A Scaled Projected Reduced-Gradient Method for Linearly Constrained Smooth Optimization[1]

Paul Tseng[2]

May 18, 2007

Abstract: We propose a feasible ascent method for linearly constrained smooth optimization that uses a scaled projected reduced-gradient search direction. This method is simple, reminiscent of reduced-gradient methods and first-order affine scaling methods, and can start anywhere in the feasible set. We establish global convergence and, under a Hölderian error bound assumption, sublinear convergence rate for this method. Numerical experience on simplex constrained problems with 10000 variables suggests that the method can be effective for solving large problems.

Key words. Linearly constrained optimization, reduced gradient, affine scaling, global convergence, sublinear convergence rate, error bound.

1 Introduction

Consider the linearly constrained smooth optimization problem

$$\max_{x \in X} \quad f(x), \tag{1}$$

where $f : \Re^n \to \Re$ is continuously differentiable and

$$X := \{x \in \Re^n \mid Ax = b,\ x \geq 0\},$$

with $A \in \Re^{m\times n}$, $b \in \Re^m$. This problem has been well studied and many solution methods have been proposed, including Wolfe's reduced-gradient method [1, 22], Rosen's projected-gradient method and its active set variants [1, 2, 10], the gradient projection method of Goldstein and Levitin, Polyak

[1] This research is supported by the National Science Foundation, Grant No. DMS-0511283.

[2] Department of Mathematics, University of Washington, Seattle, WA 98195, U.S.A. (tseng@math.washington.edu)

and its variants [2, 16], and interior-point methods [9, 11, 18]. Some of the methods require solving a nontrivial subproblem at each iteration while others, such as the conditional gradient method of Frank and Wolfe [2, Sec. 2.2], can suffer from slow convergence.

In the special case of homogeneous quadratic f and unit simplex constraint, i.e.,

$$f(x) = x^T Qx, \qquad A = e, \qquad b = 1,$$

with $Q \in \Re^{n\times n}$ symmetric and e a row vector of 1s, Bomze [3, Sections 2-3] recently proposed a novel feasible ascent method that is simple and has features of reduced-gradient methods and first-order affine scaling methods [4, 7, 11, 25]. In the basic version of Bomze's method, given a current feasible point x, it computes a reduced gradient

$$r(x) = Qx - e^T x^T Qx$$

and a corresponding search direction

$$d(x) = r(x)^+ - xer(x)^+,$$

where $y^+ := \max\{0, y\}$ componentwise for any $y \in \Re^n$. Then a feasible line search is performed from x along $d(x)$ to obtain a new feasible point, and this iteration is repeated. It is shown in [3] that $d(x)$ is a feasible ascent direction, and any cluster point of the sequence of x generated is a stationary point. The direction $d(x)$ is reminiscent of affine scaling direction in which a reduced gradient similar to $r(x)$ is scaled componentwise by x twice to obtain a feasible ascent direction; see Section 6.1. Unlike affine scaling, $d(x)$ is defined for all $x \in X$ and is scaled only partially by x and only once. This method has the additional novel property that it maintains $x^T r(x) = 0$ and terminates when $r(x) \leq 0$.

Motivated by Bomze's work, we propose a first-order feasible ascent method for solving the general problem (1), using a scaled projected reduced-gradient direction that generalizes $d(x)$ above. This method is simple, has features reminiscent of reduced-gradient methods and first-order affine scaling methods, and can start anywhere in X and achieve global convergence; see Theorem 1. Under a Hölderian error bound assumption, it has provably sublinear convergence rate; see Theorem 2. We report numerical experience with a Matlab implementation of the new method on simplex constrained problems with $n \in \{1000, 10000\}$. When combined with reduced-gradient projection, the resulting hybrid method shows better performance than either a reduced-gradient projection method or a first-order affine-scaling method. Comparison with MINOS (Version 5.5.1) [20] suggests that the hybrid method can be effective for solving large problems.

In what follows, I_m denotes the $m \times m$ identity matrix. For any $J \subseteq \{1, ..., n\}$, $x_J = (x_j)_{j \in J}$ and $A_J = [A_j]_{j \in J}$, where x_j denotes the jth component of x and A_j denotes the jth column of A. Also, $\|x\| = \sqrt{x^T x}$. We make one of the following two assumptions on A and b:

Assumption 1 *(a) For some $A' \in \Re^{(m-1)\times n}$,*

$$A = \begin{bmatrix} A' \\ e \end{bmatrix}, \qquad b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{2}$$

(b) $X \neq \emptyset$ and, for every $x \in X$, A_J has rank m, where $J = \{j \mid x_j > 0\}$.

Assumption 2

$$A = \begin{bmatrix} e^1 & 0 & \cdots & 0 \\ 0 & e^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & e^m \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

with $e^1, ..., e^m$ are row vectors of 1s.

Assumption 1(a) says that the problem is in Karmarkar's canonical form [12, Section 5], which can be made without significant loss of generality. Assumption 1(b) says that A has linearly independent rows and the problem is primal nondegenerate. This assumption is often made for affine scaling methods [4, 7, 5, 11, 18, 25, 28]. Assumption 2 says that the feasible set X is a Cartesian product of unit simplices. Under this assumption, Assumption 1(b) is automatically satisfied. While this special case of (1) can be transformed into Karmarkar's canonical form, the resulting method seems to be different. Notice that X is nonempty and compact under either assumption.

2 Feasible ascent directions

A key to the new method is a mapping $E : \Re^n \to \Re^{n\times m}$ that satisfies

$$AE(x) = I_m \qquad \text{and} \qquad [x]^\theta \in \text{span}E(x) \qquad \forall x \in X, \tag{3}$$

for some $\theta \geq 1$, where $[x]^\theta$ denotes x raised to the power θ componentwise. We define, for each $x \in X$,

$$\begin{aligned} \lambda(x) &:= E(x)^T \nabla f(x), & (4)\\ r(x) &:= \nabla f(x) - A^T\lambda(x), & (5)\\ d(x) &:= r(x)^+ - E(x)Ar(x)^+. & (6) \end{aligned}$$

(As in [3], the scalar function $t \mapsto \max\{0, t\}$ can more generally be replaced by any continuous function $\varphi : \Re \to \Re$ satisfying $\varphi(t) = 0$ for $t \leq 0$ and $\varphi(t) > 0$ for $t > 0$. Theorem 1 still holds with this variant.) The following lemma gives key properties of $r(x)$ and $d(x)$.

Lemma 1 *Let $E : \Re^n \to \Re^{n\times m}$ satisfy* (3)*. Let r and d be given by* (4)–(6)*. For any $x \in X$, we have $([x]^\theta)^T r(x) = 0$, $Ad(x) = 0$, and $\nabla f(x)^T d(x) = \|r(x)^+\|^2$.*

Proof. Using (5), (3) and (4), we have

$$\begin{aligned} E(x)^T r(x) &= E(x)^T \nabla f(x) - (AE(x))^T \lambda(x)\\ &= E(x)^T \nabla f(x) - \lambda(x)\\ &= 0. \end{aligned}$$

In fact, $\lambda(x)$ is defined so that this property holds. Since $[x]^\theta \in \text{span}E(x)$, this implies $([x]^\theta)^T r(x) = 0$.

We have from (3) that

$$Ad(x) = Ar(x)^+ - AE(x)Ar(x)^+ = 0.$$

Hence

$$\begin{aligned} \nabla f(x)^T d(x) &= \nabla f(x)^T d(x) - \lambda(x)^T Ad(x)\\ &= (\nabla f(x) - A^T\lambda(x))^T d(x)\\ &= r(x)^T d(x)\\ &= r(x)^T r(x)^+ - r(x)^T E(x)Ar(x)^+\\ &= r(x)^T r(x)^+\\ &= \|r(x)^+\|^2, \end{aligned}$$

where the fifth equality uses the property $E(x)^T r(x) = 0$. ■

Below we present three specific choices of $E(x)$ under Assumption 1 or 2, for which the corresponding $d(x)$ is a feasible ascent direction at $x \in X$ whenever $r(x) \not\leq 0$. Otherwise $r(x) \leq 0$ and $d(x) = 0$. Since we always have $x^T r(x) = 0$, the latter implies that x is a stationary point of (1).

