# A Primal Affine-Scaling Algorithm for Linearly Constrained Convex Programs *

Clovis C. Gonzaga† Luiz A. Carlos ‡

**Abstract**

The affine-scaling algorithm was initially developed for linear programming problems. Its extension to problems with a nonlinear objective performs at each iteration a scaling followed by a line search along the steepest descent direction. In this paper we prove that any accumulation point generated by this algorithm when applied to a convex function is an optimal solution, under a primal non-degeneracy hypothesis and very general line search procedures.

Keywords: Interior point methods, convex programming, affine-scaling algorithm, global convergence.

## 1 Introduction

In this paper we study the minimization of a convex differentiable function in a set defined by linear constraints by an affine-scaling algorithm.

The affine-scaling algorithm was first stated by Dikin [3], and its convergence was proved by the same author [4] under a primal non-degeneracy hypothesis, and taking steps of unit length in a certain metric. The algorithm was independently rediscovered by Barnes [1] and by Vanderbei, Meketon and Freedman [15], who proved its convergence for large steps, but with both primal and dual non-degeneracy assumptions.

Dikin's proof was reworked by Vanderbei and Lagarias [14], who divide it in a sequence of clear steps. Gonzaga [5] followed their methodology to find a convergence proof for the algorithm with arbitrarily large steps and

*Manuscript completed in 1990, revised in September 2002.

†Department of Mathematics, Federal University of Santa Catarina, Florianópolis, Brazil

‡Department of Mathematics, Federal University of Rio Grande do Norte, Natal, Brazil

primal non-degeneracy. Tsuchiya [13], proved its convergence for degenerate problems using a steplength equal to one eighth of Dikin's unit step. Tseng and Luo [12] proved that the sequences generated by the affine-scaling algorithm for linear programming are always convergent, and this simplified the treatment very much. It was possible to prove convergence to optimal solutions in degenerate problems for steps of 2/3 to the boundary. A clear presentation of this result is in Monteiro, Tsuchiya and Wang [8]. It is also known that the algorithm may converge to non-optimal solutions for large steps, as shown by Mascarenhas [7].

There were several papers extending the affine-scaling algorithm to convex and non-convex quadratic programming problems. For a recent reference surveying these results, see Tseng [11]. The application of the method to linearly constrained convex problems was done for the first time in the present report, which appeared in 1990. The extension of the results to convex objective functions is not immediate, because in this case the sequences are no more necessarily convergent. In fact, for unconstrained problems the method becomes Cauchy's algorithm, for which an example with several different accumulation points is described in Gonzaga [6].

The affine-scaling algorithm is a trust region method in which each iteration minimizes a linear model of the function in a Dikin ellipsoid. The method can be improved by using quadratic models instead of the linear model. This is done in two references: Bonnans and Pola [2] and Monteiro and Wang [9].

In this paper we study the affine-scaling method in its simplest form: each iteration scales the problem, constructs the steepest descent direction (the negative projected gradient direction) and performs a line search along this direction. No auxiliary functions are used, and the line search procedure uses arbitrarily large intervals, ensuring a good improvement in the objective value by a perfect minimization. The algorithm and the convergence analysis can be modified to accommodate practical line search methods like an Armijo or a Goldstein search. We use a primal non-degeneracy hypothesis, but no second derivatives are needed.

Section 2 describes the algorithm and derives general properties of the algorithm mappings, associating with each feasible point a Karush-Kuhn-Tucker estimate. Section 3 shows that the multiplier estimates associated with any convergent subsequence generated by the algorithm converge to optimal multipliers at the limit of the subsequence, thus establishing its optimality.

**Remark on notation:** We denote by $P_M$ the projection matrix into the null space of a matrix $M$. We shall also reserve the letter $e$ for the vector of ones, $e = [1\,1 \cdots 1]^T$, with dimension given by the context. Sub-indices will usually denote components of a vector or different scalars, and super-indices single out different vectors.

## 2 The algorithm

Let $f : I\!R^n \mapsto I\!R$ be a convex continuously differentiable function. The linearly constrained convex programming problem to be studied in this paper is stated as:

$$\begin{array}{lrcl} \text{minimize} & f(x) & & \\ \text{subject to} & Ax & = & b \\ & x_Q & \geq & 0, \end{array} \tag{1}$$

where $b \in I\!R^m$, $A$ is an $m \times n$ full-rank matrix, $0 < m < n$, and $Q \subset \{1, \ldots, n\}$. To simplify the notation, we assume that $Q = \{1, \ldots, q\}$, $q \leq n$, and define $\bar{Q} = \{1, \ldots, n\} - Q$.

**Hypotheses.** (i) The feasible set of (1) has is nonempty, and an initial feasible point $x^o$ such that $x^o_Q > 0$ is available.

Given the initial feasible point $x^o$, we define the following sets:

$$\begin{array}{rcl} S & = & \{x \in I\!R^n \mid Ax = b, x_Q \geq 0, f(x) \leq f(x^o)\}, \\ S^o & = & \{x \in S \mid x_Q > 0\}. \end{array}$$

The set $S$ is the effective feasible set for an algorithm that (as is the case in the present paper) generates a sequence of points with decreasing objective values. $S^o$ is the set of effective feasible interior points.

