Interior-Point Gradient Methods with Diagonal-Scalings for Simple-Bound Constrained Optimization

Yin Zhang

CAAM Technical Report TR04-06
Department of Computational
and Applied Mathematics
Rice University, Houston, Texas

April, 2004

Abstract

In this paper, we study diagonally scaled gradient methods for simple-bound constrained optimization in a framework almost identical to that for unconstrained optimization, except that iterates are kept within the interior of the feasible region. We establish a satisfactory global convergence theory for such interior-point gradient methods applied to Lipschitz continuously differentiable functions without any further assumption. Moreover, a strong convergence result is obtained for a class of so-called L-nonlinear functions introduced in this paper which includes virtually all nonlinear functions that do not contain linear pieces.

Key Words. Interior point gradient method, consistent scaling, L-nonlinear function, global convergence.

1 Introduction

A fundamental class of methods in optimization is gradient methods. Without requiring second-order derivative information, these methods are matrix-free, simple to understand and easy to implement. Despite the disadvantage of often having a slow convergence rate, gradient methods are still widely used in practice, especially in large-scale applications where high-accuracy is not essential and the alternatives, such as Newton-type methods, are not affordable.

In this paper we consider a general framework to extend gradient methods from unconstrained optimization problems to simple-bound constrained problems,

$$\min\{f(x) : a \le x \le b\}, \tag{1}$$

where $f : \Re^n \to \Re$ is a continuously differentiable function, the two vectors a and b satisfy $a < b$ component-wise, and take values in extended $\Re^n$ space which includes the negative and positive infinity. In other words, for any variable $[x]_i$, the i-th component of x, there may exist no (finite) bound, a single lower or upper bound, or two bounds one on each side. It is well-known that the necessary optimality (Karush-Kuhn-Tucker or KKT) conditions for this problem are that for a variable $x \in \Re^n$ together with multipliers $u, v \in \Re^n$,

$$\nabla f(x) = u - v,\ (x-a) \circ u = 0,\ (b-x) \circ v = 0,\ x-a, b-x, u, v \ge 0, \tag{2}$$

where the symbol "$\circ$" denotes the Hadamard (or element-wise) multiplication, and the convention $\infty * 0 = 0$ is assumed. A point x is called a KKT point if it satisfies the KKT conditions.

Generalizations of gradient methods from unconstrained to constrained optimization is a subject that has been intensively studied in the last five decades or so. When feasibility sets are convex, most generalizations fall into the framework of gradient projection methods, such as those proposed in [7, 10, 11, 12, 16], to cite a few early works in this direction. In the gradient projection framework, iterates travel on the boundary of feasibility sets. In this paper, however, we consider an interior-point approach that keeps iterates in the interior of feasibility sets.

We will analyze the convergence of interior-point algorithms in a simple line search framework that can be considered an extension of *primal affine-scaling gradient algorithms*. Affine-scaling was first introduced by Dikin [6] for linear and quadratic programming, followed by a sizable amount of literature mostly on linear and quadratic programming (see [18] for a recent list of references). In addition, there are also quite a number of works on primal affine-scaling algorithms for nonlinear programming by Coleman and Li [2, 3] and others (see [4, 5, 9], for example), that are either Newton-type or trust-region methods using quadratic approximations. On the other hand, there seem to exist fewer works on extending primal affine-scaling gradient methods to nonlinear programming in a line search framework. One such work is by Gonzaga and Carlos [8] who established convergence for a primal affine-scaling gradient algorithm for convex programming with linear constraints. In this paper we will consider an extensions to more general objective functions and more general diagonal scalings; but we only consider simple-bound constraints.

1.1 Standard form

We note that through simple linear transformations, any one-sided bound, either lower or upper, can be converted into the form "$[x]_i \geq 0$", and any pair of two-sided bounds can be shifted and scaled into the form "$0 \leq [x]_i \leq 1$". To simplify our notations, we will consider a standard form of simple-bound constrained optimization problems that we describe now. Let us first partition the index set into three mutually exclusive sets, namely,

$$\{i = 1, 2, \cdots, n\} = I_0 \cup I_1 \cup I_2,$$

where I_0 is the set of indices for those variables without a bound, I_1 for those with a single bound, and I_2 for those with bounds on both sides. We note that one or two of these three sets may be empty. In addition, we will make use the following four index sets later,

$$I_1^+(x) = \{i \in I_1 : [\nabla f(x)]_i > 0\}, \quad I_1^-(x) = \{i \in I_1 : [\nabla f(x)]_i \leq 0\}, \tag{3}$$

$$I_2^+(x) = \{i \in I_2 : [\nabla f(x)]_i > 0\}, \quad I_2^-(x) = \{i \in I_2 : [\nabla f(x)]_i \leq 0\}, \tag{4}$$

which obviously all vary with f and x.

Without loss of generality, from now on we will consider the following set of *standard* simple-bound constraints,

$$\mathcal{F} := \{x \in \Re^n : 0 \leq x_{I_1},\ 0 \leq x_{I_2} \leq 1\}, \tag{5}$$

where x_{I_j} represents the sub-vector of x corresponding to the index set I_j for $j = 1$ and 2. Accordingly, our *standard* simple-bound constrained optimization problem takes the form:

$$\min\ \{f(x) : x \in \mathcal{F}\}. \tag{6}$$

It is not difficult to verify the following simple result, but it will play a central role in the development of this paper.