2.1 A reduced-gradient-like direction

Under Assumption 1(b), there exists a scalar $\delta > 0$ such that, for any $x \in X$, A_J has rank m, where $J = \{j \mid x_j \geq \delta\}$. We choose any $m \times m$ submatrix B of A_J (obtained by dropping linearly dependent columns of A_J) and without loss of generality, we assume that B comprises the first m columns of A. In practice, such B can be found (without knowing δ) by ordering the components of x in decreasing order and checking the corresponding columns of A for linear independence (within numerical tolerance) until we obtain m linearly independent columns. Let

$$E' = \begin{bmatrix} B^{-1} \left[\begin{smallmatrix} I_{m-1} \\ 0 \end{smallmatrix}\right] \\ 0 \end{bmatrix}, \qquad E(x) = [\, E' \quad x \,]. \tag{7}$$

Here B and $E' \in \Re^{n \times (m-1)}$ both depend on x through J. By construction, $AE' = \left[\begin{smallmatrix} I_{m-1} \\ 0 \end{smallmatrix}\right]$. Under Assumption 1(a), we also have $Ax = b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, so together they imply $AE(x) = I_m$. Moreover, $x \in \text{span}E(x)$. Thus, this choice of E satisfies (3) with $\theta = 1$.

In addition,

$$E(x)Ar(x)^+ = [\, E' \quad x \,] \begin{bmatrix} A' \\ e \end{bmatrix} r(x)^+ = \begin{bmatrix} p(x) \\ 0 \end{bmatrix} + x \; er(x)^+,$$

with $p(x) = B^{-1} \begin{bmatrix} A' \\ 0 \end{bmatrix} r(x)^+ \in \Re^m$. Also, the first m components of x exceed $\delta > 0$. Since $x \geq 0$ and $er(x)^+ \geq 0$, this shows that each of the last $n - m$ components of $d(x)$ given by (6) is nonnegative whenever the corresponding component of x is zero. Thus $d(x)$ is a feasible direction at x. In particular,

$$d(x)_j < 0 \qquad \Longrightarrow \qquad x_j \geq \delta \;\text{ or }\; x_j > \frac{r(x)_j^+}{er(x)^+}. \tag{8}$$

This choice of $d(x)$ is reminiscent of reduced-gradient methods [1, 22] except for the partial scaling by x.

2.2 An affine-scaling-like direction

Under Assumption 1(b), the matrix $A\mathrm{diag}(x)^\theta A^T$ is invertible for all $x \in X$, where $\theta \geq 1$. Let

$$E(x) = \mathrm{diag}(x)^\theta A^T (A\mathrm{diag}(x)^\theta A^T)^{-1}. \tag{9}$$

Then $AE(x) = I_m$. Under Assumption 1(a), we also have

$$E(x)(A\mathrm{diag}(x)^\theta A^T)\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \mathrm{diag}(x)^\theta A^T \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \mathrm{diag}(x)^\theta e^T = [x]^\theta,$$

so that $[x]^\theta \in \mathrm{span} E(x)$. Thus, this choice of E satisfies (3). In fact, under Assumption 1, $x \mapsto (A\mathrm{diag}(x)^\theta A^T)^{-1}$ is continuous on X [7, 11], and hence so are $E(\cdot)$, $\lambda(\cdot)$, $r(\cdot)$, and $d(\cdot)$.

In addition,

$$d(x) = r(x)^+ - \mathrm{diag}(x)^\theta A^T (A\mathrm{diag}(x)^\theta A^T)^{-1} A r(x)^+,$$

so $d(x)_j \geq 0$ whenever $x_j = 0$. In particular,

$$d(x)_j < 0 \qquad \Longrightarrow \qquad u(x)_j > 0 \quad \text{and} \quad x_j \geq \frac{-d(x)_j}{u(x)_j}, \tag{10}$$

where $u(x) := \mathrm{diag}(x)^{\theta-1} A^T (A\mathrm{diag}(x)^\theta A^T)^{-1} A r(x)^+$. This choice of $d(x)$ is reminiscent of first-order affine-scaling methods, especially if we take $\theta = 2$; see Section 6.1. Unlike affine-scaling methods, $d(x)$ is defined even if $x \not> 0$. This is advantageous for warm start and parameteric optimization. If $x > 0$, then $\mathrm{diag}(x)^{-\theta/2} d(x)$ equals the orthogonal projection of $\mathrm{diag}(x)^{-\theta/2} r(x)^+$ on to the null space of $A\mathrm{diag}(x)^{\theta/2}$.

2.3 Another affine-scaling-like direction

Under Assumption 2, we choose

$$E(x) = \begin{bmatrix} x_{J_1} & 0 & \cdots & 0 \\ 0 & x_{J_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & x_{J_m} \end{bmatrix}, \tag{11}$$

where $J_1, J_2, ..., J_m$ is a partition of $\{1, 2, ..., n\}$ corresponding to $e^1, e^2, ..., e^m$. This choice of E satisfies (3) with $\theta = 1$ (since $E(x)e^T = x$).

In addition,

$$E(x)Ar(x)^+ = \begin{bmatrix} x_{J_1} \; e^1 r(x)^+_{J_1} \\ x_{J_2} \; e^2 r(x)^+_{J_2} \\ \vdots \\ x_{J_m} \; e^m r(x)^+_{J_m} \end{bmatrix}.$$

Since $x_{J_i} \geq 0$ and $e^i r(x)^+_{J_i} \geq 0$ for $i = 1, ..., m$, this shows that $d(x)$ given by (6) has a nonnegative component whenever the corresponding component of x is zero. In particular,

$$d(x)_j < 0, \; j \in J_i \qquad \Longrightarrow \qquad x_j > \frac{r(x)^+_j}{e^i r(x)^+_{J_i}}. \tag{12}$$

In the special case of $m = 1$, Assumptions 1 and 2 are equivalent, and (9) with $\theta = 1$ is equivalent to (7) and (11).

3 A scaled projected reduced-gradient method

We now describe our method. Starting with any $x^0 \in X$, we iteratively generate x^{k+1} from $x^k \in X$ $(k = 0, 1, ...)$ by

$$x^{k+1} = x^k + \alpha^k d^k \qquad \text{with} \qquad d^k = d(x^k), \tag{13}$$

where $d(\cdot)$ is given by (4), (5), (6), and the stepsize $\alpha^k \geq 0$ is chosen to ensure that $x^{k+1} \in X$ and $f(x^{k+1}) \geq f(x^k)$. We call this a *scaled projected reduced-gradient* (SPRG) method since $d(x)$ is obtained by scaling $r(x)^+$ with $I - E(x)A$. By Lemma 1, the SPRG method maintains $(\left[x^k\right]^\theta)^T r(x^k) = 0$ and $\nabla f(x^k)^T d^k < 0$ whenever $r(x^k) \not\leq 0$.

For each k, since $x^k \in X$ and $Ad^k = 0$ (by Lemma 1), $x^{k+1} \in X$ if and only if $\alpha^k \leq \bar{\alpha}^k$, where

$$\bar{\alpha}^k := \min_{j:d^k_j<0} \frac{x^k_j}{-d^k_j}. \tag{14}$$

We can choose α^k by the maximization rule:

$$\alpha^k \in \operatorname*{argmax}_{0\leq\alpha\leq\bar{\alpha}^k} f(x^k + \alpha d^k).$$

However, this is practical only if f is quadratic or cubic. For general f, we choose α^k by an Armijo rule [2, Section 2.2.1]: α^k is the largest $\alpha \in \{s\bar{\alpha}^k(\beta)^t\}_{t=0,1,\ldots}$ satisfying

$$f(x^k + \alpha d^k) \geq f(x^k) + \sigma\alpha\nabla f(x^k)^T d^k, \tag{15}$$

where $0 < s \leq 1$, $0 < \beta, \sigma < 1$ are constants. If $d^k = 0$, then (15) is satisfied by any $\alpha \geq 0$ and the Armijo rule yields $\alpha^k = s\bar{\alpha}^k$. In general, if d^k is a feasible ascent direction at x^k, then $\bar{\alpha}^k > 0$ and α^k is defined and positive.

4 Global convergence

The following theorem establishes the global convergence of the SPRG method. Its proof uses Lemma 1, the feasible ascent properties of the search directions of Section 3, and properties of the Armijo rule [2, pages 43-44]. This theorem will also be used for the convergence rate analysis in Section 5.

Theorem 1 *Let $\{x^k\}$ be generated by the SPRG method* (4)–(6), (13), *with α^k chosen by the Armijo rule* (14), (15) *and E chosen by either* (7) *or* (9) *under Assumption 1 or* (11) *under Assumption 2. Then $x^k \in X$ for all k, $\{f(x^k)\} \uparrow$, $\{d^k\}$ is bounded, and every cluster point of $\{x^k\}$ is a stationary point of* (1). *If ∇f is Lipschitz continuous on X, then* $\inf_k \alpha^k > 0$.