(ii) The set $S$ defined above is bounded.

In the case in which $f(\cdot)$ is convex, this hypothesis is equivalent to the hypothesis that the set of optimal solutions for (1) is non-empty and bounded.

(iii) Every point of $S$ is primal non-degenerate, according to the definition below.

**The non-degeneracy hypothesis.** Consider a point $x \in S$ and define $N = \{i \in Q \mid x_i = 0\}$, $B = \{1, \ldots, n\} - N$. We shall describe two usual hypotheses coming respectively from linear and nonlinear programming, and prove their equivalence.

(i) From linear programming, a good definition of primal non-degeneracy is the following:

"x is primal non-degenerate if and only if $A_B$ has rank $m$."

Note that only the restricted variables enter the composition of $N$. The usual linear programming problem uses $Q = \{1, \ldots, n\}$, but it can be easily reproduced by introducing dummy restrictions $x_i \geq -M$ for $i \in \bar{Q}$, where $M$ is a very large number: with this artifice the non-degeneracy condition would be as stated above.

(ii) From nonlinear programming, the most usual constraint qualification condition requires the linear independence of the inequality constraint gradients reduced or projected into $\mathcal{N}(A)$. The constraints are $I_i^T x \geq 0$, where $I_i$ is the i-th column of the identity matrix, and the projected constraint gradients are $PI_i = P_i$, where $P$ is the projection matrix into $\mathcal{N}(A)$. It follows that:

"x satisfies the constraint qualification if $\{P_i\}_{i \in N}$ is a linearly independent set."

Both conditions are related by the following lemma:

**Lemma 2.1** *A point $x \in S$ satisfies the constraint qualification if and only if $x$ is primal non-degenerate.*

**Proof:** We shall prove the equivalence of the negation of the assertions in the lemma.

($\leftarrow$): Assume that $x$ does not satisfy the constraint qualification. Then there exists $\lambda \in \mathbb{R}^n$ such that $\lambda_N \neq 0$, $\lambda_B = 0$ and $\sum_{i=1}^n P_i \lambda_i = 0$, or equivalently,

$$P\lambda = 0.$$

This implies that there exists $y \in \mathbb{R}^m$ such that

$$\lambda = A^T y.$$

In particular, $0 = \lambda_B = A_B^T y$. But $y \neq 0$ because $\lambda_N = A_N^T y \neq 0$, and it follows that $A_B$ does not have rank $m$ and $x$ is degenerate.

($\rightarrow$): Assume that $x$ is degenerate. Then for some $y \in \mathbb{R}^m$, $y \neq 0$,

$$A_B^T y = 0.$$

Defining $\lambda_N = A_N^T y$ and $\lambda_B = 0$, we have that $\lambda = A^T y$.

Since $y \neq 0$ and $A$ has maximum rank, $\lambda \neq 0$ and thus $\lambda_N \neq 0$.

On the other hand, $P\lambda = \sum_{i \in N} P_i \lambda_i = 0$, because $\lambda = A^T y \perp \mathcal{N}(A)$. It follows that $\{P_i\}_{i \in N}$ is linearly dependent, completing the proof. ∎

**The scaling matrix.** The affine-scaling algorithm will use at each iteration a scaling matrix. With each point $x \in S$ we shall associate the $n \times n$ matrix $D_x$ defined by

$$D_x \;=\; \operatorname{diag}(x_1, \ldots, x_q, 1, 1, \ldots, 1) \;=\; \operatorname{diag}(x_Q, e). \tag{2}$$

If $x_Q > 0$, then this matrix defines a scaling operation $x = D_x y$ which transfers the point $x = (x_Q, x_{\bar{Q}})$ to $(e, x_{\bar{Q}})$. That is, the constrained variables are transported to the vector of ones.

**Lemma 2.2** *Consider a point $x \in S$. If $x$ is non-degenerate, then the matrix $AD_x^2A^T$ is non-singular.*

**Proof:** The matrix

$$AD_x^2A^T = (AD_x)^T(AD_x)$$

is non-singular if and only if $AD_x$ has rank $m$. If $x$ is non-degenerate, then $A_B$ has rank $m$, and therefore full rank. Consequently $A_B(D_x)_B$ has full rank, since $(D_x)_B > 0$, completing the proof. ∎

It follows from this lemma and the compactness of $S$ that if every point $x \in S$ is non-degenerate, then the map

$$x \in S \mapsto (AD_x^2A^T)^{-1}$$

is continuous and bounded.

**The affine-scaling algorithm.** We now state the affine-scaling algorithm in its most general format. It can certainly be shortened, but this statement settles the notation to be used throughout the paper, and stresses the geometrical interpretation of scaling and projection.

We use the simplified notation $D_k = D_{x^k}$.

**Algorithm 2.3** *Affine-scaling: given $\theta \in (0, 1)$ and given $x^o \in S^o$.*

$k = 0$.

REPEAT

Multipliers in scaled space: $\bar{z} = P_{AD_k} D_k \nabla f(x^k)$.

Direction in scaled space: $\bar{d}^k = -\bar{z}/\|\bar{z}\|$.

