SIAM J. OPTIM.
Vol. 23, No. 2, pp. 1062–1091

A PROXIMAL-GRADIENT HOMOTOPY METHOD FOR THE SPARSE LEAST-SQUARES PROBLEM*

LIN XIAO† AND TONG ZHANG‡

Abstract. We consider solving the ℓ_1-regularized least-squares (ℓ_1-LS) problem in the context of sparse recovery for applications such as compressed sensing. The standard proximal gradient method, also known as iterative soft-thresholding when applied to this problem, has low computational cost per iteration but a rather slow convergence rate. Nevertheless, when the solution is sparse, it often exhibits fast linear convergence in the final stage. We exploit the local linear convergence using a homotopy continuation strategy, i.e., we solve the ℓ_1-LS problem for a sequence of decreasing values of the regularization parameter, and use an approximate solution at the end of each stage to warm start the next stage. Although similar strategies have been studied in the literature, there have been no theoretical analysis of their global iteration complexity. This paper shows that under suitable assumptions for sparse recovery, the proposed homotopy strategy ensures that all iterates along the homotopy solution path are sparse. Therefore the objective function is effectively strongly convex along the solution path, and geometric convergence at each stage can be established. As a result, the overall iteration complexity of our method is $O(\log(1/\epsilon))$ for finding an ϵ-optimal solution, which can be interpreted as global geometric rate of convergence. We also present empirical results to support our theoretical analysis.

Key words. sparse optimization, proximal gradient method, homotopy continuation

AMS subject classifications. 65C60, 65H20, 65Y20, 90C25

DOI. 10.1137/120869997

1. Introduction.

In this paper, we propose and analyze an efficient numerical method for solving the ℓ_1*-regularized least-squares* (ℓ_1-LS) problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1, \tag{1.1}$$

where $x \in \mathbb{R}^n$ is the vector of unknowns, $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are the problem data, and $\lambda > 0$ is a regularization parameter. Here $\|\cdot\|_2$ denotes the standard Euclidean norm, and $\|x\|_1 = \sum_i |x_i|$ is the ℓ_1 norm of x. This is a convex optimization problem, and we use $x^\star(\lambda)$ to denote its (global) optimal solution. Since the ℓ_1 term promotes sparse solutions, we also refer to problem (1.1) as the *sparse least-squares* problem.

The ℓ_1-LS problem has important applications in machine learning, signal processing, and statistics; see, e.g., [36, 13, 8]. It has received revived interests in recent years due to the emergence of *compressed sensing* theory, which builds upon the fundamental idea that a finite-dimensional signal having a sparse or compressible representation can be recovered from a small set of linear, nonadaptive measurements [9, 11, 17]. We are especially interested in solving the ℓ_1-LS problem in such a context with the goal of recovering a sparse vector under measurement noise. More precisely, we assume A and b in (1.1) are related by a linear model

$$b = A\bar{x} + z,$$

*Received by the editors March 14, 2012; accepted for publication (in revised form) March 21, 2013; published electronically May 28, 2013.
http://www.siam.org/journals/siopt/23-2/86999.html
†Machine Learning Group, Microsoft Research, Redmond, WA 98052 (lin.xiao@microsoft.com).
‡Statistics Department, Rutgers University, Piscataway, NJ 08854 (tzhang@stat.rutgers.edu).

where $\bar{x}$ is the sparse vector we would like to recover in statistical applications, and z is a noise vector. We assume that the noise level, measured by $\|A^T z\|_\infty$, is relatively small compared with λ. This scenario is of great modern interest, and various properties of the solution $x^\star(\lambda)$ have been investigated [10, 16, 27, 38, 48, 12, 46, 47, 6, 24, 42, 43]. In particular, it is known that under suitable conditions on A such as the *restricted isometry property* (RIP) [10], and as long as $\lambda \geq c\|A^T z\|_\infty$ (for some universal constant c), one can obtain a recovery bound of the form

$$\|x^\star(\lambda) - \bar{x}\|_2^2 = O\left(\lambda^2 \|\bar{x}\|_0\right), \tag{1.2}$$

where $\|\bar{x}\|_0$ denotes the number of nonzero elements in $\bar{x}$. The constant in $O(\cdot)$ depends only on the RIP condition, and this bound achieves the optimal order of recovery. Moreover, it is known that in this situation, the solution $x^\star(\lambda)$ is sparse [46], and the sparsity of the solution is closely related to the recovery performance.

In this paper, we develop an efficient numerical method for solving the ℓ_1-LS problem in the context of sparse recovery described above. In particular, we focus on the case when $m < n$ (i.e., the linear system $Ax = b$ is underdetermined) and the solution $x^\star(\lambda)$ is sparse (which requires the parameter λ to be sufficiently large). Under such assumptions, our method has provable lower complexity than previous algorithms.

1.1. Previous algorithms. There has been extensive research on numerical methods for solving problem (1.1) and its constrained variations. A nice survey of major practical algorithms for sparse approximation appeared in [37], and performance comparisons of various algorithms can be found in, e.g., [45, 44, 2]. Here we briefly summarize the computational complexities of several methods that are most relevant for solving the ℓ_1-LS problem (1.1) in terms of finding an ϵ-optimal solution (i.e., obtaining an objective value within ϵ of the global minimum).

Interior-point methods (IPMs) were among the first approaches used for solving the ℓ_1-LS problem [13, 41, 23]. The theoretical bound on their iteration complexity is $O\left(\sqrt{n}\log(1/\epsilon)\right)$, although their practical performance demonstrates much weaker dependence on n. The bottleneck of their performance is the computational cost per iteration. For example, with an unstructured dense matrix A, the standard approach of solving the normal equation in each iteration with a direct method (Cholesky factorization) would cost $O(m^2 n)$ flops, which is prohibitive for large-scale applications. Therefore all customized solvers [13, 41, 23] use iterative methods (such as conjugate gradients) for solving the linear equations.

Proximal gradient (*PG*) *methods* for solving the ℓ_1-LS problem take the following basic form at each iteration $k = 0, 1, \ldots$:

$$x^{(k+1)} = \arg\min_y \left\{ f(x^{(k)}) + \nabla f(x^{(k)})^T (y - x^{(k)}) + \frac{L_k}{2}\|y - x^{(k)}\|_2^2 + \lambda\|y\|_1 \right\}, \tag{1.3}$$

where we used the shorthand $f(x) = (1/2)\|Ax - b\|_2^2$, and L_k is a parameter chosen by line search. The minimization problem in (1.3) has a closed-form solution

$$x^{(k+1)} = \text{soft}\left(x^{(k)} - \frac{1}{L_k}\nabla f(x^{(k)}),\ \frac{\lambda}{L_k}\right), \tag{1.4}$$

where $\text{soft} : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^n$ is the well-known *soft-thresholding* operator, defined as

$$(\text{soft}(x, \alpha))_i = \text{sgn}(x_i)\max\{|x_i| - \alpha,\ 0\}, \quad i = 1, \ldots, n. \tag{1.5}$$

Iterative methods that use the update rule (1.4) include [15, 14, 32, 22, 45]. Their major computational effort per iteration is to form the gradient $\nabla f(x) = A^T(Ax - b)$,

which costs $O(mn)$ flops for a generic dense matrix A. With suitable choices of L_k, the PG method (1.3) has an iteration complexity $O(1/\epsilon)$.

Indeed, the iteration complexity $O(\log(1/\epsilon))$ can be established for (1.3) if $m \geq n$ and A has full column rank, since in this case the objective function in (1.1) is strongly convex [32]. Unfortunately this result is not applicable to the case $m < n$. Nevertheless, when the solution $x^\star(\lambda)$ is sparse and the active submatrix is well conditioned (e.g., when A has RIP), local linear convergence can be established [26, 22], and fast convergence in the final stage of the algorithm has also been observed [32, 22, 45]. There are also coordinate descent variants of PG methods that achieve a local linear rate of convergence (e.g., [39]).

Variations and extensions of the PG method have been proposed to speed up the convergence in practice; see, e.g., [7, 45, 44]. Nesterov's optimal gradient methods for minimizing smooth convex functions [28, 30, 31] have also been extended to minimize composite objective functions such as in the ℓ_1-LS problem [32, 40, 4, 2]. These accelerated methods have the iteration complexity $O(1/\sqrt{\epsilon})$. They typically generate two or three concurrent sequences of iterates, but their computational cost per iteration is still $O(mn)$, which is the same as simple gradient methods.

Exact homotopy path-following methods were developed in the statistics literature to compute the complete regularization path when varying the parameter λ from large to small [33, 34, 18]. These methods exploit the piecewise linearity of the solution as a function of λ and identify the next breakpoint along the solution path by examining the optimality conditions (also called *active set* or *pivoting* method in optimization). With efficient numerical implementations (using updating or downdating of submatrix factorizations), the computational cost at each breakpoint is $O(mn + ms^2)$, where s is the number of nonzeros in the solution at the breakpoint. Such methods can be quite efficient if s is small. However, in general, there is no convergence result for bounding the number of breakpoints for this class of methods.

1.2. Proposed approach and contributions. We consider an *approximate* homotopy continuation method, where the key idea is to solve (1.1) with a large regularization parameter λ first and then gradually decrease λ until the target regularization is reached. For each fixed λ, we employ a PG method of the form (1.3) to solve (1.1) up to an adequate precision (to be specified later) and then use this approximate solution to serve as the initial point for the next value of λ. We call the resulting method the *proximal gradient homotopy* (PGH) method.

This is not a new idea. Similar approximate homotopy methods has been studied in, e.g., [22, 45, 44], and superior empirical performance has been reported when the solution is sparse. However, there has been no effective theoretical analysis for their overall iteration complexity. As a result, some important algorithmic choices are mostly based on heuristics and ad hoc factors. More specifically, how do we choose the sequence of decreasing values for λ? and how accurate should we solve the problem (1.1) for each value in this sequence?

In this paper, we present a PGH method that has provable low iteration complexity, along with the following specific algorithmic choices:

- We use a decreasing geometric sequence for the values of λ. That is, we choose a λ_0 and a parameter $\eta \in (0, 1)$ and let $\lambda_K = \eta^K \lambda_0$ for $K = 1, 2, \ldots$ until the target value is reached.
- We choose a parameter $\delta \in (0, 1)$ and solve problem (1.1) for each λ_K with a proportional precision $\delta\lambda_K$ (in terms of violating the optimality condition), except that for the final target value of λ, we reach the absolute precision ϵ.

- We use Nesterov's adaptive line search strategy in [32] to choose the parameters L_k in the PG method (1.3).

Under the assumptions that the target value of λ is sufficiently large and the matrix A satisfies a RIP-like condition, our PGH method exhibits geometric convergence at each stage, and the overall iteration complexity is $O(\log(1/\epsilon))$. The constant in $O(\cdot)$ depends on the RIP-like condition. Moreover, the solution satisfies a recovery bound of the optimal form (1.2). Since each iteration of the PG method costs $O(mn)$ flops, the overall computational cost is $O(mn\log(1/\epsilon))$.

The advantage of our method over the exact homotopy path-following approach [33, 34, 18] is that there is no need to keep track of all breakpoints. In fact, for large-scale problems, the total number of proximal gradient steps in our method can be much smaller than the number of nonzeros in the target solution, which is the minimum number of breakpoints the exact homotopy methods have to compute.

Compared with IPMs, our method has a similar iteration complexity (actually better in terms of theoretical bounds) and computationally can be much more efficient for each iteration. The approximate homotopy strategy used in this paper is also analogous to the long-step path-following IPMs (e.g., [29]), in the sense that the least-squares problem becomes better conditioned near the regularization path (cf. *central path* in IPMs). However, our results hold only for problems with provable sparse solutions, and the parameters η and δ depend on the problem data A and the regularization parameter λ. In contrast, the performance of IPMs is insensitive to the sparsity of the solution or the regularization parameter.

As an important special case, our results can be immediately applied to noise-free compressed-sensing applications. Consider the *basis pursuit (BP)* problem

$$\text{minimize} \quad \|x\|_1 \quad \text{subject to} \quad Ax = b. \tag{1.6}$$

Its solution can be obtained by running our PGH method on the ℓ_1-LS problem (1.1) with $\lambda \to 0$. In terms of satisfying the condition $\lambda > c\,\|Az\|_\infty$, any $\lambda > 0$ is sufficiently large in the noise-free case because $z = 0$. Therefore, the global geometric convergence of the PGH method for BP is just a special case of the more general result for (1.1) developed in this paper.

1.3. Outline of the paper. In section 2, we review some preliminaries that are necessary for developing our method and its convergence analysis. In section 3, we present our PGH method and state the assumptions and the main convergence results. Section 4 is devoted to the proofs of our convergence results. We present numerical experiments in section 5 to support our theoretical analysis, and we conclude in section 6 with some further discussions.