Proof. Since $x^0 \in X$, an induction argument using $Ad^k = 0$ and $\alpha^k \leq \bar{\alpha}^k$ yields that $x^k \in X$ for all k. Since X is bounded, this implies $\{x^k\}$, $\{E(x^k)\}$, and $\{r(x^k)\}$ are bounded, regardless of whether E is given by (7) or (9) or (11). Thus $d^k = d(x^k)$ given by (6) is bounded. Also, $\{p^k\}$ is bounded, where $p^k := r(x^k)^+$. Since $\alpha = \alpha^k$ satisfies (15), (13) and Lemma 1 yield

$$f(x^{k+1}) - f(x^k) \geq \sigma\alpha^k\nabla f(x^k)^T d^k = \sigma\alpha^k\|p^k\|^2 \quad \forall k. \tag{16}$$

Thus $\{f(x^k)\} \uparrow$. Since f is continuous and $\{x^k\}$ is bounded, this implies $\{f(x^k)\}$ converges and so $\{f(x^{k+1}) - f(x^k)\} \to 0$. Then (16) implies $\{\alpha^k\|p^k\|^2\} \to 0$.

We claim that $\bar{\alpha}^k$ given by (14) satisfies

$$\inf_k \bar{\alpha}^k > 0. \tag{17}$$

If E is given by (7) under Assumption 1, then for each k with $d_j^k < 0$, either $x_j^k \geq \delta$ or $d_j^k = p_j^k - x_j^k e p^k$ (see (8)), so that

$$\text{either} \quad \bar{\alpha}^k \geq \frac{\delta}{\max_j |d_j^k|} \quad \text{or} \quad \bar{\alpha}^k \geq \frac{x_j^k}{x_j^k e p^k - p_j^k} \geq \frac{1}{e p^k}$$

for some j with $x_j^k e p^k > p_j^k \geq 0$. Since $\{d^k\}$ and $\{p^k\}$ are bounded, this implies (17). If E is given by (9) under Assumption 1, then (10) yields

$$\bar{\alpha}^k \geq \min_{j : u(x^k)_j > 0} \frac{1}{u(x^k)_j}.$$

Since $u(\cdot)$ is continuous on X and $\{x^k\}$ is bounded, this implies (17). If E is given by (11), then for each k we have (see (12))

$$\bar{\alpha}^k = \frac{x_j^k}{x_j^k \ e^i p_{J_i}^k - p_j^k} \geq \frac{1}{e^i p_{J_i}^k},$$

for some i and $j \in J_i$ with $x_j^k \ e^i p_{J_i}^k > p_j^k \geq 0$. Since $\{p^k\}$ is bounded, this again implies (17).

Let $\bar{x}$ be any cluster point of $\{x^k\}$. Consider any subsequence $\{x^k\}_{k \in K}$ ($K \subseteq \{0, 1, ...\}$) converging to $\bar{x}$. By further passing to a subsequence if necessary, we will assume that either (i) $\inf_{k \in K} \alpha^k > 0$ or (ii) $\{\alpha^k\}_{k \in K} \to 0$.

In case (i), since $\{\alpha^k \|p^k\|^2\} \to 0$, we have $\{\|p^k\|^2\}_{k \in K} \to 0$. If E is given by (7), then since E' is from a finite set, we can assume by further passing to a subsequence that E' is constant for all $k \in K$. Then $\{E(x^k)\}_{k \in K} \to [E' \ \ \bar{x}]$, which together with the continuity of ∇f implies that

$$\{r(x^k)\}_{k \in K} \to \bar{r} := \nabla f(\bar{x}) - A^T \bar{\lambda},$$

with $\bar{\lambda} := [E' \ \ \bar{x}]^T \nabla f(\bar{x})$. Moreover, $([x^k]^\theta)^T r(x^k) = 0$ for all k (see Lemma 1) and $\{\|p^k\|^2\}_{k \in K} \to 0$ imply, respectively, $([\bar{x}]^\theta)^T \bar{r} = 0$ and $\bar{r} \leq 0$. Since $\bar{x} \geq 0$, this implies $\bar{x}^T \bar{r} = 0$ so $\bar{x}$ is a stationary point of (1). If E is given by (9) or (11), then $\{E(x^k)\}_{k \in K} \to E(\bar{x})$, and a similar argument shows $\bar{x}$ to be a stationary point of (1).

In case (ii), we have from (17) that $\alpha^k < s\bar{\alpha}^k$ for all $k \in K$ sufficiently large, implying that the ascent condition (15) is violated by $\alpha = \alpha^k/\beta$, i.e.,

$$\frac{f(x^k + (\alpha^k/\beta) d^k) - f(x^k)}{\alpha^k/\beta} < \sigma \nabla f(x^k)^T d^k. \tag{18}$$

Since $\{d^k\}$ is bounded, by further passing to a subsequence if necessary, we can assume that $\{d^k\}_{k\in K} \to$ some $\bar{d}$. Since $\{\alpha^k\}_{k\in K} \to 0$ and f is continuously differentiable, the above inequality yields in the limit that

$$\nabla f(\bar{x})^T \bar{d} \leq \sigma \nabla f(\bar{x})^T \bar{d}.$$

Since $0 < \sigma < 1$, this implies $\nabla f(\bar{x})^T \bar{d} \leq 0$. Thus $\lim_{k\in K, k\to\infty} \nabla f(x^k)^T d^k \leq 0$. By Lemma 1, this implies $\{\|p^k\|^2\}_{k\in K} \to 0$, so the same argument as in case (i) yields that $\bar{x}$ is a stationary point of (1).

Suppose that ∇f is Lipschitz continuous on X with Lipschitz constant $L \geq 0$. Then the mean value theorem yields

$$f(y) - f(x) \geq \nabla f(x)^T (y - x) - \frac{L}{2}\|y - x\|^2 \qquad \forall x, y \in X$$

(see [2, page 667]). For each $k \in \{0, 1, ...\}$, either $\alpha^k = s\bar{\alpha}^k$ or else (15) is violated by $\alpha = \alpha^k/\beta$, i.e., (18) holds. In the second case, applying the above inequality with $x = x^k, y = x^k + (\alpha^k/\beta)d^k$ and using (18) yields

$$\frac{\alpha^k}{\beta}\nabla f(x^k)^T d^k - \frac{L}{2}\left(\frac{\alpha^k}{\beta}\right)^2 \|d^k\|^2 \leq f\left(x^k + \frac{\alpha^k}{\beta}d^k\right) - f(x^k) < \frac{\alpha^k}{\beta}\sigma\nabla f(x^k)^T d^k.$$

Dividing both sides by α^k/β and rearranging terms yields

$$(1 - \sigma)\nabla f(x^k)^T d^k \leq \frac{L}{2}\frac{\alpha^k}{\beta}\|d^k\|^2.$$

Since $\{E(x^k)\}$ is bounded, (6) implies $\|d^k\| \leq C\|p^k\|$ for some constant $C > 0$. Also, $\nabla f(x^k)^T d^k = \|p^k\|^2$ by Lemma 1, so the above inequality further implies

$$(1 - \sigma)\|p^k\|^2 \leq \frac{L}{2}\frac{\alpha^k}{\beta}C^2\|p^k\|^2.$$

Since $\alpha^k \neq s\bar{\alpha}^k$, we have $d^k \neq 0$ and hence $p^k \neq 0$, so this yields $1 - \sigma \leq \frac{L}{2}\frac{\alpha^k}{\beta}C^2$. Thus in both cases we have

$$\alpha^k \geq \min\left\{s\bar{\alpha}^k, \frac{2\beta(1-\sigma)}{LC^2}\right\}.$$

This and (17) show that $\inf_k \alpha^k > 0$. $\blacksquare$

5 Sublinear convergence rate

In this section we analyze the asymptotic convergence rate of the SPRG method of Section 3. For our analysis, we define the projection residual

$$R(x) := \mathrm{P}_X[x + \nabla f(x)] - x,$$

where $\mathrm{P}_X[y] := \mathrm{argmin}_{x \in X} \|x - y\|$, and we make the following assumption similar to Assumptions A and B in [15]; also see [13].