Direction and multipliers in original space: $d^k = D_k \bar{d}^k$, $z^k = D_k^{-1} \bar{z}$.

The maximum step: $\lambda_m = \sup\{\lambda \mid x_Q^k + \lambda d_Q^k \geq 0\}$.

The step: Choose arbitrarily $\bar{\lambda} \in [\theta, \lambda_m)$ and set

$$\begin{aligned} \lambda_k &\in \underset{\lambda}{\operatorname{argmin}}\{f(x^k + \lambda d^k) \mid \lambda \in [0, \bar{\lambda}]\}, \\ h^k &= \lambda_k d^k, \\ x^{k+1} &= x^k + h^k. \end{aligned}$$

$k = k + 1$.

UNTIL convergence.

We must show that the step is well defined in any iteration of the algorithm. Given $x^o$, the set $S$ is well defined, and by construction all iterates $x^k$ belong to $S$. Given $x^k \in S^o$, scaling by $x = D_k y$ transfers $x_Q^k$ to the vector of ones. Since $e + \bar{d}_Q^k \geq 0$ because $\|\bar{d}^k\| = 1$, it follows that $x_Q^k + d_Q^k \geq 0$. Thus $\lambda_m \geq 1$ and the interval $(\theta, \lambda_m)$ is non-empty for $\theta < 1$. The existence of a minimizer along the line is a consequence of hypothesis (ii).

The algorithm allows a very free choice of $\bar{\lambda}$. As particular cases, the two most usual choices are allowed: the first one consists in taking $\bar{\lambda} = \theta < 1$, a fixed value. The value $\theta = 1$ can be used in linear programming, but not in the case of a general convex function. This corresponds to Dikin's unit step. The second choice is $\bar{\lambda} = \nu \lambda_m$, where $\nu$ is a fixed percentage of the maximum step, usually taken for linear programming around $\nu = 0.99$.

A fixed step has another interesting interpretation: it corresponds to the minimization of the linear approximation of the function in an ellipsoidal trust region, with radius characterized by $\theta$.

**KKT multiplier estimates.** The problem (1) has $m$ equality and $q$ inequality constraints. The gradient of the Lagrangean function at a point $x \in S$ can be written as

$$\nabla f(x) - A^T w - \sum_{i \in Q} z_i I_i.$$

A better format is obtained by defining a vector $z \in I\!R^n$ with $z_{\bar{Q}} = 0$, and setting

$$\nabla l(x, w, z) = \nabla f(x) - A^T w - z.$$

The Karush-Kuhn-Tucker conditions now say that $\hat{x} \in S$ is an optimal solution for (1) if and only if there exist $w \in I\!R^m$ and $z \in I\!R^n$ such that

$$\begin{aligned} \nabla f(\hat{x}) - A^T w - z &= 0 \\ z_Q &\geq 0 \\ z_{\bar{Q}} &= 0 \\ z_i x_i &= 0, \ i = 1, 2, \ldots, n. \end{aligned}$$

A simplified set of optimality conditions are obtained by projecting the first equation into $\mathcal{N}(A)$, giving

$$P_A \nabla f(\hat{x}) = P_A z.$$

These conditions correspond to a linearization of the problem around $\hat{x}$. The vector $z$ can then be interpreted as a dual slack for the linearized problem. The components $z_{\bar{Q}}$ can be interpreted as multipliers for the dummy constraints $x_i \geq -M$. We shall disregard the multipliers $w$, and work only with $z$: from here on, the expression "multiplier estimates" will refer to the multipliers $z \in I\!R^n$. The first $q$ multipliers are meaningful, and the other will hopefully converge to zero as the algorithm proceeds.

More formally, we shall consider a "multiplier estimate" at a point $x \in S$ any vector $z \in I\!R^n$ such that

$$P_A z = P_A \nabla f(x),$$

and we will show that the algorithm naturally constructs good multiplier estimates. A good estimate should in some way approach the other conditions above, namely $z_i x_i = 0$, $z_Q \geq 0$ and $z_{\bar{Q}} = 0$.

**The choice of multiplier estimates.** Consider the scaled problem at any iteration $k$ of the algorithm. We have $\bar{z} = P_{AD_k} D_k \nabla f(x^k)$, and obviously $P_{AD_k} \bar{z} = \bar{z}$. It follows that $\bar{z}$ is a multiplier estimate for the scaled problem.

Now we observe that $D_k^{-1} \bar{z}$ is a multiplier estimate for the original problem. In fact, by definition of projection into $\mathcal{N}(AD_k)$,

$$D_k \nabla f(x^k) = D_k A^T w + \bar{z}$$

for some $w \in I\!R^m$. Multiplying by $(D_k)^{-1}$, the result follows.

Summing up, each iteration of the algorithm provides a multiplier estimate given by

$$z^k = D_k^{-1} P_{AD_k} D_k \nabla f(x^k), \tag{3}$$

or equivalently, by expanding $P_{AD_k} = I - D_k A^T (AD_k^2 A^T)^{-1} AD_k$,

$$z^k = (I - A^T (AD_k^2 A^T)^{-1} AD_k^2) \nabla f(x^k). \tag{4}$$

Our greatest effort in this paper will be in showing that these multiplier estimates will approach optimal multipliers at accumulation points of the sequence of points generated by the algorithm.