Proposition 1. *The KKT conditions for (6) can be written as*

$$h(x) \circ \nabla f(x) = 0, \quad x \in \mathcal{F}, \tag{7}$$

where the i-th component of $h(x)$, $i = 1, 2, \cdots, n$, is defined as

$$[h(x)]_i = \begin{cases} 1, & i \in I_0 \cup I_1^-(x), \\ \min(1, [x]_i), & i \in I_1^+(x) \cup I_2^+(x), \\ 1 - [x]_i, & i \in I_2^-(x). \end{cases} \tag{8}$$

We observe that (i) $h(x) > 0$ for all $x \in \text{int}(\mathcal{F})$; (ii) $h(x) \leq 1$; and (iii) for $i \in I_1 \cup I_2$, $[h(x)]_i$ has discontinuities only at points where $[\nabla f(x)]_i = 0$. It is worth noting that in most affine-scaling algorithms, one usually defines $[h(x)]_i = [x]_i$ for $i \in I_1$. The theoretical results of this paper would not be affected by this definition as long as the iterate sequence $\{x_k\}$ remains bounded,

Optimality conditions of the form (7) were first used by Coleman and Li [2, 3], and later others, to construct interior-point Newton-type or trust-region algorithms. We will further discuss this connection in the final section of the paper.

For convenience, we will call the equality in (7) the *complementarity* condition, even though it may include more conditions than the usual complementarity conditions.

1.2 Notations

We will use the following notations in this paper. The norm $\|\cdot\|$ always denotes the Euclidean norm unless otherwise specified. We use $\Re^n_{++} = \{x \in \Re^n : x > 0\}$ to denote the positive orthant of $\Re^n$, $\mathcal{B}(x, \rho)$ to denote the open ball centered at x with radius $\rho > 0$, and $\mathbb{N}$ to denote the set of natural numbers. For any set Ω, $\text{cl}(\Omega)$ is the closure of Ω and $\text{int}(\Omega)$ is the interior of Ω. Subscripts are usually iteration count, while the components of a vector are denoted by subscripts after brackets; for example, $[x_k]_i$ is the i-th component of the k-th iterate vector x_k. For an index subset $N \in \{1, 2, \cdots, n\}$, $[x]_N$ is the sub-vector of $x \in \Re^n$ consisting of the components with indices in N. For any vector v, the exponential v^ν and absolute value $|v|$ are taken element-wise.

This paper is organized as follows. In Section 2, we describe the algorithmic framework under consideration and its most basic properties. We present the main theorem of the paper in Section 3, along with necessary definitions. Section 4 contains results that lead to the proof of the main theorem. In Section 5, we include a numerical example to illustrate that different diagonal scalings can result in very different convergence behavior. We conclude the paper in Section 6 with several discussions.

2 Interior-Point Gradient Algorithm

The algorithmic framework under consideration is extremely simple. Starting from a strictly feasible initial guess x_0, we update the iterates by the formula $x_{k+1} = x_k - \alpha_k[d_k \circ \nabla f(x_k)]$ and choose the step α_k so that the next iterate remains in the interior of the feasible set, where d_k is a positive diagonal-scaling vector. This so-called interior-point gradient (IPG) algorithm with diagonal-scaling is formally presented below.

IPG Algorithm:

Choose $\tau \in (0, 1)$ and $\hat{\alpha} \geq 1$. Set $k = 0$, and initialize $x^0 \in \text{int}(\mathcal{F})$.

Do

(1) Let $p_k = -d_k \circ \nabla f(x_k)$ for a scaling vector $d_k > 0$.

(2) Compute $\hat{\alpha}_k = \sup\{\alpha : x_k + \alpha p_k \in \mathcal{F}\}$ and set $\tau_k \in [\tau, 1)$.

(3) Find $\alpha_k \in (0, \min(\hat{\alpha}, \tau_k \hat{\alpha}_k)]$ by the procedure given below.

(4) Set $x_{k+1} = x_k + \alpha_k p_k$, increment k and go to step (1).

End

End Algorithm

To analyze the convergence behavior of the algorithm, we have intentionally left out a stopping criterion so that the algorithm will generate a sequence of infinitely many iterates.

We now describe the procedure for choosing the step $\alpha_k \in (0, \min(\hat{\alpha}, \tau_k \hat{\alpha}_k)]$. Similar to unconstrained optimization, we will attempt to satisfy two well-known line search conditions: the Armijo condition [1] which will always be satisfied, and the Wolfe condition [19, 20] which may or may not be satisfied in our case. Specifically, the Armijo and Wolfe conditions are, respectively,

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f(x_k)^T p_k, \tag{9}$$

$$\nabla f(x_k + \alpha p_k)^T p_k \geq c_2 \nabla f(x_k)^T p_k, \tag{10}$$

where $0 < c_1 < c_2 < 1$. It is well known that the Armijo condition (9) is satisfied by all sufficiently small $\alpha > 0$. On the other hand, although the Wolfe condition (10) can also be simultaneously satisfied by some positive values of α whenever f is bounded below in the direction of p_k, there is no guarantee in general that the smallest of such values will fall within the interval $(0, \min(\hat{\alpha}, \tau_k \hat{\alpha}_k)]$.

In our procedure we choose a value that satisfies both (9) and (10) whenever such a value exists in the interval $(0, \min(\hat{\alpha}, \tau_k \hat{\alpha}_k)]$; otherwise we set $\alpha_k = \min(\hat{\alpha}, \tau_k \hat{\alpha}_k)$. We stress that in either case the Armijo condition (9) always holds.

It is easy to see that, since $x_{k+1} - x_k = -\alpha_k d_k \circ \nabla f(x_k)$, the iterate sequence generated by the algorithm satisfies

$$x_k \in \text{int}(\mathcal{F}), \quad \nabla f(x_k) \circ (x_{k+1} - x_k) \leq 0. \tag{11}$$

Since the function value is monotonically decreasing, we also have

$$\lim_{k \to \infty} f(x_k) = \hat{f} \tag{12}$$

for some $\hat{f} \in \Re \cup \{-\infty\}$. Furthermore, we note that $\alpha_k \leq \hat{\alpha}$ for all iterations. Therefore, if $\{d_k\}$ is bounded, then we will be able to find some $\gamma > 0$ such that for all k

$$|x_{k+1} - x_k| \leq \gamma |\nabla f(x_k)|. \tag{13}$$

The following proposition contains well-known results (see [14], for example) for line search methods which satisfy the conditions (9) and (10). We include a proof for completeness.