Assumption 3 **(a)** *The stationary point set* $\bar{X} := \{x \in X \mid R(x) = 0\}$ *is nonempty.*

(b) *There exist scalars* $\tau > 0$, $\epsilon > 0$, *and* $\gamma \in (0, 1]$ *such that*

$$\mathrm{dist}(x, \bar{X}) \le \tau \|R(x)\|^{\gamma} \quad \textit{whenever} \quad x \in X, \ \|R(x)\| \le \epsilon, \tag{19}$$

where $\mathrm{dist}(x, \bar{X}) := \min_{\bar{x} \in \bar{X}} \|x - \bar{x}\|$.

(c) *There exists a scalar* $\delta > 0$ *such that*

$$\|x - y\| \ge \delta \quad \textit{whenever} \quad x \in \bar{X}, \ y \in \bar{X}, \ f(x) \ne f(y).$$

Assumption 3(a) is redundant when X is bounded. Assumption 3(b) is a local Hölderian error bound assumption, saying that the distance from x to $\bar{X}$ locally grows at most like $\|r(x)^{+}\|^{\gamma}$. Error bounds of this kind have been extensively studied and hold under very weak assumptions [8, 13, 14, 15, 21]. For example, it holds with $\gamma = 1$ when f is quadratic or $f(x) = -h(Hx) + c^T x$ for all $x \in \Re^n$, where $H \in \Re^{\ell \times n}$, $c \in \Re^n$, and h is a strongly convex differentiable function on $\Re^{\ell}$ with ∇h Lipschitz continuous on $\Re^{\ell}$. It holds with some $\gamma \in (0, 1]$ if f is analytic (e.g., polynomial). Assumption 3(c) says that the isocost surfaces of f restricted to the solution set $\bar{X}$ are "properly separated." Assumption 3(c) holds automatically if f is quadratic or concave [14, Lemma 3.1]. Thus Assumption 3 holds for any quadratic program.

The form of $r(x)$ (see (5)) hints at a connection between $r(x)^{+}$ and $R(x)$, which we will exploit. However, the search direction $d(x)$ used by the SPRG method also has features of first-order affine scaling direction due to the scaling by x This complicates the analysis. In fact, previous convergence rate analyses are only for second-order affine scaling methods and only when

f is quadratic [23, 26] or f is convex/concave with $\nabla^2 f$ having constant null space on X [18, Lemma 4.11], [24].

We begin with a key lemma which relates $\|r(x)^+\|$ to $\|R(x)\|$. In what follows, $y \perp z$ means $y^T z = 0$ for any $y, z \in \Re^n$, and, for each $\bar{x} \in \bar{X}$,

$$\Lambda(\bar{x}) := \{\bar{\lambda} \in \Re^m \mid 0 \geq \nabla f(\bar{x}) - A^T\bar{\lambda} \perp \bar{x}\}$$

denotes its set of Lagrange multiplier vectors.

Lemma 2 *Let $E : \Re^n \to \Re^{n\times m}$ satisfy* (3) *with $\theta = 1$. Let r be given by* (4) *and* (5). *Assume X is bounded. Then the following results hold.*

(a) $\nabla f(x)^T(\bar{x} - x) \leq \bar{x}^T r(x)^+$ *for all $x, \bar{x} \in X$.*

(b) *There exists a scalar $\kappa_1 > 0$ such that $\|R(x)\| \leq \kappa_1\sqrt{\|r(x)^+\|}$ for all $x \in X$.*

(c) *Suppose that every $\bar{x} \in \bar{X}$ satisfies strict complementarity in the sense that*

$$\bar{x} - (\nabla f(\bar{x}) - A^T\bar{\lambda}) > 0 \quad \forall\bar{\lambda} \in \Lambda(\bar{x}). \tag{20}$$

Then there exist scalars $\kappa_2 > 0$ and $\bar{\epsilon} > 0$ such that $\|R(x)\| \leq \kappa_2\|r(x)^+\|$ for all $x \in X$ with $\|r(x)^+\| \leq \bar{\epsilon}$.

Proof. (a). For any $x, \bar{x} \in X$, we have from (5) and $A(\bar{x} - x) = 0$ that

$$\nabla f(x)^T(\bar{x} - x) = r(x)^T(\bar{x} - x) = r(x)^T\bar{x} \leq (r(x)^+)^T\bar{x},$$

where the second equality uses Lemma 1 and the inequality uses $r(x) \leq r(x)^+$, $\bar{x} \geq 0$.

(b). In what follows, we denote $y^- = \max\{0, -y\}$. Fix any $x \in X$. For simplicity, let $g = \nabla f(x)$ and abbreviate $\lambda(x)$, $r(x)^+$, $r(x)^-$ as λ, r^+, r^-, respectively. Then (5) implies

$$g - A^T\lambda = r^+ - r^-, \qquad 0 \leq r^+ \perp r^- \geq 0.$$

Let $\tilde{J} := \{j \in \{1, ..., n\} \mid x_j < r_j^-\}$. Let $\tilde{x}, \tilde{g} \in \Re^n$ be given by

$$\tilde{x}_j := \begin{cases} 0 & \text{if } j \in \tilde{J}; \\ x_j + r_j^+ & \text{else,} \end{cases} \qquad \tilde{g}_j := \begin{cases} g_j & \text{if } j \in \tilde{J}; \\ g_j + r_j^- & \text{else.} \end{cases} \tag{21}$$

Thus if $j \in \tilde{J}$, we have $r_j^+ = 0$ and hence $\tilde{x}_j - x_j - \tilde{g}_j + A_j^T\lambda = -x_j - g_j + A_j^T\lambda = -x_j + r_j^- > 0$, as well as $\tilde{x}_j = 0$; otherwise $\tilde{x}_j - x_j - \tilde{g}_j + A_j^T\lambda = r_j^+ - g_j - r_j^- + A_j^T\lambda = 0$, as well as $\tilde{x}_j = x_j + r_j^+ \geq 0$. Thus

$$0 \leq \tilde{x} \perp \tilde{x} - x - \tilde{g} + A^T\lambda \geq 0. \tag{22}$$

Let $z = \mathrm{P}_X[x+g]$. We have that z is the optimal solution of

$$\min_y \quad \frac{1}{2}\|y - x - g\|^2 \quad \text{s.t.} \quad Ay = b, \ y \geq 0.$$

This is a convex quadratic program, so z satisfies the necessary and sufficient optimality condition

$$0 \leq z \perp z - x - g + A^T\mu \geq 0, \quad Az = Ax,$$

for some $\mu \in \Re^m$. Comparing this with (22), we see that $\tilde{x}, \tilde{g}$ defined by (21) is the optimal solution of

$$\min_y \quad \frac{1}{2}\|y - x - \tilde{g}\|^2 \quad \text{s.t.} \quad Ay = A\tilde{x}, \ y \geq 0.$$

Notice that the objective function in the above two quadratic programs have the same positive definite Hessian (the identity matrix) and differ only in their linear terms. The constraints differ only in the right-hand side. Then it is known (see [6, page 696, Exercise 7.6.10]) that

$$\|\tilde{x} - z\| = O(\|\tilde{x} - x\| + \|\tilde{g} - g\|).$$

Since $\|\tilde{x} - z\| \geq \|x - z\| - \|\tilde{x} - x\|$, this yields

$$\|R(x)\| = \|x - z\| = O(\|\tilde{x} - x\| + \|\tilde{g} - g\|). \tag{23}$$

Since $x^Tr^+ = x^Tr^- \geq x_jr_j^-$ for all j, the definition of $\tilde{J}$ implies that

$$x^Tr^+ > (x_j)^2 \quad \forall\, j \in \tilde{J}, \qquad x^Tr^+ > (r_j^-)^2 \quad \forall\, j \notin \tilde{J}.$$

Hence $\|\tilde{g} - g\| = \|(r_j^-)_{j \notin \tilde{J}}\| = O(\sqrt{x^Tr^+})$ and $\|\tilde{x} - x\| \leq \|x_{\tilde{J}}\| + \|(r_j^+)_{j \notin \tilde{J}}\| = O(\sqrt{x^Tr^+} + \|r^+\|)$. Since X is compact by assumption, $\|x\| = O(1)$ for all $x \in X$ and hence

$$\|\tilde{g} - g\| = O(\sqrt{\|r^+\|}), \qquad \|\tilde{x} - x\| = O(\sqrt{\|r^+\|}).$$

This together with (23) proves (b).