**The algorithmic maps.** The algorithm associates with each iterate $x^k \in S^o$ a multiplier estimate $z^k$ given by expression (4). Using lemma 2.2, this mapping can be extended to $S$, and we can define:

$$x \in S \;\mapsto\; z(x) = (I - A^T (AD_x^2 A^T)^{-1} AD_x^2) \nabla f(x), \tag{5}$$

$$x \in S^o \;\mapsto\; d(x) = -D_x \frac{D_x z(x)}{\|D_x z(x)\|}. \tag{6}$$

The map $z(\cdot)$ is continuous at every point $x \in S$, and $S$ is bounded. It follows that $z(\cdot)$ is bounded and uniformly continuous in $S$. The lemmas below summarize the important properties of these maps.

**Lemma 2.4** *Consider an arbitrary point $x \in S$ and an arbitrary direction $h \in \mathcal{N}(A)$. Then*

$$\nabla f(x)^T h = z(x)^T h.$$

**Proof:** From (5),

$$z(x) = \nabla f(x) - A^T w,$$

where $w \in \mathbb{R}^m$. Since $A^T w \perp \mathcal{N}(A)$, the result follows immediately, completing the proof. ∎

**Lemma 2.5** *At any iteration of the algorithm,*

$$\nabla f(x^k)^T d^k = -\|D_k z^k\|. \tag{7}$$

**Proof:** Consider an iteration $k$. Using lemma 2.4 and (6),

$$\begin{aligned} \nabla f(x^k)^T d^k &= (z^k)^T d^k \\ &= -(z^k)^T D_k \frac{D_k z^k}{\|D_k z^k\|}, \end{aligned}$$

which is equivalent to (7), completing the proof. ∎

**Lemma 2.6** *The sequences generated by the algorithm satisfy*

$$\begin{aligned} D_k z^k &\to 0, \\ x_i^k z_i^k &\to 0,\; i = 1, 2, \ldots, n. \end{aligned}$$

**Proof:** Assume by contradiction that in some subsequence with indices $\mathcal{K}^o \subset \mathbb{N}$

$$\|D_k z^k\| \geq \delta > 0.$$

Since $S$ is bounded, $(x^k)_{k\in\mathcal{K}^o}$ has a convergent subsequence, and hence we can define $\mathcal{K} \subset \mathcal{K}^o$, $\tilde{x} \in S$ and $\tilde{z} \in I\!R^n$ such that

$$x^k \stackrel{\mathcal{K}}{\to} \tilde{x} \quad , \quad z^k \stackrel{\mathcal{K}}{\to} \tilde{z}.$$

Since $d^k = D_k^2 z^k / \|D_k z^k\|$ and $\|D_k z^k\| > \delta$ for $k \in \mathcal{K}$,

$$d^k \stackrel{\mathcal{K}}{\to} \tilde{d} = D_{\tilde{x}}^2 \tilde{z} / \|D_{\tilde{x}} \tilde{z}\|.$$

Let $\theta \in (0,1)$ be the constant in the algorithm initialization, and choose an arbitrary number $\lambda \in [0, \theta]$. We have:

$$x^k + \lambda d^k \in S.$$

By construction (line search) and by the differentiability of $f(\cdot)$,

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k + \lambda d^k) \\ &= f(x^k) + \lambda \nabla f(x^k)^T d^k + o(x^k, \lambda d^k). \end{aligned}$$

Our hypothesis together with lemma 2.5 say that for $k \in \mathcal{K}$

$$\nabla f(x^k) d^k = -\|D_k z^k\| \leq -\delta,$$

and hence for $k \in \mathcal{K}$

$$f(x^{k+1}) - f(x^k) \leq -\lambda\delta + o(x^k, \lambda d^k).$$

The sequence $(f(x^k))$ decreases monotonically and is bounded. Thus $f(x^{k+1}) - f(x^k) \to 0$. Taking limits for $k \in \mathcal{K}$, and noting that the map $(x, \Delta x) \in I\!R^{2n} \mapsto o(x, \Delta x)$ is continuous because $f(\cdot)$ is continuously differentiable,

$$0 \leq -\lambda\delta + o(\tilde{x}, \lambda\tilde{d}).$$

Since $\lambda$ was chosen arbitrarily in $[0, \theta]$, we can take limits in this last expression, and see that

$$\lim_{\lambda\to 0} \frac{o(\tilde{x}, \lambda\tilde{d})}{\lambda} \geq \delta,$$

contradicting the differentiability of $f(\cdot)$ at $\tilde{x}$, and proving the first result.

The second result is a straightforward consequence of the first: for $i \in Q$, $x_i^k z_i^k = (D_k z^k)_i$; for $i \in \bar{Q}$, $(D_k z^k)_i = z_i^k$ and hence $x_i^k z_i^k \to 0$ because $(x_i^k)$ is bounded. This establishes the second result, and completes the proof. ∎

This lemma shows that if $\tilde{x}$ is an accumulation point of the sequence $(x^k)$, then $D_{\tilde{x}} z(\tilde{x}) = 0$. This immediately implies for $\tilde{z} = z(\tilde{x})$,

$$\tilde{z}_{\bar{Q}} = 0 \quad \text{and} \quad \tilde{z}_i \tilde{x}_i = 0,\ i = 1, \ldots, n,$$

since $D_{\tilde{x}} = \text{diag}(\tilde{x}_Q, e)$. The only KKT condition that remains to be proved is that $z_Q \geq 0$, and this will be the purpose of the next section.