2. Preliminaries and notations. In this section, we first review composite gradient mapping and some of its key properties developed in [32]. Then we describe Nesterov's PG method with adaptive line search, which we will use at each stage of our PGH method. Finally we discuss the restricted eigenvalue conditions that allow us to show the local linear convergence of the PG method.

2.1. Composite gradient mapping. Consider the following optimization problem with *composite* objective function:

$$\underset{x}{\text{minimize}} \quad \Big\{\phi(x) \triangleq f(x) + \Psi(x)\Big\}, \tag{2.1}$$

where the function f is convex and differentiable and Ψ is closed and convex on $\mathbb{R}^n$. The optimality condition of (2.1) states that $x^\star$ is a solution if and only if there exists

$\xi \in \partial\Psi(x^\star)$ such that $\nabla f(x^\star) + \xi = 0$ (see, e.g., [35, section 27]). Therefore, a good measure of accuracy for any x as an approximate solution is the quantity

$$\omega(x) \triangleq \min_{\xi \in \partial\Psi(x)} \|\nabla f(x) + \xi\|_\infty. \tag{2.2}$$

We call $\omega(x)$ the *optimality residue* of x. We will use it in the stopping criterion of the PG method.

Composite gradient mapping was introduced by Nesterov in [32]. For any fixed point y and a given constant $L > 0$, we define a local model of $\phi(x)$ around y using a simple quadratic approximation of f but keeping Ψ intact:

$$\psi_L(y; x) = f(y) + \nabla f(y)^T(x - y) + \frac{L}{2}\|x - y\|_2^2 + \Psi(x).$$

Let $T_L(y)$ denote the unique minimizer of $\psi_L(y; x)$, i.e.,

$$T_L(y) = \arg\min_x \ \psi_L(y; x). \tag{2.3}$$

Then the *composite gradient mapping* of f at y is defined as

$$g_L(y) = L(y - T_L(y)).$$

In the case $\Psi(x) = 0$, it is easy to verify that $g_L(y) = \nabla f(y)$ for any $L > 0$, and $1/L$ can be considered as the step-size from y to $T_L(y)$ along the direction $-g_L(y)$. The following property of composite gradient mapping was shown in [32, Theorem 2].

LEMMA 2.1. *For any $L > 0$, we have*

$$\psi_L(y; T_L(y)) \le \phi(y) - \frac{1}{2L}\|g_L(y)\|_2^2.$$

The function f has a Lipschitz continuous gradient if there is a constant L_f such that

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L_f\|x - y\|_2 \quad \forall\, x, y \in \mathbb{R}^n.$$

A direct consequence of having a Lipschitz continuous gradient is the following inequality (see, e.g., [30, Theorem 2.1.5]):

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_f}{2}\|y - x\|_2^2 \quad \forall\, x, y \in \mathbb{R}^n. \tag{2.4}$$

For such functions, we can measure how close $T_L(y)$ is from satisfying the optimality condition by using the norm of the composite gradient mapping at y.

LEMMA 2.2. *If f has Lipschitz continuous gradients with a constant L_f, then*

$$\omega(T_L(y)) \le \left(1 + \frac{S_L(y)}{L}\right)\|g_L(y)\|_2 \le \left(1 + \frac{L_f}{L}\right)\|g_L(y)\|_2,$$

where $S_L(y)$ is a local Lipschitz constant defined as

$$S_L(y) = \frac{\|\nabla f(T_L(y)) - \nabla f(y)\|_2}{\|T_L(y) - y\|_2}.$$

ALGORITHM 1. $\{x^+, M\} \leftarrow \texttt{LineSearch}(\lambda, x, L)$.

input: $\lambda > 0$, $x \in \mathbb{R}^n$, $L > 0$.
parameter: $\gamma_{\text{inc}} > 1$
repeat
  $x^+ \leftarrow T_{\lambda,L}(x)$
  if $\phi_\lambda(x^+) > \psi_{\lambda,L}(x; x^+)$ **then** $L \leftarrow L\gamma_{\text{inc}}$
until $\phi_\lambda(x^+) <= \psi_{\lambda,L}(x; x^+)$
$M \leftarrow L$
return $\{x^+, M\}$

ALGORITHM 2. $\{\hat{x}, \hat{M}\} \leftarrow \texttt{ProxGrad}(\lambda, \hat{\epsilon}, x^{(0)}, L_0)$.

input: $\lambda > 0$, $\hat{\epsilon} > 0$, $x^{(0)} \in \mathbb{R}^n$, $L_0 \geq L_{\min}$.
parameters: $L_{\min} > 0$, $\gamma_{\text{dec}} \geq 1$
repeat for $k = 0, 1, 2, \ldots$
  $\{x^{(k+1)}, M_k\} \leftarrow \texttt{LineSearch}(\lambda, x^{(k)}, L_k)$
  $L_{k+1} \leftarrow \max\{L_{\min}, M_k/\gamma_{\text{dec}}\}$
until $\omega_\lambda(x^{(k+1)}) \leq \hat{\epsilon}$
$\hat{x} \leftarrow x^{(k+1)}$
$\hat{M} \leftarrow M_k$
return $\{\hat{x}, \hat{M}\}$

The proof of this lemma follows from [32, Corollary 1] and the relationship between $\omega(x)$ and the directional derivatives of ϕ [32, section 2]. The details are omitted here.

In this paper, we use the following notation to simplify presentation:

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2, \qquad \phi_\lambda(x) = f(x) + \lambda\|x\|_1.$$

Accordingly, we add a subscript λ in the above definitions related to gradient mapping:

$$\begin{aligned}
\omega_\lambda(x) &= \min_{\xi \in \partial\|x\|_1} \|\nabla f(x) + \lambda\xi\|_\infty, \\
\psi_{\lambda,L}(y; x) &= f(y) + \nabla f(y)^T(x - y) + \frac{L}{2}\|x - y\|_2^2 + \lambda\|x\|_1, \\
T_{\lambda,L}(y) &= \arg\min_x \ \psi_{\lambda,L}(y; x), \\
g_{\lambda,L}(y) &= L\big(y - T_{\lambda,L}(y)\big).
\end{aligned}$$

For the ℓ_1-LS problem, $T_{\lambda,L}(x)$ has the closed-form solution given in (1.4). Given the gradient $\nabla f(x)$, the optimality residue $\omega_\lambda(x)$ can be easily computed with $O(n)$ flops.

2.2. Nesterov's gradient method with adaptive line search. With the machinery of composite gradient mapping, Nesterov developed several variants of PG methods in [32]. We use the nonaccelerated primal-gradient version described in Algorithms 1 and 2, which correspond to (3.1) and (3.2) in [32], respectively. To use this algorithm, we need to first choose an initial optimistic estimate $L_{\min}$ for the Lipschitz constant L_f,

$$0 < L_{\min} \leq L_f,$$

and two adjustment parameters $\gamma_{\text{dec}} \geq 1$ and $\gamma_{\text{inc}} > 1$. The adaptive line search scheme always tries to use a smaller Lipschitz constant first at each iteration.

Each iteration of the PG method takes the form of

$$x^{(k+1)} = T_{\lambda,M_k}(x^{(k)}),$$

where M_k is chosen by the line search procedure in Algorithm 1. The line search procedure starts with an estimated Lipschitz constant L_k and increases its value by the factor γ_{inc} until the stopping criteria is satisfied. The stopping criteria ensures

$$\begin{aligned} \phi_\lambda(x^{(k+1)}) &\leq \psi_{\lambda,M_k}\left(x^{(k)}, x^{(k+1)}\right) \;=\; \psi_{\lambda,M_k}\left(x^{(k)}, T_{\lambda,M_k}(x^{(k)})\right) \\ &\leq \phi_\lambda(x^{(k)}) - \frac{1}{2M_k}\left\|g_{\lambda,M_k}(x^{(k)})\right\|_2^2, \end{aligned} \tag{2.5}$$

where the last inequality follows from Lemma 2.1. Therefore, we have the objective value $\phi_\lambda(x^{(k)})$ decrease monotonically with k, unless the gradient mapping $g_{\lambda,M_k}(x^{(k)}) = 0$. In the latter case, according to Lemma 2.2, $x^{(k+1)}$ is an optimal solution.

Since f has Lipschitz constant L_f, the inequality (2.4) implies that the line search procedure is guaranteed to terminate if $L \geq L_f$. Therefore, we have

$$L_{\min} \leq L_k \leq M_k < \gamma_{\text{inc}} L_f. \tag{2.6}$$

Although there is no explicit bound on the number of repetitions in the line search procedure, Nesterov showed that the total number of line searches cannot be too big. More specifically, let N_k be the number of operations $x^+ \leftarrow T_{\lambda,L}(x)$ after k iterations in Algorithm 2. Lemma 3 in [32] showed that

$$N_k \;\leq\; \left(1 + \frac{\ln \gamma_{\text{dec}}}{\ln \gamma_{\text{inc}}}\right)(k+1) + \frac{1}{\ln \gamma_{\text{inc}}} \max\left\{\ln \frac{\gamma_{\text{inc}} L_f}{\gamma_{\text{dec}} L_{\min}}, 0\right\}.$$

For example, if we choose $\gamma_{\text{inc}} = \gamma_{\text{dec}} = 2$, then

$$N_k \leq 2(k+1) + \log_2 \frac{L_f}{L_{\min}}. \tag{2.7}$$

Nesterov established the following iteration complexities of Algorithm 2 for finding an ϵ-optimal solution of the problem (2.1):

- If ϕ_λ is convex but not strongly convex, then the convergence is sublinear, with an iteration complexity $O(1/\epsilon)$ [32, Theorem 4].
- If ϕ_λ is strongly convex, then the convergence is geometric, with an iteration complexity $O(\log(1/\epsilon))$ [32, Theorem 5].

A nice property of this algorithm is that we do not need to know a priori if the objective function is strongly convex or not. It will automatically exploit the strong convexity whenever it holds. The algorithm is the same for both cases.

For our case $m < n$, the objective function in Problem (1.1) is not strongly convex. Therefore, if we directly use Algorithm 2 to solve this problem, we can only get the $O(1/\epsilon)$ iteration complexity (even though fast local linear convergence was observed in [32] when the solution is sparse). Nevertheless, we can use a homotopy continuation strategy (see section 1.2) to enforce that all iterates along the solution path are sufficiently sparse. Under a RIP-like assumption on A, this implies that the objective function is effectively strongly convex along the homotopy path, and hence a global geometric rate can be established using Nesterov's analysis. Next we explain conditions that characterize restricted strong convexity for sparse vectors.

2.3. Restricted eigenvalue conditions. We first define some standard notation for sparse recovery. For a vector $x \in \mathbb{R}^n$, let

$$\operatorname{supp}(x) = \{j : x_j \neq 0\}, \qquad \|x\|_0 = |\operatorname{supp}(x)|.$$

Throughout the paper, we denote $\operatorname{supp}(\bar{x})$ by $\bar{S}$ and use $\bar{S}^c$ for its complement. We use the notation $x_{\bar{S}}$ and $x_{\bar{S}^c}$ to denote the restrictions of a vector x to the coordinates indexed by $\bar{S}$ and $\bar{S}^c$, respectively.

Various conditions for sparse recovery have appeared in the literature. The most well-known of such conditions is the RIP introduced in [10]. In this paper, we analyze the numerical solution of the ℓ_1-LS problem under a slight generalization, which we refer to as *restricted eigenvalue condition*.

DEFINITION 2.3. *Given an integer $s > 0$, we say that A satisfies the restricted eigenvalue condition at sparsity level s if there exist positive constants $\rho_-(A,s)$ and $\rho_+(A,s)$ such that*

$$\rho_+(A,s) = \sup\left\{\frac{x^T A^T A x}{x^T x} : x \neq 0,\ \|x\|_0 \leq s\right\},$$
$$\rho_-(A,s) = \inf\left\{\frac{x^T A^T A x}{x^T x} : x \neq 0,\ \|x\|_0 \leq s\right\}.$$

Note that a matrix A satisfies the original definition of restricted isometry property with RIP constant ν at sparsity level s if and only if $\rho_+(A,s) \leq 1+\nu$ and $\rho_-(A,s) \geq 1-\nu$. More generally, the strong convexity of the objective function in (1.1), namely, $\phi_\lambda(x)$, is equivalent to $\rho_-(A,n) > 0$. However, since we are interested in the situation of $m < n$, which implies that $\rho_-(A,n) = 0$, we know that ϕ_λ is not strongly convex. Nevertheless, for $s < m$, it is still possible that the condition $\rho_-(A,s) > 0$ holds. This means that if both x and y are sparse vectors, then ϕ_λ is strongly convex along the line segment that connects x and y. Moreover, the inequality that characterize the smoothness of the function, namely, (2.4), could use a much smaller restricted Lipschitz constant instead of the global constant $L_f = \rho_+(A,n)$. The following lemma follows directly from the fact $f(x) = (1/2)\|Ax-b\|_2^2$ and the definition of restricted eigenvalues.