(c). We claim that there exist scalar constants $\bar{\epsilon} > 0$ and $\rho > 0$ such that

$$x + r(x)^- \geq \rho e^T \quad \text{whenever } x \in X,\ \|r(x)^+\| \leq \bar{\epsilon}. \tag{24}$$

If this were false, then there would exist a sequence $x^k \in X$, $k = 1, 2, ...,$ such that $\{r(x^k)^+\} \to 0$ and $x^k_{\bar{j}} + r(x^k)^-_{\bar{j}} \to 0$ for some $\bar{j} \in \{1, ..., n\}$. This would imply $\{x^k_{\bar{j}}\} \to 0$ and $\{r(x^k)^-_{\bar{j}}\} \to 0$. Since x^k lies in the compact set X, by passing to a subsequence if necessary we can assume that $\{x^k\}$ converges to some $\bar{x} \in X$. Since, by Lemma 1, $r(x^k) = \nabla f(x^k) - A^T\lambda(x^k) \perp x^k$ for all k and $\lambda(x^k) = E(x^k)^T \nabla f(x^k)$ is bounded, we would have in the limit as $k \to \infty$ that

$$0 \geq \nabla f(\bar{x}) - A^T\bar{\lambda} \perp \bar{x}$$

for some $\bar{\lambda} \in \Re^m$, implying $\bar{x} \in \bar{X}$ and $\bar{\lambda} \in \Lambda(\bar{x})$. Moreover, $\bar{x}_{\bar{j}} = 0$ and $\nabla f(\bar{x})_{\bar{j}} - A^T_{\bar{j}}\bar{\lambda} = 0$, which would contradict (20).

Fix any $x \in X$ with $\|r^+\| \leq \bar{\epsilon}$, where for simplicity we abbreviate $r(x)^+$, $r(x)^-$ as r^+, r^-, respectively. Let $\tilde{J} := \{j \in \{1, ..., n\} \mid x_j < r^-_j\}$. By (24),

$$r^-_j \geq \rho/2 \quad \forall j \in \tilde{J} \quad \text{and} \quad x_j \geq \rho/2 \quad \forall j \notin \tilde{J}.$$

Since $x^T r^+ = x^T r^- \geq x_j r^-_j$, this implies that

$$x_j \leq \frac{2}{\rho} x^T r^+ \quad \forall j \in \tilde{J} \quad \text{and} \quad r^-_j \leq \frac{2}{\rho} x^T r^+ \quad \forall j \notin \tilde{J}.$$

Hence, for $\tilde{x}, \tilde{g}$ defined by (21), $\|\tilde{g} - g\| = \|(r^-_j)_{j \notin \tilde{J}}\| = O(x^T r^+)$ and $\|\tilde{x} - x\| \leq \|x_{\tilde{J}}\| + \|(r^+_j)_{j \notin \tilde{J}}\| = O(x^T r^+ + \|r^+\|)$. Since $\|x\| = O(1)$ for all $x \in X$, this yields

$$\|\tilde{g} - g\| = O(\|r^+\|), \qquad \|\tilde{x} - x\| = O(\|r^+\|).$$

This together with (23) proves (c). ■

The strict complementarity assumption (20) is also used in the analyses of affine scaling methods [4, 5, 28]. Note that, under Assumption 1 or 2, $\Lambda(\bar{x})$ is a singleton for all $\bar{x} \in \bar{X}$. Below is our main convergence rate result. Its proof uses Theorem 1 and Lemma 2, as well as ideas from [15, Appendix] for the linear convergence rate analysis of gradient-projection-like methods.

Theorem 2 *Suppose that ∇f is Lipschitz continuous on X and Assumption 3 holds. Let $\{x^k\}$ be generated by the SPRG method* (4)–(6), (13), *with α^k chosen by the Armijo rule* (14), (15) *and E chosen by either* (7) *or* (9) *(with $\theta = 1$) under Assumption 1 or* (11) *under Assumption 2. Then there exist scalar $C > 0$ and integer $\bar{k}$ such that*

$$0 \le e^k \le \frac{C}{k^\omega} \tag{25}$$

for all $k \ge \bar{k}$, where $\omega = \frac{\bar{\gamma}}{1-\bar{\gamma}}$, $\bar{\gamma} = \min\{\gamma, 1\}/2$, $e^k = \bar{v} - f(x^k)$, and $\bar{v} = \lim_{k\to\infty} f(x^k)$. If every $\bar{x} \in \bar{X}$ satisfies (20), *then we can instead take $\bar{\gamma} = \min\{\gamma, 1/2\}$.*

Proof. By Theorem 1, $\{f(x^k)\} \uparrow$ and $x^k \in X$, $\alpha^k \ge \underline{\alpha}$ for all k, where $\underline{\alpha} > 0$. By the Armijo rule and Lemma 1 (see (16)),

$$f(x^{k+1}) - f(x^k) \ge \sigma\alpha^k \nabla f(x^k)^T d^k = \sigma\alpha^k \|p^k\|^2 \ge \sigma\underline{\alpha}\|p^k\|^2 \quad \forall k, \tag{26}$$

where $p^k := r(x^k)^+$. Also, X is nonempty and bounded by Assumption 1 or 2.

Since $\{f(x^k)\}$ converges, we have $\{f(x^{k+1}) - f(x^k)\} \to 0$ and hence (26) yields $\{p^k\} \to 0$. By Lemma 2(b), $\|R(x^k)\| \le \kappa_1\sqrt{\|p^k\|}$ for all k, where $\kappa_1 > 0$, and hence $\{R(x^k)\} \to 0$. By Assumption 3(b), there exist an index $\hat{k}$ and a scalar $\tau > 0$ such that

$$\|x^k - \bar{x}^k\| \le \tau\|R(x^k)\|^\gamma \quad \forall k \ge \hat{k},$$

for some $\bar{x}^k \in \bar{X}$. Hence

$$\|x^k - \bar{x}^k\| \le \tau\kappa_1^\gamma \|p^k\|^{\gamma/2} \quad \forall k \ge \hat{k}. \tag{27}$$

Combining (27) with $\{p^k\} \to 0$ gives

$$x^k - \bar{x}^k \to 0, \tag{28}$$

Then, Assumption 3(c) implies that $\bar{x}^k$ eventually settles down at some iso-cost surface of f, i.e., there exist an index $\bar{k} \ge \hat{k}$ and a scalar $\bar{v}$ such that

$$f(\bar{x}^k) = \bar{v} \quad \forall k \ge \bar{k}. \tag{29}$$

Fix any index $k \geq \bar{k}$. Since $x^k \in X$ and $\bar{x}^k$ is a stationary point of f over X, we have $\nabla f(\bar{x}^k)^T(x^k - \bar{x}^k) \leq 0$ and from the Mean Value Theorem that $f(\bar{x}^k) - f(x^k) = \nabla f(\psi^k)^T(\bar{x}^k - x^k)$, for some $\psi^k \in \Re^n$ lying on the line segment joining $\bar{x}^k$ with x^k. Upon summing these two relations and using (29), we obtain

$$\begin{array}{rcl} \bar{v} - f(x^k) & \geq & (\nabla f(\psi^k) - \nabla f(\bar{x}^k))^T(\bar{x}^k - x^k) \\ & \geq & -\|\nabla f(\psi^k) - \nabla f(\bar{x}^k)\| \|\bar{x}^k - x^k\| \\ & \geq & -L\|\bar{x}^k - x^k\|^2, \end{array}$$

where L is the Lipschitz constant for ∇f over X and $\|\psi^k - \bar{x}^k\| \leq \|x^k - \bar{x}^k\|$. This together with (28) yields

$$\limsup_{k\to\infty} f(x^k) \leq \bar{v}. \tag{30}$$

Fix any index $k \geq \bar{k}$. Since $x^k, \bar{x}^k \in X$, we have from Lemma 2(a) that

$$\nabla f(x^k)^T(\bar{x}^k - x^k) \leq (\bar{x}^k)^T p^k.$$

We also have from the Mean Value Theorem that

$$f(\bar{x}^k) - f(x^k) = \nabla f(\xi^k)^T(\bar{x}^k - x^k),$$

for some $\xi^k \in \Re^n$ lying on the line segment joining $\bar{x}^k$ with x^k. Combining these two relations and using (29), we obtain

$$\begin{array}{rcl} \bar{v} - f(x^k) & = & (\nabla f(\xi^k) - \nabla f(x^k))^T(\bar{x}^k - x^k) + \nabla f(x^k)^T(\bar{x}^k - x^k) \\ & \leq & \|\nabla f(\xi^k) - \nabla f(x^k)\| \|\bar{x}^k - x^k\| + (\bar{x}^k)^T p^k \\ & \leq & L\|\xi^k - x^k\| \|\bar{x}^k - x^k\| + (\bar{x}^k)^T p^k. \end{array}$$