Proposition 2. *If $\{f(x_k)\}$ is bounded below and α_k satisfies (9) for every k, then*

$$\lim_{k\to\infty} \nabla f(x_k)^T (x_{k+1} - x_k) = 0. \tag{14}$$

Moreover, if ∇f is Lipschitz continuous with constant L and α_k satisfies (10), then

$$\alpha_k \geq \frac{c_2 - 1}{L} \frac{\nabla f(x_k)^T p_k}{\|p_k\|^2}. \tag{15}$$

Proof. Condition (9) implies

$$f(x_k) - f(x_{k+1}) \geq -c_1 \nabla f(x_k)^T (x_{k+1} - x_k) \geq 0.$$

Since $\{f(x_k)\}$ is monotone and bounded below, we obtain (14) by letting $k \to \infty$.

Condition (10) implies

$$\begin{aligned} (c_2 - 1)\nabla f(x_k)^T p_k &\leq (\nabla f(x_{k+1}) - \nabla f(x_k))^T p_k \\ &\leq \|\nabla f(x_{k+1}) - \nabla f(x_k)\| \|p_k\| \\ &\leq L\|x_{k+1} - x_k\| \|p_k\| = L\alpha_k \|p_k\|^2, \end{aligned}$$

which is (15). □

Since we choose α_k to either satisfy (10) or to be $\min(\hat{\alpha}, \tau_k \hat{\alpha}_k)$, by (15) we have

$$\alpha_k \geq \min\left(\hat{\alpha}, \tau_k \hat{\alpha}_k, \frac{c_2 - 1}{L} \frac{\nabla f(x_k)^T p_k}{\|p_k\|^2}\right). \tag{16}$$

At this point, the only unspecified component in the IPG algorithm is the choice of the diagonal scaling vector d_k at each iteration. The choice $d_k = [h(x_k)]^2$ gives what is called the affine scaling that has been extensively studied in linear and quadratic programming. In the next section, we will provide general conditions for selecting d_k to guarantee convergence.

3 Main Result

We will present our main result, Theorem 1, in this section, which states that the IPG algorithm with *consistent scalings* will generate iterates that either diverge to infinity or cluster at KKT points of (6). In addition, when applied to a class of so-called *L-nonlinear functions*, the whole iterate sequence must converge to a KKT point if it does not diverge. However, we first need to introduce the concepts of consistent scalings and L-nonlinear functions.

3.1 Consistent Scalings

We define a class of scaling vectors that will be shown to guarantee the convergence of the IPG algorithm.

Definition 1 (Consistent Scaling). *Given $\{x_k\} \subset int(\mathcal{F})$, we say that $\{d_k\} \in \Re^n_{++}$ is a consistent scaling sequence (or simply consistent), if*

$$d_k = s_k \circ h(x_k) \tag{17}$$

for a bounded sequence $\{s_k\} \subset \Re^n_{++}$ such that for every index i, a subsequence of $\{[s_k]_i\}$ converges to zero only if the corresponding subsequence of $\{[h(x_k)]_i\}$ does so.

There is a wide range of choices for consistent scalings. In particular, $\{d_k\}$ is consistent if $\{s_k\}$ is bounded both above and away from zeros for each of its components. Moreover, for any $\nu \geq 0$ and any $\{c_k\} \subset \Re^n_{++}$ that is bounded above and away from zero, the choice

$$s_k = c_k \circ [h(x_k)]^\nu, \tag{18}$$

produces a consistent scaling sequence. In particular, with $\nu = 1$ and $[c_k]_i \equiv 1$, we obtain the so-called affine scaling vector $d_k = [h(x_k)]^2$. It is easy to see that in (18) the term $[h(x)]_i$ can be replaced by $\phi([h(x)]_i)$ for any function ϕ that is continuous in $[0, 1]$ and satisfies $\phi(t) > 0$ for $t \in (0, 1]$ and $\phi(0) = 0$. Another choice for s_k is

$$s_k = 1./\left(\nabla f(x_k) + r_k\right), \tag{19}$$

where the symbol "./" denotes element-wise division, and $r_k > 0$ is chosen in such a way that ensures $\nabla f(x_k) + r_k \geq \delta > 0$, thus $\{s_k\}$ being positive and bounded. Whenever $\{x_k\}$ and $\{r_k\}$ are bounded, $\{s_k\}$ will be bounded away from zero, hence consistent. This choice of s_k has the property that it always ensures $\hat{\alpha}_k > 1$, which can be seen from a direct calculation

$$[h(x_k - d_k \circ \nabla f(x_k))]_i = [d_k \circ r_k]_i > 0, \quad \forall i \in I_1 \cup I_2,$$

indicating that the step $\alpha = 1$ does not reach the boundary of $\mathcal{F}$.

We observe that a consistent scaling sequence $\{d_k\}$ is bounded because $\{h(x_k)\}$ is bounded. Therefore, condition (13) always holds for any consistent scaling sequences.

3.2 L-nonlinear Functions

For a differentiable function $f : \Re^n \to \Re$, let $\mathcal{T}_f(\hat{x})$ be the tangent space of f at $\hat{x}$ and $\mathcal{L}_f(\hat{x})$ be the level set of f at $\hat{x}$, i.e.,

$$\mathcal{T}_f(\hat{x}) := \{p \in \Re^n : \nabla f(\hat{x})^T p = 0\}, \quad \mathcal{L}_f(\hat{x}) := \{x \in \Re^n : f(x) = f(\hat{x})\}.$$

We now introduce a concept called L-nonlinearity that is related to level sets of functions.