LEMMA 2.4. *Let $f(x) = (1/2)\|Ax-b\|_2^2$. Suppose x and y are two sparse vectors such that $|\operatorname{supp}(x) \cup \operatorname{supp}(y)| \leq s$ for some integer $s < m$. Then the following two inequalities hold:*

$$f(y) \leq f(x) + \langle \nabla f(x), y-x\rangle + \frac{\rho_+(A,s)}{2}\|y-x\|_2^2, \tag{2.8}$$

$$f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle + \frac{\rho_-(A,s)}{2}\|y-x\|_2^2. \tag{2.9}$$

The inequality (2.8) represents *restricted smoothness*, and (2.9) represents *restricted strong convexity*. We also define the *restricted condition number* as

$$\kappa(A,s) = \frac{\rho_+(A,s)}{\rho_-(A,s)}. \tag{2.10}$$

In particular, if A has RIP constant ν at sparsity level s, then $\kappa(A,s) \leq (1+\nu)/(1-\nu)$.

ALGORITHM 3. $\hat{x}^{(\mathrm{tgt})} \leftarrow \texttt{Homotopy}(A, b, \lambda_{\mathrm{tgt}}, \epsilon, L_{\min})$.

input: $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^n$, $\lambda_{\mathrm{tgt}} > 0$, $\epsilon > 0$, $L_{\min} > 0$.
parameters: $\eta \in (0,1)$, $\delta \in (0,1)$
initialize: $\lambda_0 \leftarrow \|A^T b\|_\infty$, $\hat{x}^{(0)} \leftarrow 0$, $\hat{M}_0 \leftarrow L_{\min}$
$N \leftarrow \lfloor \ln(\lambda_0/\lambda_{\mathrm{tgt}}) \,/\, \ln(1/\eta) \rfloor$
for $K = 0, 1, 2, \ldots, N-1$ **do**
  $\lambda_{K+1} \leftarrow \eta\lambda_K$
  $\hat{\epsilon}_{K+1} \leftarrow \delta\lambda_{K+1}$
  $\{\hat{x}^{(K+1)}, \hat{M}_{K+1}\} \leftarrow \texttt{ProxGrad}\big(\lambda_{K+1}, \hat{\epsilon}_{K+1}, \hat{x}^{(K)}, \hat{M}_K\big)$
end
$\{\hat{x}^{(\mathrm{tgt})}, \hat{M}_{\mathrm{tgt}}\} \leftarrow \texttt{ProxGrad}\big(\lambda_{\mathrm{tgt}}, \epsilon, \hat{x}^{(N)}, \hat{M}_N\big)$
return $\hat{x}^{(\mathrm{tgt})}$

3. A PGH method. The key idea of the PGH method is to solve (1.1) with a large regularization parameter λ_0 first and then gradually decrease λ until the target regularization is reached. For each fixed λ, we employ Nesterov's PG method described in Algorithms 1 and 2 to solve problem (1.1) up to an adequate precision. Then we use this approximate solution to warm start the PG method for the next value of λ.

Our proposed PGH method is given as Algorithm 3. For convenience, we use λ_{tgt} to denote the target regularization parameter. The method starts with

$$\lambda_0 = \|A^T b\|_\infty,$$

since this is the smallest value for λ such that the ℓ_1-LS problem has the trivial solution 0 (by examining the optimality condition). Our method has two parameters $\eta \in (0,1)$ and $\delta \in (0,1)$. They control the algorithm as follows:

- The sequence of values for the regularization parameter is determined as $\lambda_K = \eta^K \lambda_0$ for $K = 1, 2, \ldots$, until the target value λ_{tgt} is reached.
- For each λ_K except λ_{tgt}, we solve problem (1.1) with a proportional precision $\delta\lambda_K$. For the last stage with λ_{tgt}, we solve the problem with the absolute precision ϵ.

As discussed in the introduction, sparse recovery by solving the ℓ_1-LS problem requires two types of conditions: the regularization parameter λ is relatively large compared with the noise level, and the matrix A satisfies certain RIPs or restricted eigenvalue conditions. It turns out that such conditions are also sufficient for fast convergence of our PGH method. More precisely, we have the following assumption.

Assumption 3.1. *Suppose* $b = A\bar{x} + z$. *Let* $\bar{S} = \mathrm{supp}(\bar{x})$ *and* $\bar{s} = |\bar{S}|$. *There exist* $\gamma > 0$ *and* $\delta' \in (0, 0.2]$ *such that* $\gamma > (1+\delta')/(1-\delta')$ *and*

$$\lambda_{\mathrm{tgt}} \;\geq\; 4\max\left\{2,\; \frac{\gamma+1}{(1-\delta')\gamma - (1+\delta')}\right\} \|A^T z\|_\infty. \tag{3.1}$$

Moreover, there exists an integer $\tilde{s}$ *such that* $\rho_-(A, \bar{s}+2\tilde{s}) > 0$ *and*

$$\tilde{s} \;>\; \frac{8\big(\gamma_{\mathrm{inc}}\rho_+(A, \bar{s}+2\tilde{s}) + \rho_+(A, \tilde{s})\big)}{\rho_-(A, \bar{s}+\tilde{s})}(1+\gamma)\bar{s}. \tag{3.2}$$

We also assume that $L_{\min} \leq \gamma_{\mathrm{inc}}\rho_+(A, \bar{s}+2\tilde{s})$.

As we will see later, the quantity δ' in the above assumption is related to the parameter δ in Algorithm 3, and γ defines a conic condition on $x - \bar{x}$, i.e.,

$$\|(x-\bar{x})_{\bar{S}^c}\|_1 \le \gamma \|(x-\bar{x})_{\bar{S}}\|_1,$$

which holds whenever $\omega_\lambda(x) \le \delta'\lambda$. According to [46], the above assumption implies that the solution $x^\star(\lambda)$ of (1.1) is sparse whenever $\lambda \ge \lambda_{\mathrm{tgt}}$; more specifically, $\|x^\star(\lambda)_{\bar{S}^c}\|_0 \le \tilde{s}$. (Here $\bar{S}^c$ denotes the complement of the support set $\bar{S}$.) In this paper, we will show that by choosing the parameters η and δ in Algorithm 3 appropriately, these conditions also imply that all iterates along the solution path are sparse. Our proof employs an argument similar to that of [46]. Before stating the main convergence results, we make some further remarks on Assumption 3.1.

- The condition (3.1) states that λ must be sufficiently large to dominate the noise. Such a condition is adequate for sparse recovery applications because recovery performance given in (1.2) achieves optimal error bound under stochastic noise model by picking λ of the order $\|A^T z\|_\infty$ [12, 46, 47, 6, 24, 42, 43]. Moreover, it is also necessary because when λ is smaller than the noise level, the solution $x^\star(\lambda)$ will not be sparse anymore, which defeats the practical purpose of using ℓ_1 regularization.
- The existence of $\tilde{s}$ satisfying condition (3.2) is necessary and standard in sparse recovery analysis. This is closely related to the RIP condition of [10] which assumes that there exist some $s > 0$, and $\nu \in (0,1)$ such that $\kappa(A,s) < (1+\nu)/(1-\nu)$. In fact, if RIP is satisfied with $\nu = 0.1$ at $s > \lceil 45(1+\gamma)\bar{s}\rceil$, then we may take $\gamma_{\mathrm{inc}} = 1.2$ and $\tilde{s} = \lceil 22(1+\gamma)\bar{s}\rceil$ so that condition (3.2) is satisfied. To see this, let $s = \bar{s} + 2\tilde{s}$ and note that

  $$\frac{1+\nu}{1-\nu} > \kappa(A, \bar{s}+2\tilde{s}) \ge \frac{\rho_+(A,\bar{s}+2\tilde{s})}{\rho_-(A,\bar{s}+\tilde{s})} \ge \frac{1.2\rho_+(A,\bar{s}+2\tilde{s}) + \rho_+(A,\tilde{s})\big)}{2.2\,\rho_-(A,\bar{s}+\tilde{s})}.$$

  Therefore we have

  $$\tilde{s} = \lceil 22(1+\gamma)\bar{s}\rceil \ge 17.6\frac{1+\nu}{1-\nu}(1+\gamma)\bar{s} > 8\frac{1.2\rho_+(A,\bar{s}+2\tilde{s}) + \rho_+(A,\tilde{s})}{\rho_-(A,\bar{s}+\tilde{s})}(1+\gamma)\bar{s}.$$

- The RIP condition in the above example looks rather strong, especially when compared with those established in the sparse recovery literature (e.g., [25] and references therein). We note that these results are only concerned about the recovery property of the optimal solution $x^\star(\lambda)$, and it can be expected that stronger conditions (larger constants) are required for maintaining restricted convexity for all intermediate iterates before converging to $x^\star(\lambda)$. In fact, in addition to the matrix A, our RIP-like condition (3.2) also depends on algorithmic parameters γ_{inc} and δ (Theorem 3.2 assumes $\delta < \delta'$). For example, if we choose $\gamma_{\mathrm{inc}} = 2$ (instead of 1.2 in the above calculation), then we need RIP with $\nu = 0.1$ at $s > \lceil 61(1+\gamma)\bar{s}\rceil$ as a sufficient condition. We could also relax the range of δ'. For example, if we allow $\delta' \in (0,1)$ in Assumption 3.1, then the constant in (3.2) needs to be increased from 8 to 16.
- If $L_{\min} > \gamma_{\mathrm{inc}}\rho_+(A, \bar{s}+2\tilde{s})$, then we may simply replace $\gamma_{\mathrm{inc}}\rho_+(A,\bar{s}+2\tilde{s})$ by $L_{\min}$ in the assumption, and all theorem statements hold with $\gamma_{\mathrm{inc}}\rho_+(A,\bar{s}+2\tilde{s})$ replaced by $L_{\min}$. Nevertheless, in practice it is natural to simply pick

  $$L_{\min} = \rho_+(A,1) = \max_{i\in\{1,\dots,n\}} \|A_i\|_2^2,$$

  where A_i is the ith column of A. It automatically satisfies the assumption.

Our first result below concerns the local geometric convergence of Algorithm 2. Basically, if the starting point $x^{(0)}$ is sparse and the optimality condition is satisfied with adequate precision, then all iterates $x^{(k)}$ are sparse, and Algorithm 2 has geometric convergence. (Similar local liner convergence has been established in, e.g., [26, 39], but without specification of the local convergence zone.) To simplify notation, we use a single symbol κ to denote the restricted condition number

$$\kappa = \kappa(A, \bar{s} + 2\tilde{s}) = \frac{\rho_+(A, \bar{s} + 2\tilde{s})}{\rho_-(A, \bar{s} + 2\tilde{s})}. \tag{3.3}$$

THEOREM 3.1. *Suppose Assumption* 3.1 *holds for some* δ', γ, *and* $\tilde{s}$. *If the initial point* $x^{(0)}$ *in Algorithm* 2 *satisfies*

$$\left\|x^{(0)}_{\bar{S}^c}\right\|_0 \le \tilde{s}, \qquad \omega_\lambda(x^{(0)}) \le \delta'\lambda, \tag{3.4}$$

then for all $k \ge 0$, *we have*

$$\left\|x^{(k)}_{\bar{S}^c}\right\|_0 \le \tilde{s}, \qquad \phi_\lambda(x^{(k)}) - \phi^\star_\lambda \;\le\; \left(1 - \frac{1}{4\gamma_{\rm inc}\kappa}\right)^k \left(\phi_\lambda(x^{(0)}) - \phi^\star_\lambda\right),$$

where $\phi^\star_\lambda = \phi_\lambda(x^\star(\lambda)) = \min_x \phi_\lambda(x)$.

Our next result gives the overall iteration complexity of the PGH method. Roughly speaking, if the parameters δ and η are chosen appropriately, then the total number of proximal gradient steps for finding an ϵ-optimal solution is $O(\ln(1/\epsilon))$.

