\tau_4, \tau_2/\|\nabla f(x)\|\}$ or, equivalently, $\bar{\alpha}$ is greater than or equal to β times the latter quantity. The second claim then follows. □

Our final lemma bounds the cost difference $f(x') - v^*$ in terms of the *inexact* residual $x - [x - \nabla f(x) + B^T p]^+$. This bound is analogous to the cost bounds used in the convergence analysis of gradient projection methods (see [Dun87, eq. (23)], [GaD88, Lemmas 2 and 3], and [LuT92b, Thms. 2.1 and 3.1]).

LEMMA 5.2. *Fix any $v \geq v^*$ and let ϵ be the corresponding scalar given in Theorem 3.2. For any $x \in \mathcal{X}$, any $p \in \Re^l$, and any $x' \in \mathcal{X}$ satisfying $f(x) \leq v$, $\|x - [x - \nabla f(x) + B^T p]_+\| \leq \epsilon$ and (4.8)–(4.9), we have*

$$f(x') - v^* \leq \tau_5 \left(1 + \frac{1}{\alpha} \right) \|x - [x - \nabla f(x) + B^T p]_+\|^2,$$

where $\tau_5 > 0$ is some scalar constant depending on v and the problem data only.

Proof. Fix any x, x', and p satisfying the hypothesis of the lemma. Let $z = [x - \nabla f(x) + B^T p]^+$. Then, $(x, p) \in \mathcal{F}_0^v \times \Re^l$ and $\|x - z\| \leq \epsilon$, so (x, p) satisfies the hypothesis of Theorem 3.2. Upon invoking Theorem 3.2, we have that there exists some $x^* \in \mathcal{X}^*$ such that

$$\|x - x^*\| \leq \kappa \|x - z\|, \tag{5.5}$$

where κ is the scalar in Theorem 3.2.

Since $Bx' = Bx^*$, then

$$\begin{aligned}\langle \nabla f(x), x' - x^* \rangle &= \langle \nabla f(x) - B^T p, x' - x^* \rangle \\ &= \langle \nabla_{\bar{J}} f(x) - B_{\bar{J}}^T p, x'_{\bar{J}} - x^*_{\bar{J}} \rangle,\end{aligned}$$

where we let $J = \{j \in \{1, \ldots, n\} \mid B_j^T p = \nabla_j f(x)\}$. Since $x'_{\bar{J}}$ is the orthogonal projection of $x_{\bar{J}} - \alpha(\nabla_{\bar{J}} f(x) - B_{\bar{J}}^T p)$ onto the nonnegative orthant (cf. (4.8)) and $x^*_{\bar{J}} \geq 0$, we also have

$$\langle x'_{\bar{J}} - x_{\bar{J}} + \alpha(\nabla_{\bar{J}} f(x) - B_{\bar{J}}^T p), x'_{\bar{J}} - x^*_{\bar{J}} \rangle \leq 0,$$

which, when combined with the previous relation, yields

$$\langle \nabla f(x), x' - x^* \rangle \leq \frac{1}{\alpha} \langle x_{\bar{J}} - x'_{\bar{J}}, x'_{\bar{J}} - x^*_{\bar{J}} \rangle.$$

Also, by the Mean Value Theorem, there exists some ζ lying on the line segment joining x' with x^* such that

$$f(x') - f(x^*) = \langle \nabla f(\zeta), x' - x^* \rangle.$$

Summing the above two relations and rearranging terms give

$$\begin{aligned}f(x') - f(x^*) &\leq \langle \nabla f(\zeta) - \nabla f(x), x' - x^* \rangle + \frac{1}{\alpha} \langle x_{\bar{J}} - x'_{\bar{J}}, x'_{\bar{J}} - x^*_{\bar{J}} \rangle \\ &\leq \left(\|\nabla f(\zeta) - \nabla f(x)\| + \frac{1}{\alpha} \|x - x'\| \right) \|x' - x^*\| \\ &\leq \left(\rho \|E^T\| \|E\| \|\zeta - x\| + \frac{1}{\alpha} \|x - x'\| \right) \|x' - x^*\| \\ &\leq \left(\rho \|E^T\| \|E\| \|x^* - x\| + \frac{1}{\alpha} \|x - x'\| \right) \|x' - x^*\|,\end{aligned}$$

where the third inequality follows from the Lipschitz continuity property of ∇f (cf. (2.2)). Using (5.5) and the fact $\|x - x'\| \leq \tau_1 \|x - z\|$ (cf. (4.9)) to bound the right-hand side of the above relation completes our proof. □

Upon using Lemmas 5.1 and 5.2, we can now establish the linear rate of convergence for RGP algorithms employing the Armijo-like stepsize rule.