Definition 2 (L-nonlinearity). *Let $f : \Re^n \to \Re$ be differentiable at $\hat{x} \in \Re^n$. We call f L-nonlinear at $\hat{x}$ if there exists a neighborhood $\mathcal{B}(\hat{x}, \rho)$ of $\hat{x}$ such that*

$$\mathcal{B}(\hat{x}, \rho) \cap (\mathcal{T}_f(\hat{x}) + \hat{x}) \cap \mathcal{L}_f(\hat{x}) = \{\hat{x}\}$$

The supremum of the radii of $\mathcal{B}(\hat{x}, \rho)$ satisfying the above condition, denoted by $\rho_f(\hat{x})$, is called the L-radius *of f at $\hat{x}$. We will say that f is L-nonlinear at $\hat{x}$ with radius $\rho_f(\hat{x})$. In addition, we say that f is L-nonlinear in a set Ω if it is so at every point in Ω.*

Geometrically, f is L-nonlinear at $\hat{x}$ with radius $\rho_f(\hat{x})$ means that within the ball $\mathcal{B}(\hat{x}, \rho_f(\hat{x}))$, the affine space $\mathcal{T}_f(\hat{x}) + \hat{x}$ and the level set $\mathcal{L}_f(\hat{x})$ intersect only at $\hat{x}$. In other words, the level set has no linear (straight) segment that contains $\hat{x}$. The following observation will be useful later.

Proposition 3. *A function f is L-nonlinear at $\hat{x}$ with radius $\rho_f(\hat{x})$ if and only if*

$$\left[x \in \mathcal{B}(\hat{x}, \rho_f(\hat{x})),\ \nabla f(\hat{x})^T (x - \hat{x}) = 0,\ x \neq \hat{x}\right] \Longrightarrow f(x) \neq f(\hat{x}). \tag{20}$$

One can readily verify the following observations:

1. If f is strictly convex, then it is everywhere L-nonlinear with radius $+\infty$ (because strictly convexity implies $f(x) > f(\hat{x}) + \nabla f(\hat{x})^T (x - \hat{x}), \forall\, x \neq \hat{x}$).

   Similarly, strictly concave functions are everywhere L-nonlinear with radius $+\infty$.

2. If f is convex and locally strictly convex at $\hat{x}$, then it is L-nonlinear at $\hat{x}$ with radius $+\infty$. The analogous property holds for the concave case too.

3. If $\hat{x}$ is an extreme point of the sub-level set $\{x \in \Re^n : f(x) \leq f(\hat{x})\}$ which happen to be convex, then f is L-nonlinear at $\hat{x}$ with radius $+\infty$.

4. If f is non-convex, but locally strictly convex at $\hat{x}$, then it is L-nonlinear at $\hat{x}$. The analogous property holds for the concave case.

5. Piecewise linear functions are not L-nonlinear anywhere.

We emphasize that there are many L-nonlinear functions that are neither convex nor concave. For example, the quadratic function $f(x) = x_1 x_2$ in $\Re^2_+$ has an indefinite Hessian and hence is neither convex nor concave at any point. However, it is L-nonlinear with radius $+\infty$ everywhere in $\Re^n_{++}$. To see this, we note that at any point $\hat{x} \in \Re^n_{++}$ the tangent line to the contour $x_1 x_2 = \hat{x}_1 \hat{x}_2 > 0$ does not intersect with the contour at any point other than $\hat{x}$. As a matter of fact, it is reasonable to say that virtually all nonlinear functions are L-nonlinear unless they contain some linear pieces.

3.3 Global Convergence

We first emphasize that the only assumption made in our convergence analysis is that the objective function f is Lipschitz continuously differentiable. In particular, we will not make any iterate-dependent assumptions such as the sub-level set $\mathcal{L}_0 := \{x \in \mathcal{F} : f(x) \leq f(x_0)\}$ is bounded. The global convergence result of this paper is the following.

Theorem 1. *Let f be differentiable in $\mathcal{F}$ whose gradient is Lipschitz continuous in any compact subset of $\mathcal{F}$, and let $\{x_k\}$ be generated by the IPG algorithm corresponding to a consistent scaling sequence. Then either $f(x_k) \to -\infty$, or $\|x_{k+1} - x_k\| \to 0$ and one of the following two cases occurs:*

(1) *either $\{x_k\}$ diverges, i.e., $\|x_k\| \to \infty$;*

(2) *or any limit point of $\{x_k\}$ is a KKT point.*

Moreover, if f is L-nonlinear, then $\{x_k\}$ must converge to a KKT point in case **(2)**.

The proof of Theorem 1 will be postponed to the next section. Here we offer a few observations about the theorem.

- Theorem 1 does not require that the iterate sequence $\{x_k\}$ be bounded; nor does it require that f be bounded below. Hence the theory allows (i) the possibility for a subsequence of an unbounded sequence to converge to a KKT point, and (ii) the possibility for convergence to a KKT point of a function which is unbounded below in $\mathcal{F}$.

- The results in Theorem 1 are sharp in the sense that the divergence case can indeed happen when $I_0 \cup I_1 \neq \emptyset$. This can be seen from the simple univariant problem: $\min\ \{1/x : x \geq 1\}$. For this problem, the IPG algorithm with a consistent scaling sequence $\{d_k\}$ would generate a sequence $\{x_k\}$ that satisfies $x_{k+1} - x_k = \alpha_k d_k / x_k^2 \to 0$ and $x_k \to \infty$.