THEOREM 3.2. *Suppose that Assumption* 3.1 *holds for some* δ', γ, *and* $\tilde{s}$, *and the parameters* δ *and* η *in Algorithm* 3 *are chosen such that*

$$\frac{1+\delta}{1+\delta'} \le \eta < 1. \tag{3.5}$$

Let $N = \left\lfloor \ln(\lambda_0/\lambda_{\rm tgt}) \,/\, \ln \eta^{-1} \right\rfloor$ *as in the algorithm. Then:*

1. *Condition* (3.4) *holds for each call of Algorithm* 2. *For* $K = 0, \ldots, N-1$, *the number of iterations in each call of Algorithm* 2 *is no more than*

$$\ln\left(\frac{C}{\delta^2}\right) \Big/ \ln\left(1 - \frac{1}{4\gamma_{\rm inc}\kappa}\right)^{-1},$$

   where $C = 8\gamma_{\rm inc}(1+\kappa)^2(1+\gamma)\kappa\bar{s}$. *Note that this bound is independent of* λ_K.
2. *For* $K = 0, \ldots, N-1$, *the outer-loop iterates* $\hat{x}^{(K)}$ *satisfy*

$$\phi_{\lambda_{\rm tgt}}(\hat{x}^{(K)}) - \phi^\star_{\lambda_{\rm tgt}} \le \eta^{2(K+1)} \, \frac{4.5\,(1+\gamma)\lambda_0^2\bar{s}}{\rho_-(A, \bar{s} + \tilde{s})}, \tag{3.6}$$

   and the following bound on sparse recovery performance holds:

$$\|\hat{x}^{(K)} - \bar{x}\|_2 \le \eta^{K+1} \, \frac{2\lambda_0\sqrt{\bar{s}}}{\rho_-(A, \bar{s} + \tilde{s})}.$$

3. *When Algorithm* 3 *terminates, the total number of iterations is no more than*

$$\left(\frac{\ln(\lambda_0/\lambda_{\rm tgt})}{\ln \eta^{-1}} \, \ln\!\left(\frac{C}{\delta^2}\right) + \ln\max\left(1, \frac{\lambda_{\rm tgt}^2 C}{\epsilon^2}\right)\right) \Big/ \ln\left(1 - \frac{1}{4\gamma_{\rm inc}\kappa}\right)^{-1},$$

   and the output $\hat{x}^{(\rm tgt)}$ *satisfies*

$$\phi_{\lambda_{\rm tgt}}(\hat{x}^{(\rm tgt)}) - \phi^\star_{\lambda_{\rm tgt}} \le \frac{4(1+\gamma)\lambda_{\rm tgt}\bar{s}}{\rho_-(A, \bar{s} + \tilde{s})}\,\epsilon.$$

We have the following remarks regarding these results:

- The precision ϵ in Algorithm 3 is measured against the optimality residue $\omega_\lambda(x)$. In terms of the objective gap, suppose $\epsilon_0 > 0$ is the target precision to be reached. Let
$$K_0 = \left\lceil \frac{1}{2} \ln \left(\frac{4.5\,(1+\gamma)\lambda_0^2 \bar{s}}{\rho_-(A, \bar{s}+\tilde{s})\epsilon_0} \right) \Big/ \ln \eta^{-1} \right\rceil - 1.$$
From the inequality (3.6), we see that if $0 \le K_0 \le N-1$, then for all $K \ge K_0$,
$$\phi_{\lambda_{\text{tgt}}}(\hat{x}^{(K)}) - \phi^\star_{\lambda_{\text{tgt}}} \le \epsilon_0.$$
If we let $\epsilon_0 \to 0$ and run the PGH method forever, then the number of iterations is no more than $O(\ln(\lambda_0/\epsilon_0))$ to achieve an ϵ_0 accuracy in terms of both the objective gap and the optimality residue $\omega_\lambda(\cdot) \le \epsilon_0$. This means that the PGH method achieves a global geometric rate of convergence.
- When the restricted condition number κ is large, we have the approximation
$$\ln \left(1 - \frac{1}{4\gamma_{\text{inc}}\kappa} \right)^{-1} \approx \frac{1}{4\gamma_{\text{inc}}\kappa}.$$
Then the overall iteration complexity can be estimated by $O\left(\kappa \ln\left(\lambda_0/\epsilon\right)\right)$, which is proportional to the restricted condition number κ.
- Even if we solve each stage to high precision with $\hat{\epsilon}_{K+1} = \min(\epsilon, \delta\lambda_{K+1})$, the global convergence rate is still near geometric, and the total number of proximal gradient steps is no more than $O((\ln(\lambda_0/\epsilon))^2)$.

Finally we remark on the relationship between the choice of δ and Assumption 3.1. In Theorem 3.2, we need $\delta < \delta'$ to satisfy condition (3.5). In order to accommodate a larger δ, i.e., to allow less accurate solutions at each stage of Algorithm 3, we can relax the interval for δ' in Assumption 3.1. As discussed in the remarks after Assumption 3.1, this would require a stronger RIP-like condition. On the other hand, using a larger δ leaves the choice for the parameter η to be very close to 1, i.e., we have to reduce the regularization weight λ slowly, which means more homotopy stages.

As we will see from the numerical experiments in section 5, the PGH method often demonstrates best performance (measured by the total number of iterations to obtain a given accuracy) when using relatively large δ and small η, which are unlikely to satisfy our assumptions for geometric convergence at each stage. In fact, with a good warm-start point and a very loose stopping criterion (i.e., $\omega_\lambda(x) \le \delta\lambda$), each intermediate stage requires only a very small number of iterations, even with a sublinear convergence rate. The overall performance of the method hinges on rapidly getting to the linear convergence zone in the final stage, where a significant number of iterations are performed to reach the final high precision. From a practical point of view, while linear convergence in the final stage is critical, it may be too restrictive for the intermediate stages. In particular, using a large η (close to 1) often leads to an unnecessarily large number of iterations before reaching the final stage.

4. Proofs of the convergence results. Our proofs are divided into the following subsections. In section 4.1, we show that under Assumption 3.1, if $x^{(0)}$ is sparse and $\omega_\lambda(x^{(0)})$ is small, then all iterates generated by Algorithm 2 are sparse. In section 4.2, we use the sparsity along the solution path and the restricted eigenvalue condition to show the local geometric convergence of Algorithm 2, thus proving Theorem 3.1. In section 4.3, we show that by setting the parameters δ and η in Algorithm 3 appropriately, we have geometric convergence at each stage of the homotopy method, which leads to the global iteration complexity $O(\log(1/\epsilon))$.

4.1. Sparsity along the solution path. First, we list some useful inequalities that are direct consequences of (3.1) and $\delta' \in (0, 0.2]$:

$$(1-\delta')\lambda - 4\|A^T z\|_\infty > 0, \tag{4.1}$$

$$(1+\delta')\lambda + \|A^T z\|_\infty \le 1.4\lambda, \tag{4.2}$$

$$\lambda + \|A^T z\|_\infty \le (1.4-\delta')\lambda, \tag{4.3}$$

$$\frac{(1+\delta')\lambda + \|A^T z\|_\infty}{(1-\delta')\lambda - \|A^T z\|_\infty} \le \gamma. \tag{4.4}$$

The following result means that if x is sparse and it satisfies an approximate optimality condition for minimizing ϕ_λ, then $\phi_\lambda(x)$ is not much larger than $\phi_\lambda(\bar{x})$.

LEMMA 4.1. *Suppose that Assumption* 3.1 *holds for some δ', γ, and $\tilde{s}$, and $\lambda \ge \lambda_{\mathrm{tgt}}$. If x is sparse, i.e., $\|x_{\bar{S}^c}\|_0 \le \tilde{s}$, and it satisfies the approximate optimality condition*

$$\min_{\xi \in \partial \|x\|_1} \left\| A^T(Ax-b) + \lambda \xi \right\|_\infty \le \delta' \lambda, \tag{4.5}$$

then we have the following inequalities:

$$\|(x-\bar{x})_{\bar{S}^c}\|_1 \le \gamma \|(x-\bar{x})_{\bar{S}}\|_1, \tag{4.6}$$

$$\|x-\bar{x}\|_2 \le \frac{1.4\lambda\sqrt{\bar{s}}}{\rho_-(A, \bar{s}+\tilde{s})}, \tag{4.7}$$

$$\phi_\lambda(x) \le \phi_\lambda(\bar{x}) + \frac{1.4\,\delta'(1+\gamma)\lambda^2 \bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}. \tag{4.8}$$

Proof. Let $\xi \in \partial\|x\|_1$ be a subgradient that achieves the minimum on the left-hand side of (4.5). Then the approximate optimality condition leads to

$$(x-\bar{x})^T \left(A^T(Ax-b) + \lambda\xi \right) \;\le\; \|x-\bar{x}\|_1 \left\| A^T(Ax-b) + \lambda\xi \right\|_\infty \;\le\; \delta'\lambda \|x-\bar{x}\|_1.$$

On the other hand, we can use $b = A\bar{x} + z$ to obtain

$$\begin{aligned}(x-\bar{x})^T \left(A^T(Ax-b) + \lambda\xi \right) &= (x-\bar{x})^T A^T \big(A(x-\bar{x}) - z \big) + \lambda (x-\bar{x})^T \xi \\ &\ge \|A(x-\bar{x})\|_2^2 - \|x-\bar{x}\|_1 \|A^T z\|_\infty + \lambda\, \xi^T (x-\bar{x}).\end{aligned}$$

Next, we break the inner product $\xi^T(x-\bar{x})$ into two parts as

$$\xi^T(x-\bar{x}) = \xi_{\bar{S}}^T (x-\bar{x})_{\bar{S}} + \xi_{\bar{S}^c}^T (x-\bar{x})_{\bar{S}^c}.$$

For the first part, we have (by noticing $\|\xi\|_\infty \le 1$)

$$\xi_{\bar{S}}^T (x-\bar{x})_{\bar{S}} \;\ge\; -\|\xi_{\bar{S}}\|_\infty \|(x-\bar{x})_{\bar{S}}\|_1 \;\ge\; -\|(x-\bar{x})_{\bar{S}}\|_1.$$

For the second part, we use the facts $\bar{x}_{\bar{S}^c} = 0$ and $\xi \in \partial\|x\|_1$ to obtain

$$\xi_{\bar{S}^c}^T (x-\bar{x})_{\bar{S}^c} \;=\; x_{\bar{S}^c}^T \xi_{\bar{S}^c} \;=\; \|x_{\bar{S}^c}\|_1 \;=\; \|(x-\bar{x})_{\bar{S}^c}\|_1.$$

Combining the inequalities above gives

$$\|A(x-\bar{x})\|_2^2 - \|A^T z\|_\infty \|x-\bar{x}\|_1 - \lambda\|(x-\bar{x})_{\bar{S}}\|_1 + \lambda\|(x-\bar{x})_{\bar{S}^c}\|_1 \;\le\; \delta'\lambda\|x-\bar{x}\|_1.$$

Using $\|x-\bar{x}\|_1 = \|(x-\bar{x})_{\bar{S}}\|_1 + \|(x-\bar{x})_{\bar{S}^c}\|_1$ and rearranging terms, we arrive at

$$\begin{aligned} \|A(x-\bar{x})\|_2^2 + \left((1-\delta')\lambda - \|A^T z\|_\infty\right)\|(x-\bar{x})_{\bar{S}^c}\|_1 \\ \leq \left((1+\delta')\lambda + \|A^T z\|_\infty\right)\|(x-\bar{x})_{\bar{S}}\|_1. \end{aligned} \tag{4.9}$$

Next using the inequalities (4.1) and (4.4), we obtain $\|(x-\bar{x})_{\bar{S}^c}\|_1 \leq \gamma\|(x-\bar{x})_{\bar{S}}\|_1$, which is the first desired result in (4.6).

With assumption $\|x_{\bar{S}^c}\|_0 \leq \tilde{s}$, the restricted eigenvalue condition implies

$$\begin{aligned} \rho_-(A, \bar{s}+\tilde{s})\|x-\bar{x}\|_2^2 &\leq \|A(x-\bar{x})\|_2^2 \\ &\leq \left((1+\delta')\lambda + \|A^T z\|_\infty\right)\|(x-\bar{x})_{\bar{S}}\|_1 \\ &\leq 1.4\lambda\|(x-\bar{x})_{\bar{S}}\|_1 \\ &\leq 1.4\lambda\sqrt{\bar{s}}\,\|(x-\bar{x})_{\bar{S}}\|_2 \\ &\leq 1.4\lambda\sqrt{\bar{s}}\,\|x-\bar{x}\|_2, \end{aligned}$$

where the second inequality is a result of (4.9), the third inequality follows from (4.2), and the fourth inequality holds because $|\bar{S}| = \bar{s}$. This proves the second result (4.7).