THEOREM 5.3 (linear convergence). *Let $\{x^0, x^1, \ldots\}$ be a sequence in $\mathcal{X}$ generated by a* RGP *algorithm (cf.* (4.8)–(4.10)*) using the Armijo-like stepsize rule (cf.* (5.1)–(5.2)*). Then, $\{x^r\}$ converges at least linearly to an element of $\mathcal{X}^*$ and $\{f(x^r)\}$ converges at least linearly to v^*.*

Proof. For each index $r \geq 0$, let α^r and p^r denote, respectively, the stepsize and the multiplier vector associated with the generation of x^{r+1} by the RGP algorithm using the Armijo-like stepsize rule. In other words, the conditions (4.8)–(4.9) and (5.1)–(5.2), as well as $x' \in \mathcal{X}$, are satisfied by $x = x^r$, $x' = x^{r+1}$, $\alpha = \alpha^r$, and $p = p^r$ for every r. By (5.2), we have

$$f(x^r) - f(x^{r+1}) \geq \tau_3 \alpha^r \|x^r - [x^r - \nabla f(x^r) + B^T p^r]_+\|^2, \quad \forall r, \tag{5.6}$$

and, by Lemma 5.1, we have

$$\alpha^r \geq \beta \min\{1/\tau_4, \tau_2 / \|\nabla f(x^r)\|\}, \quad \forall r. \tag{5.7}$$

Relation (5.6) implies $f(x^r) \leq f(x^0)$ for all r. Since in addition $x^r \in \mathcal{X}$ for all r, we obtain from (1.6) that $x^r \in \mathcal{F}_0^v$ for all r where we let $v = f(x^0)$. Then, Lemma 2.3 implies that the sequence $\{Ex^r\}$ is bounded. Since ∇g is continuous, this in turn implies that $\{\nabla g(Ex^r)\}$ is bounded, so that (cf. (2.1)) $\{\nabla f(x^r)\}$ is bounded. Combining this with (5.7), we see that $\{\alpha^r\}$ is bounded below by some positive scalar constant.

Since $\{\alpha^r\}$ is bounded away from zero and f is bounded below on $\mathcal{X}$, the relation (5.6) implies $x^r - [x^r - \nabla f(x^r) + B^T p^r]_+ \to 0$. Then, by Lemma 5.2, there exist a scalar constant $\tau_5 > 0$ and an index $\bar{r}$ such that

$$\|x^r - [x^r - \nabla f(x^r) + B^T p^r]_+\|^2 \geq \frac{\alpha^r}{\tau_5(1+\alpha^r)}(f(x^{r+1}) - v^*), \quad \forall r \geq \bar{r},$$

which, when combined with (5.6), yields

$$f(x^r) - f(x^{r+1}) \geq \frac{\tau_3(\alpha^r)^2}{\tau_5(1+\alpha^r)}(f(x^{r+1}) - v^*), \quad \forall r \geq \bar{r}.$$

Upon rearranging terms in the above relation, we obtain

$$f(x^{r+1}) - v^* \leq \frac{\tau_5(1+\alpha^r)}{\tau_5(1+\alpha^r) + \tau_3(\alpha^r)^2}\,(f(x^r) - v^*), \quad \forall r \geq \bar{r}.$$

Since $\{\alpha^r\}$ is bounded away from zero, this shows that $f(x^r) \to v^*$ at least linearly, which, together with (5.6), shows that $\|x^r - [x^r - \nabla f(x^r) + B^T p^r]_+\| \to 0$ at least linearly. Since $\|x^{r+1} - x^r\| \leq \tau_1 \|x^r - [x^r - \nabla f(x^r) + B^T p^r]_+\|$ (cf. (4.9)), it follows that $\|x^{r+1} - x^r\| \to 0$ at least linearly, so $\{x^r\}$ converges. Since $f(x^r) \to v^*$, the limit point of $\{x^r\}$ is in $\mathcal{X}^*$. □

We have just shown that any RGP algorithm using the Armijo-like stepsize rule attains a linear rate of convergence. Upon applying Theorem 5.3 to the algorithm of Bertsekas and to the active-set-type algorithm of §4, we immediately obtain the following new convergence results.

COROLLARY 5.4. *Suppose that* $C = [0, \infty)^n$, $B = [1\ 1\ \cdots\ 1]$, *and* $c = 1$. *Then, any sequence of iterates generated by the Bertsekas algorithm* (*cf.* (4.3)–(4.5)), *with stepsizes determined by the Armijo-like rule* (*cf.* (5.1)–(5.2)), *converges at least linearly to an element of* $\mathcal{X}^*$.