- The phenomenon that $\|x_{k+1} - x_k\| \to 0$, whenever $\{f(x_k)\}$ is bounded below, is an interesting one. It implies that any non-empty limit set of $\{x_k\}$ is either a singleton or a continuum (see Theorem 4.1(ii) in [17] for a proof, and see [15] for a similar result). As we will prove in Lemma 4 in the next section, for a L-nonlinear function the limit set of $\{x_k\}$ cannot be a continuum. Therefore, for a L-nonlinear function the sequence $\{x_k\}$ must converge to a KKT point if it does not diverge to infinity.

- If f is twice continuously differentiable in $\mathcal{F}$, then its gradient will be Lipschitz continuous in any compact subset of $\mathcal{F}$.

Of course, one could easily add some further assumptions such as the boundedness of the sub-level set $\mathcal{L}_0$, as is frequently assumed in analysis of this kind, and obtain stronger convergence statements. However, we prefer to keep our assumptions to the absolute minimum.

4 Proof of Convergence

In this section, we develop a proof for Theorem 1, which will consist of five technical lemmas. Lemma 1 proves that $\|x_{k+1} - x_k\| \to 0$ whenever $\{f(x_k)\}$ is bounded below. Lemmas 2 and 3 prove case **(2)** of Theorem 1, and Lemmas 4 and 5 prove the statement for L-nonlinear functions in Theorem 1.

We start by introducing a list of five conditions that will be selectively used by different lemmas. First we define $\mathcal{A}\{x_k\}$ to be the set of accumulation (or limit) points of a sequence $\{x_k\}$. Obviously, if the sequence is bounded, then $\mathcal{A}\{x_k\} \neq \emptyset$. Moreover, in our case $\{x_k\}$ satisfies the condition $f(x_{k+1}) < f(x_k)$ for some continuous function f, hence $\mathcal{A}\{x_k\} \equiv \mathrm{cl}\{x_k\} \setminus \{x_k\}$. Let $\mathcal{K} \subset \mathbb{N}$ be an infinite subsequence of indices.

C1 $\{x_k : k \in \mathcal{K}\}$ is bounded $\Rightarrow h(x_k) \circ \nabla f(x_k) \to 0$ for $k \in \mathcal{K}$.

C2 $\nabla f(x_k) \circ (x_{k+1} - x_k) \leq 0$.

C3 $f(x_k) \to \hat{f} > -\infty$.

C4 $\nabla f(x_k)^T(x_{k+1} - x_k) \to 0$.

C5 $|x_{k+1} - x_k| \leq \gamma |\nabla f(x_k)|$ for some $\gamma > 0$.

We first note that conditions C2, C3, C4 and C5 are the properties (11), (12), (14) and (13), respectively.

Proposition 4. *Let f and $\{x_k\}$ satisfy the conditions in Theorem 1 and $\{f(x_k)\}$ be bounded below, then conditions C2-C5 hold. Moreover, inequality (16) holds on any bounded subsequence of $\{x_k\}$ for some Lipschitz constant L.*

Condition C1 is an asymptotic complementarity condition for any bounded (not necessarily convergent) subsequences. Out of the five conditions C1-C5, only condition C1 is still missing.

Lemma 1. *Conditions C2, C4 and C5 imply that $\|x_{k+1} - x_k\| \to 0$.*

Proof. First, conditions C2 and C4 imply

$$(x_{k+1} - x_k) \circ \nabla f(x_k) \to 0. \tag{21}$$

Suppose that the lemma is not true, then for some index i there must exist a subsequence $\mathcal{K} \subset \mathbb{N}$ such that the set $\{[x_{k+1} - x_k]_i : k \in \mathcal{K}\} \subset \Re$ is bounded away from zero. By (21) we have $[\nabla f(x_k)]_i \to 0$ for $k \in \mathcal{K}$. However, condition C5 forces $[x_{k+1} - x_k]_i \to 0$ for $k \in \mathcal{K}$, which generates a contradiction. $\square$

We now prove condition C1 in the following lemma.

Lemma 2. *Let f and $\{x_k\}$ satisfy the conditions in Theorem 1 and $\{f(x_k)\}$ be bounded below. Then for any bounded subsequence $\{x_k : k \in \mathcal{K} \subset \mathbb{N}\}$, $h(x_k) \circ \nabla f(x_k) \to 0$ for $k \in \mathcal{K}$.*

Proof. Recall that $p_k = -d_k \circ \nabla f(x_k)$ and α_k satisfies (16). We observe that $\hat{\alpha}_k$ is the largest value of α that satisfies

$$[h(x_k - \alpha d_k \circ \nabla f(x_k))]_i \geq 0, \; \forall i.$$

For convenience, let us define

$$I_a(x) = \{i : i \in I_1^+(x) \;\; \text{or} \;\; i \in I_2 \text{ with } [\nabla f(x_k)]_i \neq 0\}.$$

Then, by the definitions of $h(x)$ and d_k (see (8) and (17), respectively),

$$\hat{\alpha}_k \equiv \min_{i \in I_a(x)} \frac{[h(x_k)]_i}{|d_k \circ \nabla f(x_k)|_i} = \min_{i \in I_a(x)} \frac{1}{|s_k \circ \nabla f(x_k)|_i} \geq \frac{1}{\|s_k \circ \nabla f(x_k)\|_\infty}. \tag{22}$$

Next, noting that $\{d_k\}$ is bounded, we calculate that

$$\frac{-\nabla f(x_k)^T p_k}{\|p_k\|^2} = \frac{p_k^T \mathrm{Diag}(d_k)^{-1} p_k}{p_k^T p_k} \geq \frac{1}{\|d_k\|_\infty} \geq \gamma_1 > 0. \tag{23}$$

Now it follows from (16), (22) and (23) that for $k \in \mathcal{K}$

$$\alpha_k \geq \min\left(\hat{\alpha}, \frac{\tau}{\|s_k \circ \nabla f(x_k)\|_\infty}, \frac{1-c_2}{L}\gamma_1\right) \geq \gamma_2 > 0, \tag{24}$$

since $\{s_k\}$ is bounded, and so is $\{\nabla f(x_k) : k \in \mathcal{K}\}$ given that $\{x_k : k \in \mathcal{K}\}$ is bounded.