Finally, since ϕ_λ is convex and $A^T(Ax-b)+\xi$ is a subgradient of ϕ at x, we have

$$\phi_\lambda(x) - \phi_\lambda(\bar{x}) \;\leq\; -\left(A^T(Ax-b)+\xi\right)^T(\bar{x}-x) \;\leq\; \delta'\lambda\|\bar{x}-x\|_1.$$

From the inequality in (4.6), we have

$$\|\bar{x}-x\|_1 \;=\; \|(\bar{x}-x)_{\bar{S}}\|_1 + \|(\bar{x}-x)_{\bar{S}^c}\|_1 \;\leq\; (1+\gamma)\|(\bar{x}-x)_{\bar{S}}\|_1.$$

Therefore,

$$\phi_\lambda(x) - \phi_\lambda(\bar{x}) \;\leq\; \delta'\lambda(1+\gamma)\|(\bar{x}-x)_{\bar{S}}\|_1 \;\leq\; \delta'\lambda(1+\gamma)\sqrt{\bar{s}}\,\|(\bar{x}-x)_{\bar{S}}\|_2,$$

which, together with (4.7), leads to the third desired result. ∎

The next lemma means that if x is sparse, and $\phi_\lambda(x)$ is not much larger than $\phi_\lambda(\bar{x})$, then both $\|x-\bar{x}\|_2$ and $\|x-\bar{x}\|_1$ are small. In fact, similar results hold under the condition $\omega_\lambda(x) \leq \delta'\lambda$ and are proved in Lemma 4.1. However, in the PG method, the optimality residue $\omega_\lambda(x^{(k)})$ may not be monotonic decreasing, but the objective function $\phi_\lambda(x^{(k)})$ is. So in order to establish the desired results for all $x^{(k)}$, we need to show them when the objective gap is sufficiently small.

LEMMA 4.2. *Suppose that Assumption* 3.1 *holds for some* δ', γ, *and* $\tilde{s}$, *and* $\lambda \geq \lambda_{\text{tgt}}$. *Consider* x *such that*

$$\|x_{\bar{S}^c}\|_0 \leq \tilde{s}, \qquad \phi_\lambda(x) \leq \phi_\lambda(\bar{x}) + \frac{1.4\,\delta'(1+\gamma)\lambda^2\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})};$$

then

$$\max\left\{\frac{1}{2.8\lambda}\|A(x-\bar{x})\|_2^2,\ \|x-\bar{x}\|_1\right\} \leq \frac{1.4(1+\gamma)\lambda\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}.$$

Proof. For notational convenience, let

$$\Delta = \frac{1.4\,\delta'(1+\gamma)\lambda^2\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}.$$

We write the assumption $\phi_\lambda(x) \leq \phi_\lambda(\bar{x}) + \Delta$ explicitly as

$$\frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1 \leq \frac{1}{2}\|A\bar{x}-b\|_2^2 + \lambda\|\bar{x}\|_1 + \Delta. \tag{4.10}$$

We can expand the least-squares part in $\phi_\lambda(x)$ as

$$\begin{aligned}\frac{1}{2}\|Ax-b\|_2^2 &= \frac{1}{2}\|(A\bar{x}-b)+A(x-\bar{x})\|_2^2 \\ &\geq \frac{1}{2}\|(A\bar{x}-b)\|_2^2+\frac{1}{2}\|A(x-\bar{x})\|_2^2-\|x-\bar{x}\|_1\|A^T(A\bar{x}-b)\|_\infty.\end{aligned}$$

Plugging the above inequality into (4.10), and noticing $A\bar{x}-b=z$, we obtain

$$\frac{1}{2}\|A(x-\bar{x})\|_2^2-\|x-\bar{x}\|_1\|A^Tz\|_\infty+\lambda\|x\|_1 \;\leq\; \lambda\|\bar{x}\|_1+\Delta.$$

Using the fact $\bar{x}_{\bar{S}^c}=0$, we have

$$\|x\|_1=\|x_{\bar{S}^c}\|_1+\|x_{\bar{S}}\|_1=\|x_{\bar{S}^c}-\bar{x}_{\bar{S}^c}\|_1+\|x_{\bar{S}}\|_1.$$

Therefore

$$\begin{aligned}\frac{1}{2}\|A(x-\bar{x})\|_2^2-\|x-\bar{x}\|_1\|A^Tz\|_\infty+\lambda\|x_{\bar{S}^c}-\bar{x}_{\bar{S}^c}\|_1 &\leq \lambda\left(\|\bar{x}_{\bar{S}}\|_1-\|x_{\bar{S}}\|_1\right)+\Delta \\ &\leq \lambda\,\|\bar{x}_{\bar{S}}-x_{\bar{S}}\|_1+\Delta.\end{aligned}$$

Further splitting $\|x-\bar{x}\|_1$ on the left-hand side as $\|(x-\bar{x})_{\bar{S}}\|_1+\|(x-\bar{x})_{\bar{S}^c}\|_1$, we get

$$\frac{1}{2}\|A(x-\bar{x})\|_2^2+\left(\lambda-\|A^Tz\|_\infty\right)\|(x-\bar{x})_{\bar{S}^c}\|_1 \;\leq\; \left(\lambda+\|A^Tz\|_\infty\right)\|(x-\bar{x})_{\bar{S}}\|_1+\Delta. \tag{4.11}$$

Now there are two possible cases. In the first case, we assume

$$\|x-\bar{x}\|_1 \;\leq\; \frac{\Delta}{\delta'\lambda} \;=\; \frac{1.4(1+\gamma)\lambda\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})}. \tag{4.12}$$

From (4.1), we know that $\left(\lambda-\|A^Tz\|_\infty\right)\|(x-\bar{x})_{\bar{S}^c}\|_1$ is nonnegative, so we can drop it from the left-hand side of (4.11) to obtain

$$\begin{aligned}\frac{1}{2}\|A(x-\bar{x})\|_2^2 &\leq \left(\lambda+\|A^Tz\|_\infty\right)\|(x-\bar{x})_{\bar{S}}\|_1+\Delta \\ &\leq (1.4\lambda-\delta'\lambda)\|(x-\bar{x})_{\bar{S}}\|_1+\Delta \\ &\leq (1.4\lambda-\delta'\lambda)\frac{1.4(1+\gamma)\lambda\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})}+\frac{1.4\delta'(1+\gamma)\lambda^2\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})} \\ &= \frac{1.4^2\lambda(1+\gamma)\lambda\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})},\end{aligned}$$

where in the second inequality we used (4.3) and in the third inequality we used (4.12). This means that the claim of the lemma holds.

In the second case, the assumption in (4.12) does not hold. Then $\Delta<\delta'\lambda\|x-\bar{x}\|_1$ and (4.11) implies

$$\frac{1}{2}\|A(x-\bar{x})\|_2^2+\left(\lambda-\|A^Tz\|_\infty\right)\|(x-\bar{x})_{\bar{S}^c}\|_1 \leq \left(\lambda+\|A^Tz\|_\infty\right)\|(x-\bar{x})_{\bar{S}}\|_1+\delta'\lambda\|x-\bar{x}\|_1.$$

Again we split $\|x-\bar{x}\|_1$ as $\|(x-\bar{x})_{\bar{S}}\|_1+\|(x-\bar{x})_{\bar{S}^c}\|_1$ to obtain

$$\begin{aligned}\frac{1}{2}\|A(x-\bar{x})\|_2^2+\left((1-\delta')\lambda-\|A^Tz\|_\infty\right)\|(x-\bar{x})_{\bar{S}^c}\|_1 \\ \leq \left((1+\delta')\lambda+\|A^Tz\|_\infty\right)\|(x-\bar{x})_{\bar{S}}\|_1.\end{aligned} \tag{4.13}$$

By further using the inequalities (4.1) and (4.4), we get

$$\|(x-\bar{x})_{\bar{S}^c}\|_1 \;\le\; \frac{(1+\delta')\lambda+\|A^Tz\|_\infty}{(1-\delta')\lambda-\|A^Tz\|_\infty}\|(x-\bar{x})_{\bar{S}}\|_1 \;\le\; \gamma\|(x-\bar{x})_{\bar{S}}\|_1. \tag{4.14}$$

This means that if we define

$$\gamma' = \frac{\|(x-\bar{x})_{\bar{S}^c}\|_1}{\sqrt{\bar{s}}\|(x-\bar{x})_{\bar{S}}\|_2},$$

then $\gamma' \le \gamma$ (note that $|\bar{S}| = \bar{s}$). Moreover, we can use the restricted eigenvalue condition and the assumption $\|x_{\bar{S}^C}\|_0 \le \tilde{s}$ to obtain

$$\begin{aligned}
\frac{1}{2}\rho_-(A,\bar{s}+\tilde{s})\|x-\bar{x}\|_2^2 &\le \frac{1}{2}\|A(x-\bar{x})\|_2^2 \\
&\le \left((1+\delta')\lambda+\|A^Tz\|_\infty\right)\left(\|(x-\bar{x})_{\bar{S}}\|_1-\gamma^{-1}\|(x-\bar{x})_{\bar{S}^c}\|_1\right) \\
&\le \left((1+\delta')\lambda+\|A^Tz\|_\infty\right)\sqrt{\bar{s}}(1-\gamma'/\gamma)\|(x-\bar{x})_{\bar{S}}\|_2 \\
&\le 1.4\lambda\sqrt{\bar{s}}(1-\gamma'/\gamma)\,\|(x-\bar{x})_{\bar{S}}\|_2 \\
&\le 1.4\lambda\sqrt{\bar{s}}(1-\gamma'/\gamma)\,\|x-\bar{x}\|_2,
\end{aligned}$$

where the second inequality follows from (4.13) and (4.4), the third inequality holds because of the definition of γ', and the forth inequality follows from (4.2). Hence

$$\|x-\bar{x}\|_2 \le \frac{2\cdot 1.4\lambda\sqrt{\bar{s}}(1-\gamma'/\gamma)}{\rho_-(A,\bar{s}+\tilde{s})}.$$

The above arguments also imply

$$\frac{1}{2}\|A(x-\bar{x})\|_2^2 \;\le\; 1.4\lambda\sqrt{\bar{s}}\,\|x-\bar{x}\|_2 \;\le\; \frac{2\cdot 1.4^2\lambda^2\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})} \;\le\; \frac{1.4^2(1+\gamma)\lambda^2\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})},$$

where the last inequality is due to $\gamma > 1$. Finally, using the definition of γ', we get

$$\|x-\bar{x}\|_1 \;\le\; (1+\gamma')\sqrt{\bar{s}}\,\|(x-\bar{x})_{\bar{S}}\|_2 \;\le\; \frac{2\cdot 1.4(1+\gamma')(1-\gamma'/\gamma)\lambda\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})} \;\le\; \frac{1.4(1+\gamma)\lambda\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})},$$

where the last inequality follows by maximizing over γ' achieved at $\gamma' = (\gamma-1)/2$. These prove the desired bound. ∎

The following lemma means that if x is sparse and $\phi_\lambda(x)$ is not much larger than $\phi_\lambda(\bar{x})$, then $T_{\lambda,L}(x)$ is sparse.

LEMMA 4.3. *Suppose that Assumption* 3.1 *holds for some* δ', γ, *and* $\tilde{s}$, *and* $\lambda \ge \lambda_{\mathrm{tgt}}$. *If* x *satisfies*

$$\|x_{\bar{S}^c}\|_0 \le \tilde{s}, \qquad \phi_\lambda(x) \le \phi_\lambda(\bar{x}) + \frac{1.4\,\delta'(1+\gamma)\lambda^2\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})} \tag{4.15}$$

and $L < \gamma_{\mathrm{inc}}\rho_+(A,\bar{s}+2\tilde{s})$, *then*

$$\left\|\left(T_{\lambda,L}(x)\right)_{\bar{S}^c}\right\|_0 < \tilde{s}.$$

Proof. Recall that $T_{\lambda,L}$ can be computed by the soft-thresholding operator, i.e.,

$$(T_L(x))_i = \mathrm{sgn}(\tilde{x}_i)\max\left\{|\tilde{x}_i| - \frac{\lambda}{L},\ 0\right\}, \qquad i = 1,\dots,n,$$

where

$$\tilde{x} = x - \frac{1}{L}A^T(Ax - b) = x - \frac{1}{L}A^TA(x - \bar{x}) + \frac{1}{L}A^Tz.$$

In order to upper bound the number of nonzero elements in $(T_L(x))_{\bar{S}^c}$, we split the truncation threshold λ/L on elements of $\tilde{x}_{\bar{S}^c}$ into three parts:

- $0.175\lambda/L$ on elements of $x_{\bar{S}^c}$,
- $0.125\lambda/L$ on elements of $(1/L)A^Tz$, and
- $0.7\lambda/L$ on elements of $(1/L)A^TA(x - \bar{x})$.