COROLLARY 5.5. *Suppose that* $C = [0, \infty)^n$. *Then, any sequence of iterates generated by the active-set-type algorithm* (*cf.* (4.11)), *with stepsizes determined by the Armijo-like rule* (*cf.* (5.1)–(5.2)), *converges at least linearly to an element of* $\mathcal{X}^*$.

6. Concluding remarks. In this paper, we studied a (new) local error bound for certain convex minimization problems over a polyhedral set. We then used this error bound to prove linear convergence for a class of reduced-gradient projection algorithms.

There are several directions in which our results may be generalized. We briefly describe two main ones below.

1. Problems with extended-real-valued cost function. In many situations, g is defined only on some open subset $\mathcal{G}$ of $\Re^m$ and ∇g is Lipschitz continuous and strongly monotone on any compact subset of $\mathcal{G}$. All of our results can be extended to this situation provided that, for some $\bar{v} > v^*$, the level set $\mathcal{F} = \{x \in \mathcal{X} \mid f(x) \leq \bar{v}\}$ satisfies

$$E\mathcal{F} \subseteq \mathcal{G}.$$

(Notice that the above condition holds automatically if dom g is open and g tends to ∞ at the boundary of dom g.) In particular, Theorem 3.2 still holds provided that v therein does not exceed $\bar{v}$. The proof of this is based on an interesting fact that, for $\delta > 0$ sufficiently small, $E\mathcal{F}_\delta^{\bar{v}}$ is a compact subset of $\mathcal{G}$, where $\mathcal{F}_\delta^{\bar{v}}$ is defined as in §2. (The proof of this is similar to that of Lemma 9.1 in [Tse91].) By using this fact in place of Lemma 2.3, we can verify that all the steps in the proof of Theorem 3.2 go through, provided that we take $v \leq \bar{v}$. Linear convergence of the algorithms described in §4 also holds, provided that the stepsize α is taken sufficiently small so as to ensure that each new iterate remains within $\mathcal{F}$. (The proof of the latter uses the boundedness of ∇f on $\mathcal{F}$ and the strict inclusion of $E\mathcal{F}$ by $\mathcal{G}$.)

2. Variational inequality problems. The error bound in §3 readily extends to the following variational inequality problem, first studied by Bertsekas and Gafni [BeG82], of finding an x^* satisfying

$$x^* = [x^* - F(x^*)]_{\mathcal{X}}^+,$$

where $F(x) = E^T G(Ex) + q$ and $G : \Re^m \mapsto \Re^m$ is a Lipschitz continuous strongly monotone function. However, it is unclear whether the bound would help in the development of algorithms for solving such a problem. The error bound also readily extends to *affine* variational inequality problems (where F in the above problem is any affine mapping). This follows from a result of Robinson [Rob81] on certain Lipschitz continuity properties of polyhedral multifunctions.

There remain many open questions which we plan to investigate. Specifically, can the local error bound described in §3 be extended to problems with general convex constraints? Can the linear convergence result of Corollary 5.4 be extended to an asynchronous version of the Bertsekas algorithm proposed by Tsitsiklis and Bertsekas [TsB86]? Some progress along this latter direction has already been made (see [LuT91]). Are there other reduced-gradient projection algorithms, different from those described here, to which our convergence analysis can be fruitfully applied?

It was pointed out to us by one of the referees that, although RGP algorithms typically require less work per iteration than the gradient projection algorithm, their rate of convergence may be slower, thus offsetting any saving in the per iteration workload. In particular, a careful examination of the convergence analysis in §5 shows that, in the worst case, the rate of convergence of an RGP algorithm may depend on n, whereas the gradient projection algorithm does not. Does this dependence exist in practice and, if yes, what are its effects on the performance of an RGP algorithm? This is yet another question that we hope to address in the future.

Acknowledgment. We thank Professor D. P. Bertsekas and an anonymous referee for their many helpful comments, which led to a number of improvements in the presentation.

REFERENCES

[Ber76] D. P. BERTSEKAS, *On the Goldstein–Levitin–Poljak gradient projection method*, IEEE Trans. Automat. Control, AC21 (1976), pp. 174–184.

[Ber80] ———, *A class of routing algorithms for communication networks*, in Proc. Fifth Internat. Conf. Comput. Commun., Atlanta, GA, 1980, pp. 71–76.

[Ber82] ———, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim., 20 (1982), pp. 221–246.