By conditions C2 and C4 and noting $x_{k+1} - x_k = \alpha_k p_k = -\alpha_k d_k \circ \nabla f(x_k)$ for some consistent $d_k > 0$, we have

$$-\alpha_k\, p_k \circ \nabla f(x_k) = \alpha_k\, [s_k \circ \nabla f(x_k)] \circ [h(x_k) \circ \nabla f(x_k)] \rightarrow 0.$$

In particular, since $\{\alpha_k : k \in \mathcal{K}\}$ is bounded away from zero by (24), there holds

$$[s_k \circ \nabla f(x_k)] \circ [h(x_k) \circ \nabla f(x_k)] \rightarrow 0,\ k \in \mathcal{K}. \tag{25}$$

Suppose the lemma is not true. Then, for some index i, there must exist a subsequence $\mathcal{K}_1 \subset \mathcal{K}$ such that both $\{[h(x_k)]_i : k \in \mathcal{K}_1\}$ and $\{[\nabla f(x_k)]_i : k \in \mathcal{K}_1\}$ are bounded away from zero. By (25), the corresponding subsequence $\{[s_k]_i : k \in \mathcal{K}_1\}$ must converge to zero. Since $\{d_k\}$ is consistent, this means that the corresponding subsequence $\{[h(x_k)]_i : k \in \mathcal{K}_1\}$ must converge to zero, which is a contradiction. $\square$

It is worth noting that since the function $h(x)$ is not everywhere continuous, case **(2)** of Theorem 1 does not automatically follow from condition C1. We need the following lemma to prove case **(2)**.

Lemma 3. *Under condition C1, any limit point of $\{x_k\} \subset \mathcal{F}$ is a KKT point of Problem (6).*

Proof. Let a subsequence $\{x_k : k \in \mathcal{K} \subset \mathbb{N}\}$ converge to a point $\hat{x} \in \mathcal{A}\{x_k\}$. Clearly $\hat{x} \in \mathcal{F}$ since $\{x_k\} \subset \mathcal{F}$. To prove that $\hat{x}$ satisfies the KKT condition in (7), we only need to verify the complementarity

$$[h(\hat{x}) \circ \nabla f(\hat{x})]_i = 0,\ \forall i.$$

Clearly, the equation holds if $[\nabla f(\hat{x})]_i = 0$. On the other hand, if $[\nabla f(\hat{x})]_i \neq 0$, then $[h(x)]_i$ is continuous at $\hat{x}$. Consequently, by condition C1

$$[h(\hat{x})]_i = \lim_{k \in \mathcal{K}} [h(x_k)]_i = 0,$$

which proves the lemma. $\square$

The statement in Theorem 1 concerning L-nonlinear functions follows from the last two lemmas below.

Lemma 4. *Under conditions C1 and C3, if f is L-nonlinear in $\mathcal{A}\{x_k\}$, then every limit point of $\{x_k\}$ is isolated in $\mathcal{A}\{x_k\}$.*

Proof. Let $\hat{x}$ be a limit point of $\{x_k\}$. Then condition C1, together with Lemma 3, and condition C3 imply that

$$h(\hat{x}) \circ \nabla f(\hat{x}) = 0, \quad f(\hat{x}) = \hat{f}.$$

Define the index set

$$N = \{i : [\nabla f(\hat{x})]_i \neq 0\}.$$

It follows from the complementarity that $[h(\hat{x})]_N = 0$.

By continuity of $\nabla f(x)$, there exists a neighborhood $\mathcal{B}(\hat{x}, \epsilon)$ of $\hat{x}$ in which all points satisfy $[\nabla f(x)]_i \neq 0, i \in N$. This means that in $\mathcal{B}(\hat{x}, \epsilon)$ the functions $[\nabla f(x)]_i$, $i \in N$, do not change signs. In addition, we insist on choosing $\epsilon \leq \rho_f(\hat{x})$, the L-radius of the L-nonlinear function f at $\hat{x}$. The lemma will be proved if we show that there is no limit point other than $\hat{x}$ in $\mathcal{B}(\hat{x}, \epsilon)$. We will prove this fact by contradiction.

Suppose that $x' \in \mathcal{B}(\hat{x}, \epsilon) \cap \mathcal{A}\{x_k\}$ but $x' \neq \hat{x}$. We should have (i) $f(x') = \hat{f}$, (ii) $h(x') \circ \nabla f(x') = 0$, and (iii) $[h(x')]_N = [h(\hat{x})]_N = 0$ since $[\nabla f(x')]_N \neq 0$. We note that (iii) implies $[x' - \hat{x}]_N = 0$, which in turn implies $\nabla f(\hat{x})^T (x' - \hat{x}) = 0$. Since f is L-nonlinear at $\hat{x}$ with radius $\rho_f(\hat{x})$ and $\|x' - \hat{x}\| < \rho_f(\hat{x})$, by (20) we have $f(x') \neq f(\hat{x})$ which is a contradiction. $\square$

Lemma 5. *Under conditions C1-C5, if $\mathcal{A}\{x_k\} \neq \emptyset$ and f is L-nonlinear in $\mathcal{A}\{x_k\}$, then $\{x_k\}$ must converge.*