By assumption (3.1), we have $\|A^Tz\|_\infty \le \lambda/8$; hence

$$\big|\{j : ((1/L)A^Tz)_j > 0.125\lambda/L\}\big| = 0.$$

Therefore,

$$\big\|\big(T_L(x)\big)_{\bar{S}^c}\big\|_0 \;\le\; \big|\big\{j \in \bar{S}^c : |x_j| > 0.175\lambda/L\big\}\big| + \big|\big\{j : \big|\big(A^TA(x - \bar{x})\big)_j\big| \ge 0.7\lambda\big\}\big|.$$

Note that

$$\begin{aligned}\big|\{j \in \bar{S}^c : |x_j| \ge 0.175\lambda/L\}\big| &= \big|\{j \in \bar{S}^c : |(x - \bar{x})_j| \ge 0.175\lambda/L\}\big| \\ &\le \big|\{j : |(x - \bar{x})_j| \ge 0.175\lambda/L\}\big| \\ &\le L(0.175\lambda)^{-1}\|x - \bar{x}\|_1 \\ &\le \frac{L}{0.175\lambda}\frac{1.4(1+\gamma)\lambda\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})} = \frac{8L(1+\gamma)\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})},\end{aligned} \tag{4.16}$$

where the last inequality follows from Lemma 4.2.

For the last part, consider S' with maximum size $s' = |S'| \le \tilde{s}$ such that

$$S' \subset \{j : |(A^TA(x - \bar{x}))_j| \ge 0.7\lambda\}.$$

Then there exists u such that $\|u\|_\infty = 1$ and $\|u\|_0 = s'$, and $0.7s'\lambda \le u^TA^TA(x - \bar{x})$. Moreover,

$$0.7s'\lambda \;\le\; u^TA^TA(x-\bar{x}) \;\le\; \|Au\|_2\|A(x-\bar{x})\|_2 \;\le\; \sqrt{\rho_+(A, s')}\sqrt{s'}\sqrt{\frac{2\cdot 1.4^2(1+\gamma)\lambda^2\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}},$$

where the last inequality again follows from Lemma 4.2. Taking squares of both sides of the above inequality gives

$$s' \le \frac{8\,\rho_+(A, s')(1+\gamma)\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})} \le \frac{8\,\rho_+(A, \tilde{s})(1+\gamma)\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})} < \tilde{s},$$

where the last inequality is due to (3.2). Since $s' = |S'|$ achieves the maximum possible value such that $s' \le \tilde{s}$ for any subset S' of $\{j : |(A^TA(x - \bar{x}))_j| \ge 0.7\lambda\}$, and the above inequality shows that $s' < \tilde{s}$, we must have

$$S' = \{j : |(A^T A(x - \bar{x}))_j| \geq 0.7\lambda\},$$

and thus

$$\big|\{j : |(A^T A(x - \bar{x}))_j| \geq 0.7\lambda\}\big| = s' \leq \left\lfloor \frac{8\,\rho_+(A, \tilde{s})(1+\gamma)\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})} \right\rfloor.$$

Finally, combining the above bound with the bound in (4.16) gives

$$\big\|\big(T_{\lambda,L}(x)\big)_{\bar{S}^c}\big\|_0 \;\leq\; \frac{8\,(L + \rho_+(A, \tilde{s}))}{\rho_-(A, \bar{s}+\tilde{s})}(1+\gamma)\bar{s}.$$

Under the assumption $L < \gamma_{\text{inc}}\rho_+(A, \bar{s}+2\tilde{s})$ and (3.2), the right-hand side of the above inequality is less than $\tilde{s}$. This proves the desired result. ∎

Recall that each iteration of Algorithm 2 takes the form $x^{(k+1)} = T_{\lambda,M_k}(x^{(k)})$. According to (2.5), the objective value $\phi_\lambda(x^{(k)})$ is monotone decreasing. So if $x^{(0)}$ satisfies the condition (4.15), so does every iterate $x^{(k)}$. In order to show

$$\|(x^{(k)})_{\bar{S}^c}\|_0 < \tilde{s} \quad \forall\, k > 0,$$

we only need to note that the line search in Algorithm 1 always terminates with

$$M_k \leq \gamma_{\text{inc}}\rho_+(A, \bar{s}+2\tilde{s}). \tag{4.17}$$

Indeed, as long as

$$M_k \in [\rho_+(A, \bar{s}+2\tilde{s}),\ \gamma_{\text{inc}}\rho_+(A, \bar{s}+2\tilde{s})],$$

Lemma 4.3 implies that $\big\|\big(T_{\lambda,L}(x)\big)_{\bar{S}^c}\big\|_0 < \tilde{s}$ and the restricted smoothness property (2.8) implies the termination of line search.

4.2. Proof of Theorem 3.1. In this subsection, we show that for any fixed λ, the sequence $\{x^{(k)}\}_{k=0}^{\infty}$ generated by Algorithm 2 (without invoking the stopping criteria) has a limit and the local rate of convergence is geometric.

First, since the sublevel set $\{x : \phi_\lambda(x) \leq \phi_\lambda(x^{(0)})\}$ is bounded and $\phi_\lambda(x^{(k)})$ is monotone decreasing, the sequence $\{x^{(k)}\}_{k=0}^{\infty}$ is bounded. By the Bolzano–Weierstrass theorem, it has a convergent subsequence and a corresponding accumulation point. Moreover, from (2.5) and the fact that $\phi_\lambda(x)$ is bounded below, we conclude that

$$\lim_{k\to\infty} \|g_{\lambda,L}(x^{(k)})\|_2 = 0.$$

By Lemma 2.2, this implies that any accumulation point of the sequence $\{x^{(k)}\}_{k=0}^{\infty}$ satisfies the optimality condition and therefore is a minimizer of ϕ_λ.

Let $x^\star(\lambda)$ denote an accumulation point of the sequence $\{x^{(k)}\}_{k=0}^{\infty}$. By Lemma 4.3, any accumulation point is also sparse. In particular, we have $\|(x^\star(\lambda))_{\bar{S}^c}\|_0 \leq \tilde{s}$.

Now using the restricted strong convexity property (2.9), we have

$$f(x) \geq f(x^\star) + \langle \nabla f(x^\star(\lambda)), x - x^\star(\lambda)\rangle + \frac{\rho_-(A, \bar{s}+2\tilde{s})}{2}\|x - x^\star(\lambda)\|_2^2. \tag{4.18}$$

Since $x^\star(\lambda) = \arg\min_x\{f(x) + \lambda\|x\|_1\}$, there must exist $\xi \in \partial\|x^\star(\lambda)\|_1$ such that

$$\nabla f(x^\star(\lambda)) + \lambda\xi = 0. \tag{4.19}$$

Since $\xi \in \partial\|x^\star(\lambda)\|_1$, we also have (by convexity of $\lambda\|\cdot\|_1$)

$$\lambda\|x\|_1 \geq \lambda\|x^\star(\lambda)\|_1 + \langle \lambda\xi, x - x^\star(\lambda)\rangle. \tag{4.20}$$

Adding the two inequalities (4.18) and (4.20) and using (4.19), we get

$$\phi_\lambda(x) - \phi_\lambda(x^\star(\lambda)) \geq \frac{\rho_-(A, \bar{s}+2\tilde{s})}{2}\|x - x^\star(\lambda)\|_2^2 \qquad \forall\, x : \|x_{\bar{S}^c}\|_0 \leq \tilde{s}. \tag{4.21}$$

Since any accumulation point satisfies $\|x_{\bar{S}^c}\|_0 \leq \tilde{s}$, we conclude that $x^\star(\lambda)$ is a unique accumulation point, in other words, the limit, of the sequence $\{x^{(k)}\}_{k=0}^{\infty}$.

Next we show that under the assumptions in Lemma 4.3, especially with $x^{(0)}$ satisfying (4.15), Algorithm 2 has a geometric convergence rate. We start with the stopping criteria in the line search procedure:

$$\begin{aligned}
\phi_\lambda(x^{(k+1)}) &\leq \psi_{\lambda,M_k}(x^{(k)}, x^{(k+1)}) \\
&\leq \min_x \left\{ f(x) + \frac{M_k}{2}\|x - x^{(k)}\|_2^2 + \lambda\|x\|_1 \right\} \\
&= \min_x \left\{ \phi_\lambda(x) + \frac{M_k}{2}\|x - x^{(k)}\|_2^2 \right\},
\end{aligned}$$

where the second inequality follows from the convexity of f. We can further relax the right-hand side of the above inequality by restricting the minimization over the line segment $x = \alpha x^\star(\lambda) + (1-\alpha)x^{(k)}$, where $\alpha \in [0,1]$. This leads to

$$\begin{aligned}
\phi_\lambda(x^{(k+1)}) &\leq \min_\alpha \left\{ \phi_\lambda\big(\alpha x^\star(\lambda) + (1-\alpha)x^{(k)}\big) + \frac{M_k}{2}\|\alpha(x^{(k)} - x^\star(\lambda))\|_2^2 \right\} \\
&= \min_\alpha \left\{ \phi_\lambda(x^{(k)}) - \alpha\big(\phi_\lambda(x^{(k)}) - \phi_\lambda(x^\star(\lambda))\big) + \frac{\alpha^2 M_k}{2}\|x^{(k)} - x^\star(\lambda)\|_2^2 \right\}.
\end{aligned}$$

Since the conclusion of Lemma 4.3 implies that $\|x^{(k)}_{\bar{S}^c}\|_0 \leq \tilde{s}$ for all $k \geq 0$, we can use the "restricted" strong convexity property (4.21) to obtain

$$\phi_\lambda(x^{(k+1)}) \leq \min_\alpha \left\{ \phi_\lambda(x^{(k)}) - \alpha\left(1 - \frac{\alpha M_k}{\rho_-(A, \bar{s}+2\tilde{s})}\right)\Big(\phi_\lambda(x^{(k)}) - \phi_\lambda(x^\star(\lambda))\Big) \right\}.$$

The minimizing value is $\alpha = \rho_-(A, \bar{s}+2\tilde{s})/(2M_k)$, which gives

$$\phi_\lambda(x^{(k+1)}) \;\leq\; \phi_\lambda(x^{(k)}) - \frac{\rho_-(A, \bar{s}+2\tilde{s})}{4M_k}\Big(\phi_\lambda(x^{(k)}) - \phi_\lambda(x^\star(\lambda))\Big).$$

Let $\phi_\lambda^\star = \phi_\lambda(x^\star(\lambda))$. Subtracting $\phi_\lambda^\star$ from both side of the above inequality gives

$$\begin{aligned}
\phi_\lambda(x^{(k+1)}) - \phi_\lambda^\star &\leq \left(1 - \frac{\rho_-(A, \bar{s}+2\tilde{s})}{4M_k}\right)\Big(\phi_\lambda(x^{(k)}) - \phi_\lambda^\star\Big) \\
&\leq \left(1 - \frac{\rho_-(A, \bar{s}+2\tilde{s})}{4\gamma_{\mathrm{inc}}\rho_+(A, \bar{s}+2\tilde{s})}\right)\Big(\phi_\lambda(x^{(k)}) - \phi_\lambda^\star\Big),
\end{aligned}$$

where the second inequality follows from (4.17). Therefore, we have

$$\phi_\lambda(x^{(k)}) - \phi_\lambda^\star \;\leq\; \left(1 - \frac{1}{4\gamma_{\mathrm{inc}}\kappa}\right)^k \Big(\phi_\lambda(x^{(0)}) - \phi_\lambda^\star\Big),$$

where κ is the restricted condition number defined in (3.3). Note that the above convergence rate does not depend on λ. This completes the proof of Theorem 3.1.

4.3. Proof of Theorem 3.2. In Algorithm 3, $\hat{x}^{(K)}$ denotes an approximate solution for minimizing the function ϕ_{λ_K}. A key idea of the homotopy method is to use $\hat{x}^{(K)}$ as the starting point in the PG method for minimizing the next function $\phi_{\lambda_{K+1}}$. The following lemma shows that if we choose the parameters δ and η appropriately, then $\hat{x}^{(K)}$ satisfies the approximate optimality condition for λ_{K+1} that guarantees local geometric convergence.