Take $\bar{k}$ sufficiently large so that $\|p^k\| \leq 1$ for all $k \geq \bar{k}$. Then the above inequality, together with $\|\xi^k - x^k\| \leq \|\bar{x}^k - x^k\|$ and (27), yields

$$\bar{v} - f(x^k) \leq \kappa_3 \|p^k\|^{\min\{\gamma,1\}} \quad \forall k \geq \bar{k},$$

where $\kappa_3 > 0$ depends on $L, \max_k \|\bar{x}^k\|, \kappa_1, \tau, \gamma$. Combining this with (26) and $\bar{\gamma} = \min\{\gamma, 1\}/2$ yields

$$\bar{v} - f(x^k) \leq \bar{\kappa}(f(x^{k+1}) - f(x^k))^{\bar{\gamma}} \quad \forall k \geq \bar{k},$$

where $\bar{\kappa} = \kappa_3/(\sigma\underline{\alpha})^{\bar{\gamma}}$. Using $e^k = \bar{v} - f(x^k)$ and rearranging terms, we have

$$e^{k+1} \leq e^k - \left(\frac{e^k}{\bar{\kappa}}\right)^{1/\bar{\gamma}} \qquad \forall k \geq \bar{k}.$$

We also have from (30) and the fact $\{f(x^k)\} \uparrow$ that $f(x^k) \leq \bar{v}$ and hence $e^k \geq 0$ for all k.

Take $\bar{k}$ sufficiently large so that $e^k \leq (\bar{\gamma}^{\bar{\gamma}}\bar{\kappa})^{\frac{1}{1-\bar{\gamma}}}$ for all $k \geq \bar{k}$. Take $C \geq \bar{\kappa}^{\frac{1}{1-\bar{\gamma}}}$ sufficiently large so that (25) holds for $k = \bar{k}$. Then an induction argument shows that (25) holds for all $k \geq \bar{k}$. In particular, if (25) holds for some $k \geq \bar{k}$, then we have from $C \geq \bar{\kappa}^{\frac{1}{1-\bar{\gamma}}}$ that $(C/\bar{\kappa})^{\frac{1}{\bar{\gamma}}} \geq C$ and hence

$$e^{k+1} \leq e^k - \left(\frac{e^k}{\bar{\kappa}}\right)^{\frac{1}{\bar{\gamma}}} \leq \frac{C}{k^\omega} - \left(\frac{C}{\bar{\kappa}k^\omega}\right)^{\frac{1}{\bar{\gamma}}} \leq C\left(\frac{1}{k^\omega} - \frac{1}{k^{\frac{\omega}{\bar{\gamma}}}}\right) \leq C\,\frac{1}{(k+1)^\omega},$$

where the first inequality uses (25) and $t \mapsto t - (t/\bar{\kappa})^{\frac{1}{\bar{\gamma}}}$ being increasing on $[0, (\bar{\gamma}^{\bar{\gamma}}\bar{\kappa})^{\frac{1}{1-\bar{\gamma}}}]$; the last inequality, which is equivalent to $1 - \dfrac{1}{k^{\omega/\bar{\gamma}-\omega}} \leq \left(1 - \dfrac{1}{k+1}\right)^\omega$, holds since $\omega/\bar{\gamma} - \omega = 1$ and $(1 - \frac{1}{k+1})^\omega \geq 1 - \frac{1}{k+1}$ (using $\omega \leq 1$).

If every $\bar{x} \in \bar{X}$ satisfies (20), then $\{p^k\} \to 0$ and Lemma 2(c) yield that $\|R(x^k)\| \leq \kappa_2\|p^k\|$ for all k sufficiently large, where $\kappa_2 > 0$. Then the same argument applies, but with $\bar{\gamma} = \min\{\gamma, 1/2\}$ instead. ■

Why can we prove only sublinear convergence instead of linear convergence, as was done in [14, 15] for gradient-projection-like methods under Assumption 3? We see from the proof of Theorem 2 that linear convergence is achieved if $\bar{v} - f(x^k) = O(\|p^k\|^2)$. In contrast, the proof of Theorem 2 only shows that $\bar{v} - f(x^k) = O(\|p^k\|^{\min\{\gamma,1\}})$. This is because the upper bound on $\nabla f(x)^T(\bar{x} - x)$ in Lemma 2(a) involves $\bar{x}^T r(x)^+$ which is linear, not quadratic, in $\|r(x)^+\|$. Notice that Theorem 2 requires E to satisfy (3) with $\theta = 1$ in order to apply Lemma 2. If $\theta > 1$, can sublinear convergence still be proved?

6 Numerical Experience

In order to better understand its practical performance, we have implemented the SPRG method in Matlab to solve the special case of (1) with simplex constraint, i.e.,

$$A = e, \qquad b = 1.$$

In this section, we describe our implementation and report our numerical experience on test problems with $n \in \{1000, 10000\}$ and objective functions f from Moré et al. [19], negated for maximization. We compare the performance of the method with a reduced-gradient projection method and a first-order affine scaling method, which are also feasible ascent methods with simple iterations, as well as MINOS [20], a well-known Fortran code for constrained smooth optimization.

6.1 Implementations and Test Functions

Since $A = e$, we have $E(x) = x$ using either (7) or (11) and $E(x) = [x]^\theta / e[x]^\theta$ using (9). We use the latter more general choice, for which

$$d^k = p^k - [x^k]^\theta \frac{ep^k}{e[x^k]^\theta} \quad \text{with} \quad p^k = r(x^k)^+. \tag{31}$$

In our implementation of the SPRG method, we use the standard setting of $s = 1$, $\beta = .5$, $\sigma = .1$ for the Armijo rule (15). To save on function evaluations, we note that typically $\alpha^k \approx \alpha^{k-1}$, so we use instead

$$\bar{\alpha}^k = \min\left\{\max\left\{10^{-5}, \frac{\alpha^{k-1}}{\beta}\right\}, \min_{j:d_j^k<0} \frac{x_j^k}{-d_j^k}\right\},$$

with $\alpha^{-1} = \infty$. It can be checked that this modification to $\bar{\alpha}^k$ does not affect the global convergence or convergence rate results in Theorems 1 and 2.

Tsitsiklis and Bertsekas [27] proposed, in the context of data network routing, a reduced-gradient projection method for simplex constrained smooth optimization. This method is as simple as the SPRG method, and is provably linearly convergent under Assumption 3 and Lipschitz continuity of ∇f on X [17]. Specifically, for a given $x^k \in X$ and $g^k = \nabla f(x^k)$, this method chooses an index $j^k \in \operatorname{argmax}_{j=1,\ldots,n} g_j^k$ and a stepsize $\alpha^k > 0$, and updates $x^{k+1} = z^k[\alpha^k]$, where

$$z^k[\alpha]_j := \begin{cases} x_j^k + \alpha^k r_j^k & \text{if } j \neq j^k; \\ 1 - \sum_{j \neq j^k} z^k[\alpha]_j & \text{if } j = j^k, \end{cases}$$

with $r^k = g^k - e^T g_{j^k}^k$. The stepsize can be chosen by an Armijo-like rule along the projection arc [2, page 226]: α^k is the largest $\alpha \in \{\bar{\alpha}^k(\beta)^t\}_{t=0,1,\ldots}$ satisfying

$$f(z^k[\alpha]) \geq f(x^k) + \sigma(g^k)^T(z^k[\alpha] - x^k), \tag{32}$$

where $\bar{\alpha}^k = \min\left\{\max\{10^{-5}, \alpha^{k-1}/\beta\}, s\right\}$ and $s > 0, 0 < \beta, \sigma < 1$ are constants. We use the setting $s = 1$, $\beta = .5$, $\sigma = .1$.