Proof. Let $\hat{x} \in \mathcal{A}\{x_k\}$ and $\mathcal{B}(\hat{x}, \epsilon)$ be such that its closure $\mathrm{cl}(\mathcal{B}(\hat{x}, \epsilon))$ contains no other limit point, which is possible for ϵ sufficiently small because $\hat{x}$ is isolated in $\mathcal{A}\{x_k\}$ by Lemma 4. We will prove that $\{x_k\}$ converges to $\hat{x}$. The proof is by contradiction. Suppose that $\{x_k\}$ does not converge to $\hat{x}$. Then there must be infinitely many points of $\{x_k\}$ inside of $\mathcal{B}(\hat{x}, \epsilon)$ and infinitely many points of $\{x_k\}$ outside of $\mathcal{B}(\hat{x}, \epsilon)$. We can select an infinite subset $\mathcal{K} \subset \mathbb{N}$ such that

$$\{x_k : k \in \mathcal{K}\} \subset \mathcal{B}(\hat{x}, \epsilon), \quad \{x_{k+1} : k \in \mathcal{K}\} \subset \Re^n \setminus \mathcal{B}(\hat{x}, \epsilon).$$

Moreover, since $\hat{x}$ is the only limit point in $\mathrm{cl}(\mathcal{B}(\hat{x}, \epsilon))$,

$$\lim_{k \in \mathcal{K}} x_k = \hat{x} \in \mathcal{B}(\hat{x}, \epsilon).$$

It is now evident

$$\liminf_{k\in\mathcal{K}} \|x_{k+1} - x_k\| = \liminf_{k\in\mathcal{K}} \|x_{k+1} - \hat{x}\| \geq \epsilon > 0,$$

which contradicts Lemma 1. $\square$

5 A Numerical Example

In this section, we present a simple numerical example to demonstrate that the choices of scaling sequences can indeed significantly influence the convergence behavior of the IPG algorithm.

We consider an instance of the nonnegative linear least squares problem

$$\min_{x\geq 0} f(x) := \frac{1}{2}\|Ax - b\|^2,$$

where the data (A, b) are generated by the Matlab scripts:

```
rand('state',0);
m = 800; n = 400; density = 0.05; condA = 1.e+2;
A = sprand(m,n,density,1/condA); b = rand(m,1);
```

and we set the random-number generator to the state of 0 so that the same data can be re-generated. The elements of the matrix A and the right-hand side vector b are uniformly distributed pseudo-random numbers between zero and one. In fact, we have $A \geq 0$ and $b > 0$.

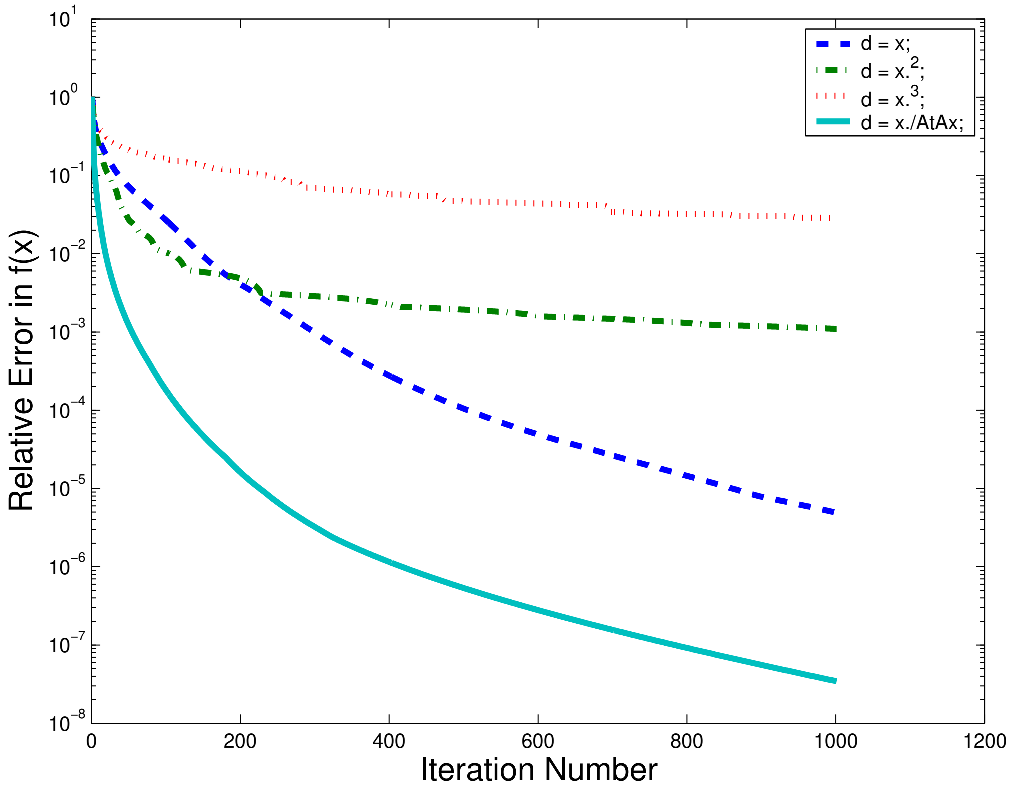
For the IPG algorithm, we use four different scalings: $d = x^p$ for $p = 1, 2, 3$, and $d = x./(A^T Ax)$. We note that $d = x^2$ is the affine scaling, and that the last scaling is a special case of (19) corresponding to $r = A^T b$. Indeed, in this case,

$$\nabla f(x) + r = (A^T Ax - A^T b) + A^T b = A^T Ax.$$

Since $A \geq 0$ and $b > 0$ in our particular case, we do have the properties that $r \equiv A^T b > 0$ and $\nabla f(x) + r \equiv A^T Ax > 0$ for $x > 0$. For this convex quadratic function, instead of doing a line search at each iteration, we always take the minimum in the search direction whenever it is inside the permissible interval of the IPG algorithm. We mention that the scaling $d = x./(A^T Ax)$ was studied in [13] as a special case of the IPG algorithm for solving nonnegative least squares problems with data $A \geq 0$ and $b > 0$.