LEMMA 4.4. *Suppose $\hat{x}^{(K)}$ satisfies the approximate optimality condition*

$$\omega_{\lambda_K}(\hat{x}^{(K)}) \le \delta\lambda_K$$

for some $\delta < \delta'$. Let $\lambda_{K+1} = \eta\lambda_K$ for some η that satisfies

$$\frac{1+\delta}{1+\delta'} \le \eta < 1. \tag{4.22}$$

Then we have

$$\omega_{\lambda_{K+1}}(\hat{x}^{(K)}) \le \delta'\lambda_{K+1}.$$

Proof. If $\omega_{\lambda_K}(\hat{x}^{(K)}) \le \delta\lambda_K$, then there exists $\xi \in \partial\|\hat{x}^{(K)}\|_1$ such that

$$\|\nabla f(\hat{x}^{(K)}) + \lambda_K\xi\|_\infty \le \delta\lambda_K.$$

Then we have

$$\begin{aligned}
\omega_{\lambda_{K+1}}(\hat{x}^{(K)}) &\le \|\nabla f(\hat{x}^{(K)}) + \lambda_{K+1}\xi\|_\infty \\
&= \|\nabla f(\hat{x}^{(K)}) + \lambda_K\xi + (\lambda_{K+1} - \lambda_K)\xi\|_\infty \\
&\le \|\nabla f(\hat{x}^{(K)}) + \lambda_K\xi\|_\infty + |\lambda_{K+1} - \lambda_K| \cdot \|\xi\|_\infty \\
&\le \delta\lambda_K + (1-\eta)\lambda_K.
\end{aligned}$$

Since (4.22) implies $\delta\lambda_K + (1-\eta)\lambda_K \le \delta'\lambda_{K+1}$, we have the desired result. ∎

LEMMA 4.5. *Suppose that Assumption 3.1 holds for some δ', γ, and $\tilde{s}$. Let $\lambda \ge \lambda_{\mathrm{tgt}}$, and assume that x satisfies*

$$\omega_\lambda(x) \le \delta'\lambda.$$

Then for all $\lambda' \in [\lambda_{\mathrm{tgt}}, \lambda]$, we have

$$\phi_{\lambda'}(x) - \phi_{\lambda'}(x^\star(\lambda')) \le \frac{2(1+\gamma)(\lambda+\lambda')(\omega_\lambda(x) + \lambda - \lambda')\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}.$$

Proof. Let $\xi(\lambda) = \arg\min_{\xi \in \partial\|x\|_1} \|\nabla f(x) + \lambda\xi\|_\infty$. Thus $\omega_\lambda(x) = \|\nabla f(x) + \lambda\xi(\lambda)\|_\infty$. By the convexity of $\phi_{\lambda'}$, we have

$$\begin{aligned}
\phi_{\lambda'}(x) - \phi_{\lambda'}(x^\star(\lambda')) &\le \langle \nabla f(x) + \lambda'\xi(\lambda), x - x^\star(\lambda')\rangle \\
&\le (\|\nabla f(x) + \lambda\xi(\lambda)\|_\infty + \lambda - \lambda')\|x - x^\star(\lambda')\|_1 \\
&= (\omega_\lambda(x) + \lambda - \lambda')\,\|x - x^\star(\lambda')\|_1. \qquad (4.23)
\end{aligned}$$

Since $\omega_{\lambda'}(x^\star(\lambda')) = 0 < \delta'\lambda'$, by Lemma 4.1, we have

$$\|x^\star(\lambda') - \bar{x}\|_1 \le (1+\gamma)\sqrt{\bar{s}}\,\|x^\star(\lambda') - \bar{x}\|_2 \le \frac{2(1+\gamma)\lambda'\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}.$$

Similarly, because of the assumption $\omega_\lambda(x) \le \delta'\lambda$, we have

$$\|x - \bar{x}\|_1 \le (1+\gamma)\sqrt{\bar{s}}\,\|x - \bar{x}\|_2 \le \frac{2(1+\gamma)\lambda\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}.$$

Therefore, we have

$$\|x - x^\star(\lambda')\|_1 \le \|x - \bar{x}\|_1 + \|\bar{x} - x^\star(\lambda')\|_1 \le \frac{2(1+\gamma)(\lambda+\lambda')\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}.$$

Now we obtain from (4.23) that

$$\phi_{\lambda'}(x) - \phi_{\lambda'}(x^\star(\lambda')) \;\le\; \frac{2(1+\gamma)(\lambda+\lambda')(\omega_\lambda(x) + \lambda - \lambda')\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}.$$

This proves the desired result. □

Now we are ready to estimate the overall complexity of the PGH method. First, we need to bound the number of iterations within each call of Algorithm 2.

Using Lemma 2.2, we can upper bound the optimality residue as

$$\begin{aligned}
\omega_\lambda(x^{(k+1)}) &\le \left(1 + \frac{S_{M_k}(x^{(k)})}{M_k}\right) \left\|g_{\lambda,M_k}(x^{(k)})\right\|_2 \\
&\le \left(1 + \frac{\rho_+(A, \bar{s}+2\tilde{s})}{\rho_-(A, \bar{s}+2\tilde{s})}\right) \left\|g_{\lambda,M_k}(x^{(k)})\right\|_2 \\
&= (1+\kappa)\left\|g_{\lambda,M_k}(x^{(k)})\right\|_2,
\end{aligned}$$

where the second inequality follows from

$$S_{M_k}(x^{(k)}) \le \rho_+(A, \bar{s}+2\tilde{s}), \qquad M_k \ge \rho_-(A, \bar{s}+2\tilde{s}),$$

which are direct consequences of the line search termination criterion, the restricted smoothness property (2.8), and the restricted strong convexity property (2.9).

To bound the norm of $g_{\lambda,M_k}(x^{(k)})$, we use (2.5) and Theorem 3.1 to obtain

$$\begin{aligned}
\left\|g_{\lambda,M_k}(x^{(k)})\right\|_2^2 &\le 2M_k\left(\phi_\lambda(x^{(k)}) - \phi_\lambda(x^{(k+1)})\right) \\
&\le 2M_k\left(\phi_\lambda(x^{(k)}) - \phi_\lambda^\star\right) \\
&\le 2\gamma_{\mathrm{inc}}\,\rho_+(A, \bar{s}+2\tilde{s})\left(1 - \frac{1}{4\gamma_{\mathrm{inc}}\kappa}\right)^k \left(\phi_\lambda(x^{(0)}) - \phi_\lambda^\star\right),
\end{aligned}$$

where $\phi_\lambda^\star = \phi_\lambda(x^\star(\lambda)) = \min_x \phi_\lambda(x)$, and the last inequality is due to (4.17). Therefore, in order to satisfy the stopping criteria

$$\omega_\lambda(x^{(k+1)}) \;\le\; \delta\lambda,$$

it suffices to ensure

$$(1+\kappa)\sqrt{2\gamma_{\mathrm{inc}}\rho_+(A, \bar{s}+2\tilde{s})\left(1 - \frac{1}{4\gamma_{\mathrm{inc}}\kappa}\right)^k \left(\phi_\lambda(x^{(0)}) - \phi_\lambda^\star\right)} \;\le\; \delta\lambda,$$

which requires

$$k \;\ge\; \ln\left(\frac{2\gamma_{\mathrm{inc}}(1+\kappa)^2\rho_+(A, \bar{s}+2\tilde{s})}{\delta^2\lambda^2}\left(\phi_\lambda(x^{(0)}) - \phi_\lambda^\star\right)\right) \Big/ \ln\left(1 - \frac{1}{4\gamma_{\mathrm{inc}}\kappa}\right)^{-1}.$$

We still need to bound the gap $\phi_\lambda(x^{(0)}) - \phi_\lambda^\star$. Since Lemma 4.4 implies that $\omega_\lambda(x^{(0)}) \le \delta'\lambda$, we can obtain the following inequality directly from Lemma 4.5 by setting $\lambda' = \lambda$ and $x = x^{(0)}$:

$$\phi_\lambda(x^{(0)}) - \phi_\lambda^\star \le \frac{4(1+\gamma)\lambda^2\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}.$$

Therefore, the number of iterations in each call of Algorithm 2 is no more than

$$\ln\left(\frac{8\gamma_{\rm inc}(1+\kappa)^2(1+\gamma)\bar{s}}{\delta^2}\frac{\rho_+(A,\bar{s}+2\tilde{s})}{\rho_-(A,\bar{s}+\tilde{s})}\right) \Big/ \ln\left(1-\frac{1}{4\gamma_{\rm inc}\kappa}\right)^{-1}.$$

To simplify presentation, we note that

$$C \;=\; 8\gamma_{\rm inc}(1+\kappa)^2(1+\gamma)\bar{s}\kappa \;\ge\; 8\gamma_{\rm inc}(1+\kappa)^2(1+\gamma)\bar{s}\frac{\rho_+(A,\bar{s}+2\tilde{s})}{\rho_-(A,\bar{s}+\tilde{s})}.$$

Thus the previous iteration bound is no more than

$$\ln\left(\frac{C}{\delta^2}\right) \Big/ \ln\left(1-\frac{1}{4\gamma_{\rm inc}\kappa}\right)^{-1}.$$

This proves part 1 of Theorem 3.2. We note that this bound is independent of λ.

In the PGH method (Algorithm 3), after K outer iterations for $K \le N-1$, we have from Lemma 4.4 that $\omega_{\lambda_{K+1}}(\hat{x}^{(K)}) \le \delta'\lambda_{K+1}$. The sparse recovery performance bound

$$\|\hat{x}^{(K)} - \bar{x}\|_2 \le 2\eta^{K+1}\lambda_0\sqrt{\bar{s}}/\rho_-(A,\bar{s}+\tilde{s})$$

follows directly from Lemma 4.1 and $\lambda_{K+1} = \eta^{K+1}\lambda_0$. Moreover, from Lemma 4.5 with $\lambda' = \lambda_{\rm tgt}$, $\lambda = \lambda_{K+1}$, and $x = \hat{x}^{(K)}$, we obtain

$$\phi_{\lambda_{\rm tgt}}(\hat{x}^{(K)}) - \phi_{\lambda_{\rm tgt}}^\star \le \frac{2(1+\gamma)(\lambda_{K+1}+\lambda_{\rm tgt})(\delta'\lambda_{K+1}+\lambda_{K+1}-\lambda_{\rm tgt})\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})}.$$

Next, we use $\delta' < 1$ and maximize $(\lambda_{K+1}+\lambda_{\rm tgt})(2\lambda_{K+1}-\lambda_{\rm tgt})$ over $\lambda_{\rm tgt}$ to obtain

$$\phi_{\lambda_{\rm tgt}}(\hat{x}^{(K)}) - \phi_{\lambda_{\rm tgt}}^\star \le \frac{4.5(1+\gamma)\lambda_{K+1}^2\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})} = \eta^{2(K+1)}\frac{4.5(1+\gamma)\lambda_0^2\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})}.$$

This proves part 2 of Theorem 3.2.

In Algorithm 3, the number of outer iterations, excluding the last one for $\lambda_{\rm tgt}$, is

$$N = \left\lfloor \frac{\ln(\lambda_0/\lambda_{\rm tgt})}{\ln(1/\eta)} \right\rfloor.$$

The last iteration for $\lambda_{\rm tgt}$ uses an absolute precision ϵ instead of the relative precision $\delta\lambda_{\rm tgt}$. Therefore, the overall complexity is bounded by

$$\left(\frac{\ln(\lambda_0/\lambda_{\rm tgt})}{\ln(1/\eta)}\ln\left(\frac{C}{\delta^2}\right) + \ln\max\left(1, \frac{\lambda_{\rm tgt}^2 C}{\epsilon^2}\right)\right) \Big/ \ln\left(1-\frac{1}{4\gamma_{\rm inc}\kappa}\right)^{-1}.$$

Finally, when the PGH method terminates, we have $\omega_{\lambda_{\rm tgt}}(\hat{x}^{(\rm tgt)}) \le \epsilon$. Therefore we can apply Lemma 4.5 with $\lambda = \lambda' = \lambda_{\rm tgt}$ and $x = \hat{x}^{(\rm tgt)}$ to obtain the last result in part 3.

5. Numerical experiments. In this section, we present numerical experiments to support our theoretical analysis. For comparison purposes, we implemented the following methods for solving the ℓ_1-LS problem:

- PG: the PG method with adaptive line search (Algorithm 2).
- PGH: our proposed PGH method described in Algorithm 3.
- ADG: Nesterov's accelerated dual gradient method, i.e., [32, algorithm (4.9)].
- ADGH: the PGH method in Algorithm 3, but with PG replaced by ADG.

In section 5.1, we first demonstrate the numerical properties of PGH by comparing it with other methods listed above and then investigate the effects of varying the homotopy parameters δ and η. In section 5.2, we compare it with two very similar implementations of approximate homotopy method: SpaRSA [45] and FPC [22].