The next lemma is very similar to the last one.

**Lemma 2.7** *The sequences generated by the algorithm satisfy*

$$(z^k)^T h^k \to 0.$$

**Proof:** The proof is very similar to the proof of lemma 2.6. By lemma 2.4, $(z^k)^T h^k = \nabla f(x)^T h^k$ and since $h^k$ is a descent direction, $(z^k)^T h^k < 0$.

Assume by contradiction that in some subsequence with indices $\mathcal{K}^o \subset \mathbb{N}$

$$(z^k)^T h^k \leq -\delta < 0.$$

Since $S$ is bounded,the sequences $(x^k)$, $(z^k)$ and $(h^k)$ with $h^k = x^{k+1} - x^k$ are bounded, and we can choose a subsequence $\mathcal{K} \subset \mathcal{K}^o$ such that

$$x^k \stackrel{\mathcal{K}}{\to} \tilde{x}\ ,\ z^k \stackrel{\mathcal{K}}{\to} \tilde{z}\ ,\ h^k \stackrel{\mathcal{K}}{\to} \tilde{h}\ ,\ (z^k)^T h^k \stackrel{\mathcal{K}}{\to} \tilde{z}^T \tilde{h} \leq -\delta,$$

where $\tilde{x} \in S$, $\tilde{z} \in \mathbb{R}^n$, $\tilde{h} \in \mathcal{N}(A)$.

From lemma 2.4, $(z^k)^T h^k = \nabla f(x^k)^T h^k$, and we can write

$$f(x^k + \lambda h^k) = f(x^k) + \lambda (z^k)^T h^k + o(x^k, \lambda h^k).$$

By construction (line search), for any $\lambda \in [0, 1]$,

$$f(x^{k+1}) \leq f(x^k + \lambda h^k),$$

and hence

$$\begin{aligned} f(x^{k+1}) - f(xk) &\leq \lambda (z^k)^T h^k + o(x^k, \lambda h^k) \\ &\leq -\lambda\delta + o(x^k, \lambda h^k). \end{aligned}$$

As we did in the proof of lemma 2.6, we can take limits of this expression for $k \in \mathcal{K}$, obtaining

$$0 \leq -\delta\lambda + o(\tilde{x}, \lambda\tilde{h}).$$

for all $\lambda \in [0,1]$. This contradicts the differentiability of $f(\cdot)$ at $\tilde{x}$, completing the proof. ∎

**Remark on convexity:** Up to now, the convexity of $f(\cdot)$ was not used, unless in the point where we stated the KKT conditions as necessary and sufficient optimality conditions. If the objective function is not convex, then these conditions are only necessary. All the lemmas proved in this section remain valid in the non-convex case. This is unfortunately not the case with respect to the the material in the next section, where the convexity of the objective function will be explicitly used.

## 3 Convergence of the algorithm

Consider an application of the algorithm, and the sequences $(x^k)$, $(z^k)$, $(d^k)$, $(\lambda_k)$, $(h^k)$ generated by it. We shall now choose an accumulation point $\tilde{x}$ of $(x^k)$, which will be fixed to the end of the convergence proof. We shall prove that the multiplier estimate $\tilde{z} = z(\tilde{x})$ is a Karush-Kuhn-Tucker multiplier associated with $\tilde{x}$, thus proving that this is an optimal solution to the convex programming problem. Note that we do not prove that the whole sequence converges: we prove that any accumulation point of the sequence is optimal.

Let then $\tilde{x}$ be an accumulation point of $(x^k)$, and define $\tilde{z} = z(\tilde{x})$. Define the sets

$$N = \{i = 1, 2, \ldots, n \mid \tilde{z}_i \neq 0\}, \; \bar{N} = \{1, \ldots, n\} - N, \tag{8}$$

and the constant

$$\epsilon = \frac{1}{2} \min_{i \in N} |\tilde{z}_i|. \tag{9}$$

From the last section we know that $\tilde{z}_{\bar{Q}} = 0$ and that $\tilde{x}_i \tilde{z}_i = 0$ for all $i = 1, \ldots, n$. Hence $N \subset Q$ and $\tilde{x}_N = 0$.

**The terminal set.** The point $\tilde{x}$ belongs to the face of the feasible set characterized by $x_N = 0$. In this face we define the following *terminal set*:

$$\Omega = \{x \in S \mid x_N = 0, f(x) = f(\tilde{x})\}. \tag{10}$$

**Lemma 3.1** *The set $\Omega$ defined above is convex.*

**Proof:** Let $x$ be an arbitrary point in the convex hull $\operatorname{co}\Omega$. We must prove that $x \in \Omega$.

Since $x$ is a combination of points in $\Omega$, $x_N = 0$, and from the convexity of $f(\cdot)$, $f(x) \leq f(\tilde{x})$. It remains to be proved that $f(x) \geq f(\tilde{x})$.