Starting from the vector of all ones, we ran the IPG algorithm with the above four scalings and a maximum number of 1000 iterations. In Figure 1, we plot the relative

Figure 1: Relative errors $[f(x_k) - f(x_*)]/f(x_*)$ for 4 scaling sequences



errors, $[f(x_k) - f(x_*)]/f(x_*)$, against the iteration number, where x_* is the optimal solution computed by an active set method.

As can be seen from Figure 1, the four scalings have led to very different convergence behaviors on this test problem. The scaling $d = x^3$ has the slowest convergence, followed by $d = x^2$ and then $d = x$, while the scaling $d = x./(A^T Ax)$ has the fastest convergence. For this problem, the differences in the speed of convergence for these four scalings are quite significant. In fact, our numerical experiments indicate that the advantage of the last scaling tends to be more pronounced for matrices with larger condition numbers. However, the study of different scalings, as interesting as it may be, is out of the scope of the current paper.

6 Discussions

The IPG algorithm studied in this paper can be considered an extension to primal affine-scaling gradient algorithms. A closely related work in this direction is Gonzaga and Carlos [8] which studies convergence for convex programming with linear constraints. Our results also have connections to the works of Coleman and Li [3], and Dennis and Vicente [4] that study trust-region methods for simple-bound constrained optimization using scalings related to $h(x_k)$. Since some consistently scaled gradient directions are permissible approximate solutions to the trust-region subproblems in [3] or [4] when the matrix (Hessian approximation) in the quadratic model is set to zero, it is reasonable to argue that the convergence results in [3] and [4] have a certain overlap with ours. However, we point out that, unlike our results, the convergence results in both [3] and [4] assumed the boundedness of the sub-level set at the initial point (hence the boundedness of the entire iterate sequence).

In our view, Theorem 1 gives a fairly complete characterization of the behavior of the IPG algorithm for Lipschitz continuously differentiable functions. The most interesting result, though, is perhaps the strong convergence behavior proven for L-nonlinear functions. We are not aware of any similar results of this kind in the literature.

The class of consistent scalings introduced in this paper offers flexibility to the choices of diagonal scalings besides the usual affine-scaling, which may prove useful in exploiting problem structures and improving convergence behavior. This topic deserves further research.

Acknowledgments

The author would like to thank Matthias Heinkenschloss, Florian Jarre, and Richard Tapia for many useful comments on early drafts of the paper that have helped improve the paper.

References

[1] L. Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. Pacific Journal of Mathematics, 16, 1-3, 1966.

[2] T. F. Coleman and Y. Li. On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to bounds. Mathematical Programming 67, Series A (1994), 189-224.

[3] T. F. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. SIAM Journal on Optimization 6 (1996) pp. 418-445.

[4] J. E. Dennis and L. N. Vicente. Trust-Region Interior-Point Algorithms for Minimization Problems with Simple Bounds. H. Fischer (Ed.), Applied Mathematics and Parallel Computing, Festscrift for Klans Ritter, Springer, New York, 1996, pp. 97-107.

[5] J. E. Dennis, M. Heinkenschloss and L. N. Vicente. Trust-Region Interior-Point SQP Algorithms for a Class of Nonlinear Programming Problems. SIAM J. Control Optim. 36 (1998), no. 5, 1750-1794.

[6] I. I. Dikin. Iterative solution of problems of linear and quadratic programming. Sov. Math. Doklady 8, pp. 674-675, 1967.

[7] A. A. Goldstein. Convex programming in Hilbert space. Bull. Amer. Math. Soc. 70 (1964): 709–710.

[8] C. Gonzaga and L. Carlos. A primal affine-scaling algorithm for constrained convex programs. Report ES-238/90, COPPE - Federal University of Rio de Janeiro, 1990. Revised in September 2002.

[9] M. Heinkenschloss, M. Ulbrich, and S. Ulbrich. Superlinear and quadratic convergence of affine-scaling interior-point Newton methods for problems with simple bounds without strict complementarity assumption. Mathematical Programming, Vol 86 (1999), No. 3, pp. 615-635.

[10] E. S. Levitin and B. T. Polyak. Constrained minimization methods. USSR Comput. Math. Math. Phys., vol. 6, pp. 1-50, 1966.

[11] G. P. McCormick. Anti-zig-zagging by bending. Management Science, 15 (1969): 315–319.

[12] G. P. McCormick and R. A. Tapia. The gradient projection method under mild differentiability conditions. SIAM J. Control 10 (1972), 93–98.

[13] M. Merritt and Y. Zhang. An interior-point gradient method for large-scale totally nonnegative least squares problems. Technical Report TR04-08, Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005, U.S.A., 2004.

[14] J. Nocedal and S. Wright. *Numerical Optimization.* Springer, New York, 1999.

[15] A. M. Ostrowski. *Solution of Equations and Systems of Equations.* Academic Press, New York, 1966.

[16] J. B. Rosen. The gradient projection method for nonlinear programming. I. Linear constraints. J. Soc. Indust. Appl. Math. 8 (1960): 181–217.

[17] R. Tapia, Y. Zhang and Y. Ye. On the convergence of the iteration sequence in primal-dual interior-point methods. Mathematical Programming, Vol.68:141-154, 1995.

[18] P. Tseng. Convergence Properties of Dikin's Affine Scaling Algorithm for Nonconvex Quadratic Minimization. Manuscript, Department of Mathematics, University of Washington, Seattle, Washington, 2001 (revised June, 2002), to appear in J. Global Optimization.

[19] P. Wolfe. Convergence conditions for ascent methods. SIAM Review, 11, 226-235, 1969.

[20] P. Wolfe. Convergence conditions for ascent methods II: some corrections. SIAM Review, 13, 185-188, 1971.