Another simple method for simplex constrained smooth optimization is the first-order affine-scaling method [4, 11, 25]. Specifically, for a given $x^k \in X$ with $x^k > 0$ and $g^k = \nabla f(x^k)$, we compute

$$d^k = \operatorname{diag}(x^k)^2 r^k \quad \text{with} \quad r^k = g^k - e^T \frac{([x^k]^2)^T g^k}{\|x^k\|^2},$$

and updates $x^{k+1} = x^k + \alpha^k d^k$ for some $\alpha^k > 0$ so that $x^{k+1} > 0$. (Equivalently, d^k has the form (31) with $\theta = 2$ and $p^k = \operatorname{diag}(x^k)^2 g^k$ instead.) We choose α^k by the Armijo rule (15) with $s = 1$, $\beta = .5$, $\sigma = .1$, and

$$\bar{\alpha}^k = \min\left\{\max\left\{10^{-5}, \frac{\alpha^{k-1}}{\beta}\right\},\ .95 \min_{j:d_j^k<0} \frac{x_j^k}{-d_j^k}\right\},$$

where $\alpha^{-1} = \infty$.

The above three methods can all be efficiently implemented in Matlab using vector operations. Since the starting points given in [19] may not satisfy the simplex constraint and the affine scaling method requires $x^0 > 0$, we use the starting point

$$x^0 = e^T/n$$

for the above three methods. We also experimented with other positive starting points, and the results are qualitatively similar. We terminate the methods when

$$\|\min\{x^k, -r(x^k)\}\| \leq tol.$$

We set $tol = 10^{-3}$ in our tests, which yields solutions whose objective values are accurate to 5 significant digits. Roundoff error in Matlab occasionally causes this termination criterion never to be met, in which case we exit. In particular, we exit at iteration k if either $x^k = x^{k-1}$ or if the Armijo ascent condition (i.e., (15) or (32)) is still not met when α reaches below 10^{-20}.

MINOS (Version 5.5.1) is a well-known Fortran implementation of an active-set method for constrained smooth optimization [20]. To accomodate problems with $n = 10000$, we set Superbasics limit to $\min\{2n+1, 3000\}$ and Workspace to 8,000,000 in MINOS.

For f, we choose 9 test functions from the set of nonlinear least square functions used by Moré et al. [19] and negate them for maximization. These functions, listed in Table 1 with the numbering from [19, pages 26-28] shown in parentheses, are chosen for their diverse characteristics: convex or nonconvex, sparse or dense Hessian, well-conditioned or ill-conditioned Hessian, and are grouped accordingly. The first three functions `ER`, `DBV`, `BT` are nonconvex with sparse Hessians; `TRIG`, `BAL` are nonconvex with dense Hessians; `EPS` is convex with sparse Hessian; `VD` is strongly convex with dense Hessian; `LR1`, `LR1Z` are convex quadratic with dense Hessians of rank 1. The functions `ER` and `EPS` have block-diagonal Hessians, and `VD`, `LR1`, `LR1Z` have ill-conditioned Hessians. Upon negation, these convex functions become concave functions. The functions and gradients are coded in Fortran and in Matlab using vector operations.

Problem	**SPRG**	**RPG**	**AS**
	iter/nf/cpu/obj	iter/nf/cpu/obj	iter/nf/cpu/obj
`ER` (21)	1/2/.01/498.00	1253/2515/1.8/498.00	13/14/.02/498.00
`DBV` (28)	$328/667/.8/4.9{\cdot}10^{-7}$	$627/1267/1.7/5.1{\cdot}10^{-7}$	$8/18/.03/2.9{\cdot}10^{-7}$
`BT` (30)	4193/8387/3.6/999.03	58/122/.07/999.03	$2{\cdot}10^{6}/4.0{\cdot}10^{6}/2433.5/999.03$
`TRIG` (26)	$8/17/.03/2.7{\cdot}10^{-6}$	$541/1090/1.1/4.2{\cdot}10^{-6}$	$23/44/.05/1.2{\cdot}10^{-6}$
`BAL` (27)	$1/2/.0/9.9899{\cdot}10^{8}$	[†]$5/63/.03/9.9899{\cdot}10^{8}$	[†]$20/106/.03/9.9899{\cdot}10^{8}$
`EPS` (22)	$4191/8379/11.6/1.0{\cdot}10^{-6}$	$24340/48688/54.7/8.5{\cdot}10^{-6}$	$1162/2322/3.42/3.6{\cdot}10^{-6}$
`VD` (25)	$7/8/.01/6.2250{\cdot}10^{22}$	$1/2/.01/6.2250{\cdot}10^{22}$	$23/164/.04/6.2250{\cdot}10^{22}$
`LR1` (33)	$7/8/.01/3.3283{\cdot}10^{8}$	$1/2/0/3.3283{\cdot}10^{8}$	$82/214/.07/3.3282{\cdot}10^{8}$
`LR1Z` (34)	22/69/.02/251.12	24/790/.19/251.15	33/105/.05/251.12

Table 1: Comparing the SPRG method with reduced-gradient projection and first-order affine scaling on test functions from [19], with $n = 1000$ and $x^0 = e^T/n$.

[†] Exits due to roundoff error causing $x^k = x^{k-1}$, with $\alpha^k \approx 10^{-16}$.

6.2 Numerical Results

We now report on the performance of the SPRG method, and compare it with the performances of reduced-gradient projection (RGP), affine scaling (AS) methods, and MINOS. All runs are performed on an HP DL360 workstation, running Red Hat Linux 3.5 and Matlab (Version 7.0). MINOS is compiled using the Gnu F-77 compiler (Version 3.2.57). Tables 1 and 2 show the number of iterations, number of f-evaluations, cpu time (in seconds), and final objective value (before negation). For the SPRG method, we experimented with $\theta = 1$ and $\theta = 2$ and found $\theta = 1$ to perform consistently better. Thus we report the results for $\theta = 1$ only.

Problem (n)	**SPRG-RGP** $x^0 = e^T/n$	**SPRG-RGP** $x^0 = (1, 0, ..., 0)^T$	**MINOS** x^0 by default initialization
	iter/nf/cpu/obj	iter/nf/cpu/obj	iter/nf/cpu/obj
ER (1000)	1/16/.02/498.00	52/196/.16/498.00	1049/2105/2.49/498.00
(10000)	1/19/.11/4998.00	23/89/.57/4998.00	4555/9118/312.3/4998.00
DBV (1000)	$328/1335/1.5/4.9{\cdot}10^{-7}$	$348/1394/1.56/5.8{\cdot}10^{-7}$	$10/48/.01/5.96{\cdot}10^{-2}$
(10000)	$0/1/.03/2.0{\cdot}10^{-8}$	$2/22/.2/7.6{\cdot}10^{-8}$	$10/48/.06/5.96{\cdot}10^{-2}$
BT (1000)	180/725/.3/999.03	25/104/.05/999.03	11/24/.00/999.03
(10000)	97/393/1.98/9999.03	25/104/.56/9999.03	11/24/.02/9999.03
TRIG (1000)	$8/41/.03/2.7{\cdot}10^{-6}$	$23/84/.08/4.7{\cdot}10^{-6}$	$2023/4051/27.23/1.3{\cdot}10^{-6}$
(10000)	$3/24/.16/8.5{\cdot}10^{-7}$	$21/77/.68/7.2{\cdot}10^{-7}$	$6017/12805/763.2/2.16{\cdot}10^{-4}$
BAL (1000)	$1/3/.0/9.9899{\cdot}10^{8}$	$6/152/.04/9.9899{\cdot}10^{8}$	$2/32/1.9/9.9899{\cdot}10^{8}$
(10000)	$^*2/59/.1/9.9990{\cdot}10^{11}$	$1/6/.03/9.9990{\cdot}10^{11}$	$1/5/2.22/9.9989{\cdot}10^{11}$
EPS (1000)	$2469/9884/10.6/1.0{\cdot}10^{-6}$	$1110/4439/4.9/1.4{\cdot}10^{-6}$	$3199/6534/10.08/3.0{\cdot}10^{-7}$
(10000)	$99/407/4.2/6.7{\cdot}10^{-7}$	$88/351/3.4/1.3{\cdot}10^{-6}$	$3199/6534/13.6/3.0{\cdot}10^{-7}$
VD (1000)	$1/3/.0/6.2250{\cdot}10^{22}$	$1/3/.01/6.2250{\cdot}10^{22}$	$^\dagger 0/6/.01/6.2749{\cdot}10^{22}$
(10000)	$1/3/.01/6.2475{\cdot}10^{30}$	$1/3/.01/6.2475{\cdot}10^{30}$	$^\dagger 0/6/.01/6.2524{\cdot}10^{30}$
LR1 (1000)	$1/3/.02/3.3283{\cdot}10^{8}$	$0/1/0/3.3283{\cdot}10^{8}$	$0/5/.00/3.3283{\cdot}10^{8}$
(10000)	$1/3/.02/3.3328{\cdot}10^{11}$	$0/1/.0/3.3328{\cdot}10^{11}$	$^\ddagger 0/16/.02/3.3328{\cdot}10^{11}$
LR1Z (1000)	6/283/.09/251.12	5/279/.06/251.12	2/8/.00/251.12
(10000)	$^*2/87/.22/2571.81$	$^*1/84/.16/2571.81$	$^\ddagger 1/29/.01/3756.89$

Table 2: Comparing the SPRG-RGP hybrid method and MINOS on test functions from [19], with $n \in \{1000, 10000\}$ and different x^0.