Consider the direction $h = x - \tilde{x}$. Note that $h_N = 0$. From the convexity of $f(\cdot)$,

$$f(x) \geq f(\tilde{x}) + \nabla f(\tilde{x})^T h.$$

But $h \in \mathcal{N}(A)$ and by lemma 2.4, $\nabla f(\tilde{x})^T h = \tilde{z}^T h = \tilde{z}_N^T h_N = 0$.
This shows that $f(x) \geq f(\tilde{x})$, completing the proof. ∎

The next lemma is a known result in convex analysis, but we provide a proof for completeness:

**Lemma 3.2** *Let $f : I\!R^n \to I\!R$ be a convex continuously differentiable function. If $f(\cdot)$ is constant on a convex set $\Omega \subset I\!R^n$ then $\nabla f(\cdot)$ is constant on $\Omega$.*

**Proof:** Let epi$f = \{z = (x, y) \in I\!R^{n+1} \mid x \in I\!R^n, y \geq f(x)\}$ be the epigraph of $f(\cdot)$. It is known (see for instance Stoer and Witzgall [10]) that given $x \in I\!R^n$, $g = \nabla f(x)$ if and only if $\gamma = (g, -1)$ defines a hyperplane $\gamma^T z = \alpha$ that separates $(x, f(x))$ from epi$f$.

Let $f(\cdot)$ be constant on the convex set $\Omega$, and let $\bar{x}$ be a point in its relative interior relint$\Omega$. Define the set

$$\Psi = \{(x, f(\bar{x})) \mid x \in \Omega\}.$$

Now $\bar{z} = (\bar{x}, f(\bar{x})) \in$ relint$\Psi$, $\bar{\gamma} = (\nabla f(\bar{x}), -1)$ defines a separating hyperplane $\Pi$ at $\bar{z}$, with equation $\bar{\gamma}^T z = \bar{\alpha}$. In particular, the hyperplane separates $\bar{z}$ from $\Psi$.

It is known that if a hyperplane $\Pi$ separates $z \in$ relint$\Psi$ from $\Psi$, then $\Psi \subset \Pi$, and $\Pi$ separates any $z \in \Psi$ from $\Psi$. It follows that $\bar{\gamma} = (\nabla f(\bar{x}), -1)$ defines a separator at any point $z \in \Psi$. Hence $\nabla f(x) = \nabla f(\bar{x})$ for all $x \in \Omega$, completing the proof. ∎

**Lemma 3.3** *For any $x \in \Omega$, $z(x) = \tilde{z}$.*

**Proof:** $\Omega$ is a convex set on which $f(\cdot)$ is constant. By lemma 3.2, $\nabla f(\cdot)$ is constant in $\Omega$.

Let $x \in \Omega$ be chosen arbitrarily.

Note initially that from (5), $\tilde{z} = \nabla f(\tilde{x}) + A^T w$ for some $w \in I\!R^m$. It follows that $\nabla f(x) = \nabla f(\tilde{x}) = \tilde{z} - A^T w$. Substituting this into the definition (5), we get

$$z(x) = (I - A^T (AD_x^2 A^T)^{-1} AD_x^2)(\tilde{z} - A^T w).$$

To simplify this expression, note initially that $D_x\tilde{z} = 0$, because for $i \in N$, $(D_x\tilde{z})_i = x_i\tilde{z}_i = 0$ and for $i \notin N$, $\tilde{z}_i = 0$ by definition of $N$. It follows that

$$\begin{aligned} z(x) &= \tilde{z} - A^T w + A^T(AD_x^2A^T)^{-1}AD_x^2A^T w \\ &= \tilde{z} - A^T w + A^T w \\ &= \tilde{z}, \end{aligned}$$

completing the proof. ∎

Now consider the terminal set $\Omega$. Given a number $\tau > 0$, we shall define a neighborhood $\Omega_\tau$ of $\Omega$ by

$$\Omega_\tau = \{x \in S \mid \|x - y\|_\infty \leq \tau \text{ for some } y \in \Omega\}. \quad (11)$$

The next lemma characterizes a small neighborhood of $\Omega$, in the sense that the multiplier estimates have small variations. And we know what is "small" for the multiplier estimates: the value $\epsilon$ defined in (9). The second part of the lemma shows the use of such a small neighborhood: the direction computed by the algorithm at a point in it disagrees in sign with $\tilde{z}$ for all components in $N$. This property will be crucial for proving that a negative component in $\tilde{z}$ prevents the sequence from accumulating in $\tilde{x}$.

**Lemma 3.4** *Let $\epsilon$ be as defined in (9). There exists a number $\delta > 0$ such that for all $x \in \Omega_\delta$,*
*(i) $\|z(x) - \tilde{z}\| \leq \epsilon$,*
*(ii) If $x > 0$, then $d_i(x)\tilde{z}_i < 0$ for all $i \in N$.*

**Proof:**
(i) Since $z(\cdot)$ is uniformly continuous in $S$, there exists $\delta > 0$ such that for all $x, y \in S$,

$$\|x - y\|_\infty \leq \delta \Rightarrow \|z(x) - z(y)\| \leq \epsilon.$$

From the definition of $\Omega_\delta$, for each $x \in \Omega_\delta$ there exists $y \in \Omega$ such that $\|x - y\|_\infty \leq \delta$. Using lemma 3.3, we obtain $z(y) = \tilde{z}$, and it follows that $\|z(x) - \tilde{z}\| \leq \epsilon$, proving (i).