[BeG82] D. P. BERTSEKAS AND E. M. GAFNI, *Projection methods for variational inequalities with application to the traffic assignment problem*, in Math. Programming Stud., D. C. Sorensen and R. J.-B. Wets, eds., 17 (1982), pp. 139–159.

[BeG83] D. P. BERTSEKAS AND E. M. GAFNI, *Projected Newton methods and optimization of multicommodity flows,* IEEE Trans. Automat. Control, AC28 (1983), pp. 1090–1096.

[BeG87] D. P. BERTSEKAS AND R. GALLAGER, *Data Networks*, Prentice–Hall, Englewood Cliffs, NJ, 1987.

[BeT89] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Parallel and Distributed Computation: Numerical Methods*, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[Dun81] J. C. DUNN, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, SIAM J. Control Optim., 19 (1981), pp. 368–400.

[Dun87] ———, *On the convergence of projected gradient processes to singular critical points*, J. Optim. Theory Appl., 55 (1987), pp. 203–216.

[GaB84] E. M. GAFNI AND D. P. BERTSEKAS, *Two-metric projection methods for constrained optimization,* SIAM J. Control Optim., 22 (1984), pp. 936–964.

[GaD88] M. GAWANDE AND J. C. DUNN, *Variable metric gradient projection processes in convex feasible sets defined by nonlinear inequalities,* Appl. Math. Optim., 17 (1988), pp. 103–119.

[Gol64] A. A. GOLDSTEIN, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.

[Gol74] ———, *On gradient projection*, in Proc. 12th Allerton Conference Circuits and System Theory, Univ. of Illinois, Allerton Park, IL, 1974, pp. 38–40.

[Hof52] A. J. HOFFMAN, *On approximate solutions of systems of linear inequalities*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 263–265.

[LeP65] E. S. LEVITIN AND B. T. POLJAK, *Constrained minimization methods,* Z. Vychisl. Mat. i Mat. Fiz., 6 (1965), pp. 787–823. (In Russian.) Translation in USSR Comput. Math. and Math. Phys., 6 (1965), pp. 1–50.

[LuT91] Z.-Q. LUO AND P. TSENG, *On the rate of convergence of a class of distributed asynchronous routing algorithms*, Tech. Rep., Dept. of Electrical and Computer Engineering, McMaster Univ., Hamilton, Ontario and Dept. of Mathematics, Univ. of Washington, Seattle, WA, May 1991.

[LuT92a] ———, *On the convergence of the coordinate descent method for convex differentiable minimization*, J. Optim. Theory Appl., 72 (1992), pp. 7–35.

[LuT92b] ———, *On the linear convergence of descent methods for convex essentially smooth minimization,* SIAM J. Control Optim., 30 (1992), pp. 408–425.

[LuT92c] ———, *Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem*, SIAM J. Optimization, 2 (1992), pp. 43–54.

[MaD88] O. L. MANGASARIAN AND R. DE LEONE, *Error bounds for strongly convex programs and (super)linearly convergent iterative schemes for the least 2-norm solution of linear programs*, Appl. Math. Optim., 17 (1988), pp. 1–14.

[MaS87] O. L. MANGASARIAN AND T.-H. SHIAU, *Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems*, SIAM J. Control Optim., 25 (1987), pp. 583–595.

[Mor89] J. J. MORÉ, *Gradient projection techniques for large-scale optimization problems,* in Proc. 28th Conf. Decision and Control, Tampa, FL, December 1989.

[OrR70] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[Pan87] J.-S. PANG, *A posteriori error bounds for the linearly-constrained variational inequality problem*, Math. Oper. Res., 12 (1987), pp. 474–484.

[Rob73] S. M. ROBINSON, *Bounds for error in the solution set of a perturbed linear program*, Linear Algebra Appl., 6 (1973), pp. 69–81.

[Rob81] ———, *Some continuity properties of polyhedral multifunctions*, Math. Programming Stud., 14 (1981), pp. 206–214.

[Rob82] ———, *Generalized equations and their solutions, part* II: *Applications to nonlinear programming*, Math. Programming Stud., 14 (1982), pp. 200–221.

[Roc70] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[Tsa89] W. K. TSAI, *Convergence of gradient projection routing methods in an asynchronous stochastic quasi-static virtual circuit network*, IEEE Trans. Automat. Control, AC34 (1989), pp. 20–33.

[Tse91] P. TSENG, *Descent methods for convex essentially smooth minimization,* J. Optim. Theory Appl., 71 (1991), pp. 425–463.

[TsB86] J. N. TSITSIKLIS AND D. P. BERTSEKAS, *Distributed asynchronous optimal routing in data networks*, IEEE Trans. Automat. Control, AC31 (1986), pp. 325–332.