† MINOS exits due to problem being badly scaled.

‡ MINOS exits due to current point cannot be improved upon.

* SPRG-RGP exits due to roundoff error causing Armijo ascent condition not met when $\alpha < 10^{-20}$.

We see from Table 1 that none of SPRG, RGP, AS method performs better than the others on all problems. The stepsizes for SPRG and RGP typically range between 10 and 10^{-4}. In contrast, stepsize for AS can vary widely between 10^8 to 10^{-10}. Interestingly, SPRG never performs the worst on any of the problems (either the best or second best), unlike RGP and AS. Thus SPRG has more robust performance. This motivated a hybrid method that at each iteration generates new x using both SPRG or RGP (which, unlike AS, do not require $x > 0$) and chooses the better (higher f-value) of the two. Table 2 reports the performances of this SPRG-RGP hybrid method and of MINOS, with $n \in \{1000, 10000\}$ and different x^0. In MINOS, x^0 is chosen by its default initialization procedure. MINOS has better performance (lower iter and nf) on `BT` while SPRG-RGP has better performance on `ER`, `TRIG`, `VD`, and is slightly better on `LR1`. On the remaining problems, neither clearly outperforms the other. On `DBV`, `VD`, and `LR1Z` with $n = 10000$, MINOS returns inaccurate solutions. In contrast, SPRG-RGP seems to return reasonably accurate solutions. Of course, we must keep in mind that MINOS is a general-purpose NLP solver, whereas SPRG-RGP is specialized to simplex constrained problems. For $n = 1000$ and $x^0 = e^T/n$, SPRG-RGP is outperformed by RGP on `BT` only and is outperformed by AS on `DBV`, `EPS` only.

Theorem 2 suggests that the convergence rate of SPRG may be influenced by strict complementarity (20) at stationary points. To check this, we test the methods on

$$f(x) = -(ex)^2 - \sum_{j=1}^{n-1} x_j^2,$$

with $n = 1000$. The problem has a unique stationary point at $\bar{x} = (0, ..., 0, 1)^T$, which violates strict complementary (20) at all except one component. From three positive starting points e^T/n, $(.5, \frac{.5}{n-1}, ..., \frac{.5}{n-1})^T$, $\frac{2}{n(n+1)}(n, n-1, ..., 1)^T$, SPRG takes 1, 27, 56 iterations, respectively, to converge to $\bar{x}$ with $tol = 10^{-3}$. In contrast, RGP and hybrid take 1, 1, 1 iterations and AS takes 2, 19, 228 iterations, respectively. From starting point $(1, 0, ..., 0)^T$ (not positive), SPRG takes 29 iterations while RGP and SPRG-RGP hybrid take, respectively, 1007 and 12 iterations. Thus SPRG seems not adversely affected by a lack of strict complementarity and the hybrid method has best overall performance.

7 Conclusions and Future Directions

We have proposed a new feasible ascent method, based on scaled projected reduced-gradient direction, for linearly constrained smooth optimization. This method has features of reduced-gradient methods and first-order affine scaling methods, and can start anywhere in the feasible set and achieve global convergence. Under a Hölderian error bound assumption, it achieves sublinear convergence. Numerical experience on simplex constrained problems suggests that the method is promising for solving large problems, especially when it is combined with reduced-gradient projection.

There are various directions for future study. Can the primal nondegeneracy assumption (Assumption 1(b)) be relaxed? Can the sublinear convergence result in Theorem 2 be improved to linear convergence? Can the new method be extended to conic programs? What about implementation and testing of the method for general linear constraints?

References

[1] Avriel, M., Nonlinear Programming: Analysis and Methods, Prentice-Hall, Inglewood Cliffs, 1976.

[2] Bertsekas, D. P., Nonlinear Programming, 2nd edition, Athena Scientific, Belmont, 1999.

[3] Bomze, I. M., Branch-and-bound approaches to standard quadratic optimization problems, J. Global Optim. 22 (2002), 17–37.

[4] Bonnans, J. F. and Pola, C., A trust region interior point algorithm for linearly constrained optimization, SIAM J. Optim. 7 (1997), 717–731.

[5] Coleman, T. F. and Li, Y., A trust region and affine scaling interior point method for nonconvex minimization with linear inequality constraints, Math. Program. 88 (2000), 1–32.

[6] Cottle, R., Pang, J.-S., and Stone R., The Linear Complementarity Problem, Academic Press, New York, 1992.

[7] Dikin, I. I., Iterative solution of problems of linear and quadratic programming, Soviet Math. Dokl. 8 (1967), 674-675.

[8] Facchinei, F. and Pang, J.-S., Finite-Dimensional Variational Inequalities and Complementarity Problems, Vols. I and II, Springer-Verlag, New York, 2003.

[9] Forsgren, A., Gill, P. E., and Wright, M. H., Interior methods for nonlinear optimization, SIAM Rev. 44 (2002), 525-597.

[10] Gill, P. E., Murray, W., and Wright, M. H., Practical Optimization, Academic Press, New York, 1981.

[11] Gonzaga, C. C. and Carlos, L. A., A primal affine scaling algorithm for linearly constrained convex programs, Tech. Report ES-238/90, Department of Systems Engineering and Computer Science, COPPE Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, December 1990.

[12] Karmarkar, N., A new polynomial-time algorithm for linear programming, Combinatorica 4 (1984), 373-395.

[13] Luo, Z.-Q., New error bounds and their applications to convergence analysis of iterative algorithms, Math. Program. 88 (2000), 341–355.

[14] Luo, Z.-Q. and Tseng, P., Error bounds and the convergence analysis of matrix splitting algorithms for the affine variational inequality problem, SIAM J. Optim. 2 (1992), 43–54.

[15] Luo, Z.-Q. and Tseng, P., Error bounds and convergence analysis of feasible descent methods: a general approach, Ann. Oper. Res. 46 (1993), 157–178.

[16] Luo, Z.-Q. and Tseng, P., Error bound and reduced gradient projection algorithms for convex minimization over a polyhedral set, SIAM J. Optim. 3 (1993), 43–59.

[17] Luo, Z.-Q. and Tseng, P., On the rate of convergence of a distributed asynchronous routing algorithm, IEEE Trans. Automat. Control 39 (1994), 1123–1129.

[18] Monteiro, R. D. C. and Wang, Y., Trust region affine scaling algorithms for linearly constrained convex and concave programs, Math. Program. 80 (1998), 283–313.

[19] Moré, J. J., Garbow, B. S., and Hillstrom, K. E., Testing unconstrained optimization software, ACM Trans. Math. Software 7 (1981), 17–41.

[20] Murtagh, B. A. and Saunders, M. A., MINOS 5.5 user's guide, Report SOL 83-20R, Department of Operations Research, Stanford University, Stanford (Revised July 1998).

[21] Pang, J.-S., Error bounds in mathematical programming, Math. Program. 79 (1997), 299–332.

[22] Smeers, Y., Generalized reduced gradient method as an extension of feasible direction methods, J. Optim. Theory Appl. 22 (1977), 209–226.

[23] Sun, J., A convergence proof for an affine-scaling algorithm for convex quadratic programming without nondegeneracy assumptions, Math. Program. 60 (1993), 69–79.

[24] Sun, J., A convergence analysis for a convex version of Dikin's algorithm, Ann. Oper. Res. 62 (1996), 357–374.

[25] Tseng, P., Partial affine-scaling for linearly constrained minimization, Math. Oper. Res. 20 (1995), 678–694

[26] Tseng, P., Convergence properties of Dikin's affine scaling algorithm for nonconvex quadratic minimization, J. Global Optim. 30 (2004), 285–300.

[27] Tsitsiklis J. N. and Bertsekas, D. P., Distributed asynchronous optimal routing in data networks, IEEE Trans. Automat. Contr. AC-31 (1986), 325–332.

[28] Ye, Y., On affine scaling algorithms for nonconvex quadratic programming, Math. Program. 56 (1992), 285–300.