(ii) Assume that $x > 0$. By definition of $\epsilon$, for any $i \in N$

$$|z_i(x) - \tilde{z}_i| \leq \|z(x) - \tilde{z}\| \leq \epsilon \leq \frac{1}{2}|\tilde{z}_i|,$$

and thus $z_i(x)$ and $\tilde{z}_i$ agree in sign. From the definition of $d(x)$ in (6), $d_i(x)$ and $z_i(x)$ have opposite signs, completing the proof. ∎

We are now ready for the main results, to be proved in three lemmas. The first lemma relates other accumulation points of $(x^k)$ to the terminal set $\Omega$ associated with $\tilde{x}$. Although it is quite trivial, this lemma is the cornerstone of the whole treatment.

**Lemma 3.5** *Let $\hat{x}$ be an accumulation point of $(x^k)$. Either $\hat{x} \in \Omega$ or $\hat{x} \notin \Omega_\delta$.*

**Proof:** Assume by contradiction that $\hat{x}$ is an accumulation point of $(x^k)$ such that $\hat{x} \in \Omega_\delta$ and $\hat{x} \notin \Omega$.
Since the sequence $(f(x^k))$ is decreasing, we must have $f(\hat{x}) = f(\tilde{x})$, and thus the hypothesis requires that for some $i \in N$, $\hat{x}_i > 0$. By lemma 3.4, $z_i(\hat{x}) \neq 0$, and it follows that $\hat{x}_i z_i(\hat{x}) \neq 0$. This contradicts lemma 2.6, and completes the proof. ∎

**Lemma 3.6** *Let $\tilde{x}$, $\tilde{z} = z(\tilde{x})$ and $\delta$ be as constructed above. If $\tilde{z}$ has a negative component, then $(x^k)$ has an accumulation point $\bar{x}$ such that $\bar{x} \notin \Omega_\delta$.*

**Proof:** Assume that $\tilde{z}_N$ has a negative component, and assume by contradiction that there exists $\bar{k} \in \mathbb{N}$ such that for $k \geq \bar{k}$, $x^k \in \Omega_\delta$. Choose $i \in N$ such that $\tilde{z}_i < 0$.
By lemma 3.4, $d_i^k \tilde{z}_i < 0$ for $k \geq \bar{k}$, and thus $d_i^k > 0$. It follows that for $k \geq \bar{k}$,

$$x_i^{k+1} = x_i^k + \lambda_k d_i^k \geq x_i^k,$$

and the sequence $(x_i^k)_{k \geq \bar{k}}$ increases monotonically. This contradicts the fact that $\liminf x_i^k = 0$, completing the proof. ∎

The last lemma completes the proof that $\tilde{z}$ is an optimal multiplier, showing that $\tilde{x}$ is an optimal solution of problem (1).

**Lemma 3.7** *Let $\tilde{x}$ and $\tilde{z} = z(\tilde{x})$ be as constructed above. Then $\tilde{z}_N > 0$.*

**Proof:** Assume by contradiction that $\tilde{z}_N$ has a negative component. We shall establish a sequence of facts.

(i) Fact: There exists $\mathcal{K} \subset \mathbb{N}$ such that $x^k \xrightarrow{\mathcal{K}} \bar{x}^1$ and $x^{k+1} \xrightarrow{\mathcal{K}} \bar{x}^2$, with $\bar{x}^1 \in \Omega$ and $\bar{x}^2 \notin \Omega_\delta$.
Proof: Since there are accumulation points both in $\Omega$ and out of $\Omega_\delta$, there must exist $\mathcal{K}^o \subset \mathbb{N}$ such that for all $k \in \mathcal{K}^o$, $x^k \in \Omega_\delta$ and $x^{k+1} \notin \Omega_\delta$.

Since the sequence $(x^k)$ is bounded, there must exist $\mathcal{K} \subset \mathcal{K}^o$, $\bar{x}^1$ and $\bar{x}^2$ such that $x^k \xrightarrow{\mathcal{K}} \bar{x}^1$ and $x^{k+1} \xrightarrow{\mathcal{K}} \bar{x}^2$. We must prove that $\bar{x}^1 \in \Omega$ and $\bar{x}^2 \notin \Omega_\delta$.

Since $\Omega_\delta$ is closed, $\bar{x}^1 \in \Omega_\delta$, and by lemma 3.5 $\bar{x}^1 \in \Omega$.
Assuming by contradiction that $\bar{x}^2 \in \Omega_\delta$, lemma 3.5 requires that $\bar{x}^2 \in \Omega$. Then for sufficiently large $k \in \mathcal{K}$, $\|x^{k+1} - \bar{x}^2\|_\infty < \delta$, which implies that $x^{k+1} \in \Omega_\delta$, contradicting the definition of $\mathcal{K}^o$ and proving that $\bar{x}^2 \notin \Omega_\delta$.

(ii) Fact: There exists $j \in N$ such that $\tilde{z}_j < 0$ and $\bar{x}_j^2 > \bar{x}_j^1 = 0$.
Proof: Since $f(\bar{x}^1) = f(\bar{x}^2)$ and $\bar{x}^2 \notin \Omega$, we must have $\bar{x}_N^2 \neq 0$. Hence for some $j \in N$, $\bar{x}_j^2 > \bar{x}_j^1 = 0$.

Since $x_j^{k+1} - x_j^k \xrightarrow{\mathcal{K}} \bar{x}_j^2 - \bar{x}_j^1$, for large $k \in \mathcal{K}$

$$h_j^k = x_j^{k+1} - x_j^k > 0.$$

Thus for large $k \in \mathcal{K}$, $x^k \in \Omega_\delta$ and $d_j^k = h_j^k/\lambda_k > 0$. Using lemma 3.4, $\tilde{z}_j < 0$, proving (iii).

Now consider the steps $h^k = x^{k+1} - x^k$, and study the sequence of values $(z^k)^T h^k$. From lemma 2.7

$$(z^k)^T h^k \to 0.$$

Taking limits in the set $\mathcal{K}$ defined above, we obtain

$$(z^k)^T h^k \xrightarrow{\mathcal{K}} \tilde{z}^T(\bar{x}^2 - \bar{x}^1) = \tilde{z}_n^T(\bar{x}_N^2 - \bar{x}_N^1),$$

and it follows that

$$\tilde{z}_n^T(\bar{x}_N^2 - \bar{x}_N^1) = 0. \tag{12}$$

For large $k \in \mathcal{K}$, $x^k \in \Omega_\delta$ and by lemma 3.4 $z_i^k h_i^k < 0$ for $i \in N$. Taking limits of these products for $k \in \mathcal{K}$, we conclude that

$$\tilde{z}_i(\bar{x}_i^2 - \bar{x}_i^1) \leq 0$$

for $i \in N$. From (iii), one of these inequalities must be strict, leading to

$$\tilde{z}_N^T(\bar{x}_N^2 - \bar{x}_N^1) < 0,$$

contradicting (12) and completing the proof. ■

Thus, $\tilde{x}$ is a feasible solution and $\tilde{s}$ is an optimal KKT multiplier associated with $\tilde{x}$, which completes the proof that all accumulation points of the sequence $(x^k)$ are optimal solutions for our convex optimization problem.

## References

[1] E. R. Barnes. A variation on Karmarkar's algorithm for solving linear programming problems. *Mathematical Programming*, 36:174-182, 1986.

[2] J. F. Bonnans and C. Pola. A trust region interior point algorithm for linearly constrained optimization. *SIAM J. Optim.*, 7:717-731, 1997

[3] I. I. Dikin. Iterative solution of problems of linear and quadratic programming. *Soviet Mathematics Doklady*, 8:674-675, 1967.

[4] I. I. Dikin. On the speed of an iterative process. *Upravlyaemye Sistemi*, 12:54-60, 1974.

[5] C. Gonzaga. *Convergence of the Large Steps Primal Affine Scaling Algorithm for Primal Non-degenerate Linear Programming.* Internal report, COPPE - Federal University of Rio de Janeiro, Rio de Janeiro, Brasil, 1990.

[6] C. C. Gonzaga. Two facts on the convergence of the Cauchy algorithm. *Journal on Optimization Theory and Applications*, 107:593–602, 2000. The affine-scaling algorithm fails for stepsize 0.999. *SIAM J. Optim.*, 7:34-46, 1997

[7] W. F. Mascarenhas. The affine-scaling algorithm fails for stepsize 0.999. *SIAM J. Optim.*, 7:34-46, 1997

[8] R. D. C. Monteiro, T. Tsuchiya and Y. Wang. A simplified global convergence proof of the affine-scaling algorithm. *Ann. Oper. Res.*, 46/47:443-482, 1993.

[9] R. D. C. Monteiro and Y. Wang. Trust region affine-scaling algorithms for linearly constrained convex and concave programs. *Math. Programming*, 80:283-313, 1998.

[10] J. Stoer and C. Witzgall. *Convexity and Optimization in Finite Dimensions I.* Springer Verlag, Berlin, 1970.

[11] P. Tseng. *Convergence properties of Dikin's affine scaling algorithm for nonconvex quadratic minimization*, Report, Dept. of Mathematics, University of Washington, Seattle, 2001.

[12] P. Tseng and Z.-Q. Luo. On the convergence of the affine-scaling algorithm. *Math. Programming*, 56:301-319, 1992.

[13] T. Tsuchiya. Global convergence of the affine scaling methods for degenerate linear programming problems. *Math. Programming*, 52:377-404, 1991.

[14] R. Vanderbei and J. C. Lagarias. I. I. Dikin's convergence result for the affine-scaling algorithm. *Contemp. Math*, 114:109-119, 1990.

[15] R. J. Vanderbei, M. J. Meketon, and B. A. Freedman. A modification of Karmarkar's linear programming algorithm. *Algorithmica*, 1:395-407, 1986.