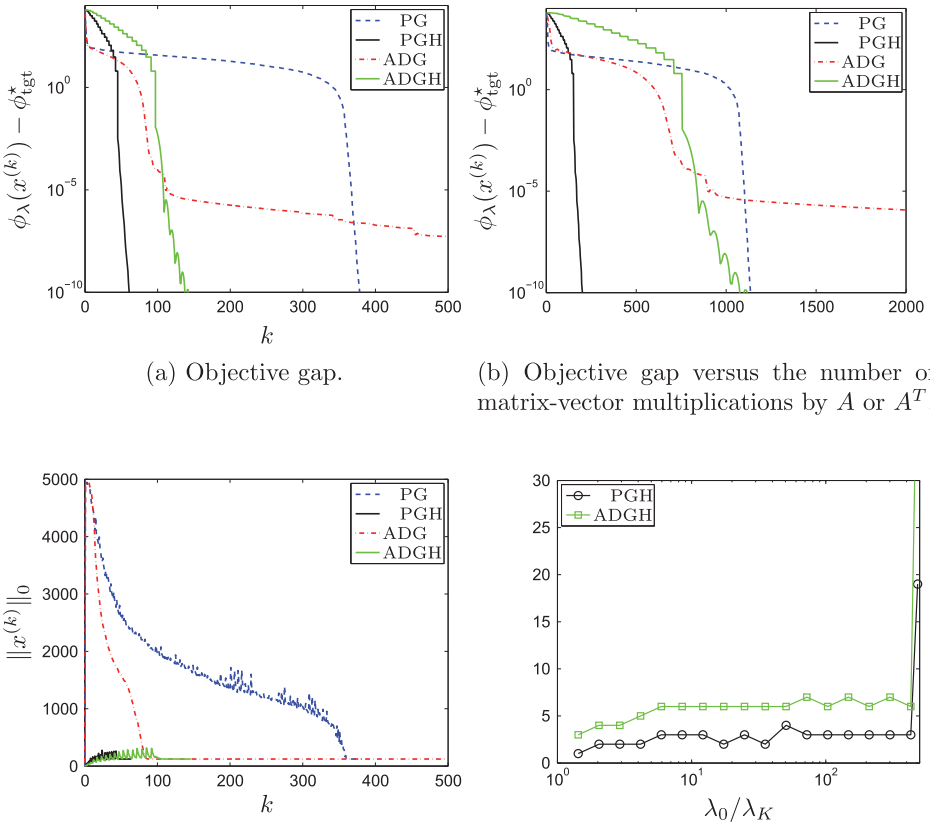
5.1. Numerical properties of PGH. We generated a random instance of (1.1) with dimensions $m = 1000$ and $n = 5000$. The entries of the matrix $A \in \mathbb{R}^{m\times n}$ are generated independently with the uniform distribution over the interval $[-1, +1]$. The vector $\bar{x} \in \mathbb{R}^n$ was generated with the same distribution at 100 randomly chosen coordinates (i.e., $\bar{s} = |\operatorname{supp}(\bar{x})| = 100$). The noise $z \in \mathbb{R}^m$ is a dense vector with independent random entries with the uniform distribution over the interval $[-\sigma, \sigma]$, where σ is the noise magnitude. Finally the vector b was obtained as $b = A\bar{x} + z$. In our experiment, we set $\sigma = 0.01$ and choose $\lambda_{\mathrm{tgt}} = 1$. For this instance we have roughly $\|A^T z\|_\infty = 0.411$. To start the PGH method, we have $\lambda_0 = \|A^T b\|_\infty = 483.4$.

Figure 5.1 illustrates various numerical properties of the four different methods for solving this random instance. We used the parameters $\gamma_{\mathrm{inc}} = 2$ and $\gamma_{\mathrm{dec}} = 2$ in all four methods. For the two homotopy methods (whose acronyms end with the letter H), we used the parameters $\eta = 0.7$ and $\delta = 0.2$. In Figures 5.1(a) and 5.1(b), the horizontal axes show the cumulative count of proximal gradient iterations. For the two homotopy methods, the vertical line segments in Figures 5.1(a) and 5.1(b) indicate switchings of homotopy stages (when the value of λ is reduced by the factor η)—they reflect the jump of objective function for the same vector $x^{(k)}$.

Figure 5.1(a) shows the objective gap $\phi_\lambda(x^{(k)}) - \phi^\star_{\mathrm{tgt}}$ versus the total number of iterations k. The PG method solves the problem with the target regularization parameter λ_{tgt} directly. For the first 350 or so iterations, it demonstrated a slow sublinear convergence rate (theoretically $O(1/k)$) but converged rapidly for the last 30 iterations with a linear rate. Referring to Figure 5.1(c), we see that the slow phase of PG is associated with relatively dense iterates (with $\|x^{(k)}\|_0$ ranging from 5,000 to several hundred), while the fast linear convergence in the end coincides with sparse iterates with $\|x^{(k)}\|_0$ around 100. In contrast, all iterates in the PGH method are very sparse (always less than 300), and it converges much faster.

Also plotted in Figure 5.1 are numerical characteristics of the ADG and ADGH methods. We see that the ADG method is much faster than the PG method in the early phase, which can be explained by its better convergence rate, i.e., $O(1/k^2)$ instead of $O(1/k)$ for PG. However, it stays with the sublinear rate even when the iterates $x^{(k)}$ becomes very sparse. The reason is that ADG cannot automatically exploit the local strong convexity as PG does, so it eventually lagged behind when the iterates became very sparse (see discussions in [32]). The ADGH method combines the homotopy strategy with the ADG method. It is much faster than ADG but still does not have linear convergence and thus is much slower than the PGH method.

Figure 5.1(d) shows the number of proximal gradient steps performed at each stage (corresponding to each λ_K) of the two homotopy methods. We see that the final stage of the PGH method took 19 inner iterations to reach the absolute precision $\epsilon = 10^{-5}$,



(a) Objective gap.

(b) Objective gap versus the number of matrix-vector multiplications by A or A^T.

(c) Sparsity along solution path.

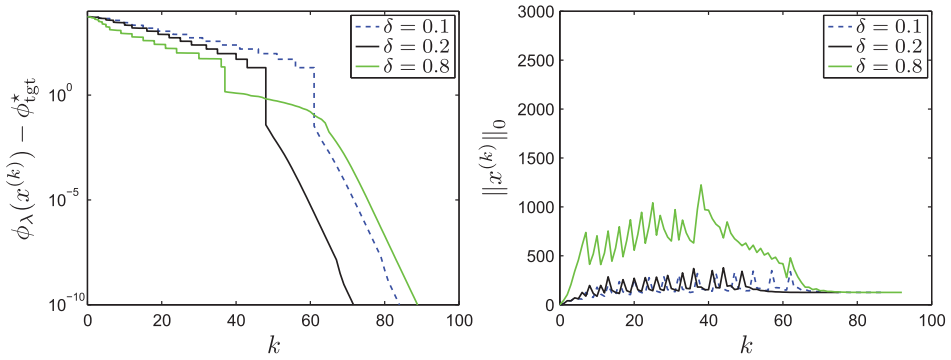
(d) Number of iterations for each λ_K.

FIG. 5.1. *Solving a random instance of the ℓ_1-LS problem.*

and all earlier stages took only 1 to 4 inner iterations to reach the relative precision $\delta\lambda_K$. We note that the number of inner iterations at each intermediate stage stayed relatively constant, even though the tolerance for the optimality residue decreases as $\delta\lambda_k = \eta^K \delta\lambda_0$. This is predicted by part 1 of Theorem 3.2. The ADGH method took more inner iterations at each stage.

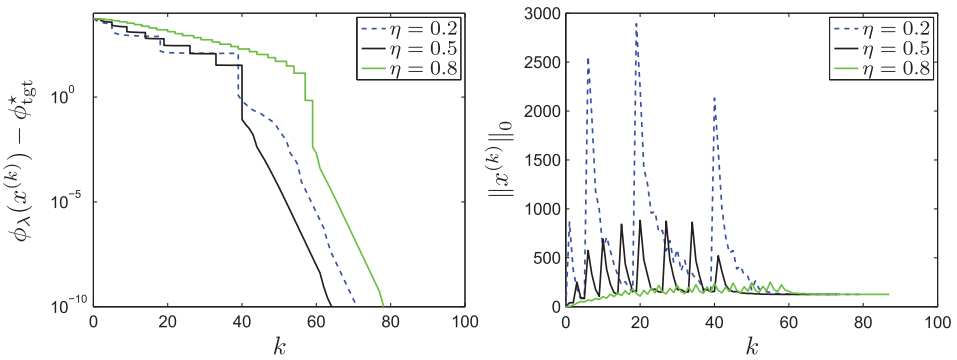
Figure 5.1(b) shows the objective gap versus the total number of matrix-vector multiplications with either A or A^T. Evaluating the objective function $f(x^{(k)})$ costs one matrix-vector multiplication, and evaluating the gradient $\nabla f(x^{(k)})$ costs an additional multiplication. The estimate in (2.7) states that each step in the PG method needs on average two calls of the oracle. But one of them is done in the line search procedure, and it requires only the function value. Therefore each inner iteration on average costs roughly three matrix-vector multiplications. On the other hand, each iteration of the ADG method on average costs eight matrix-vector multiplications [32]. These factors are confirmed by comparing the horizontal scales of Figures 5.1(a) and 5.1(b). We found that the number of matrix-vector multiplications is a very precise indicator for the running time of each algorithm. From this perspective, the advantage of the PGH method is more pronounced.

Next we conducted experiments to test the sensitivity of the PGH method with respect to the choices of parameters δ and η. Figure 5.2 shows the objective gap and sparsity of the iterates along the solution path for different δ while keeping $\eta = 0.7$.



(a) Objective gap. (b) Sparsity along solution path.

FIG. 5.2. *Performance of the PGH method by varying δ while keeping $\eta = 0.7$.*



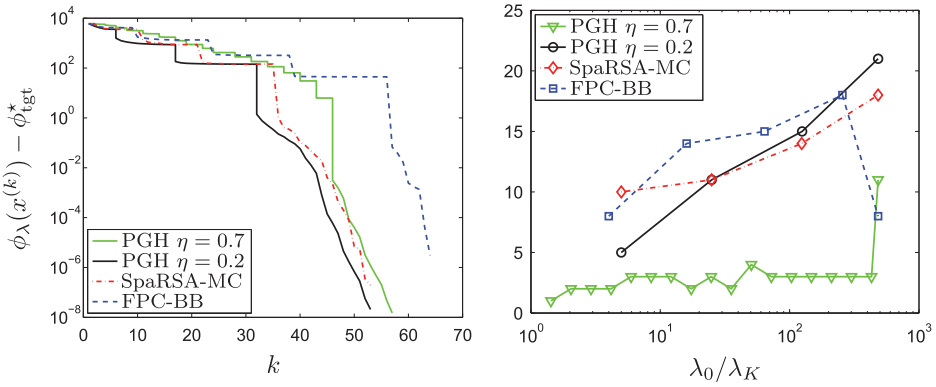
(a) Objective gap. (b) Sparsity along solution path.

FIG. 5.3. *Performance of the PGH method by varying η while keeping $\delta = 0.2$.*

We see that when δ is reduced from 0.2 to 0.1, the iterates became slightly more sparse, hence the convergence rate at each stage can be slightly faster due to better conditioning. However, this was countered by more iterations at each stage required by reaching more stringent precision, and the overall number of proximal gradient steps increased. On the other hand, increasing δ to 0.8 made the intermediate stages faster by requiring loose precision. However, this comes at the cost of less sparse iterates, and the final stage suffers a slow sublinear convergence in the beginning.

Figure 5.3 shows the effects of varying η while keeping $\delta = 0.2$. We see relatively big variations of the sparsity of the iterates, but these did not affect much of the overall iteration count. The intermediate stages may suffer from slow convergence with less sparsity, but they only need to be solved to a very rough precision. It is more important to start the last stage with a sparse vector and enjoy the fast convergence to the final precision. (See the discussions at the end of section 3.)

5.2. Comparison with SpaRSA and FPC. As mentioned in the introduction, similar approximate homotopy/continuation methods have been studied for the ℓ_1-LS problem. Here we compare the PGH method with the two most relevant ones: sparse reconstruction by separable approximation (SpaRSA) [45] and fixed point continuation (FPC) [22]. They are considered state of the art for solving sparse



(a) Objective gap. (b) Number of iterations for each λ_K.

FIG. 5.4. *Comparison with SpaRSA and FPC on a randomly generated instance.*

optimization problems. (See the performance comparisons in [45].) Both of them use the same proximal gradient step (1.3) in each iteration but with different methods for choosing the step-size. In addition, their continuation strategies are also based on reducing λ by a constant factor at each stage.

SpaRSA uses variants of the Barzilai–Borwein (spectral) method [1] for choosing L_k at each step. More specifically, at each iteration the parameter L_k is initialized as

$$L_k = \frac{\left\|A\left(x^{(k)} - x^{(k-1)}\right)\right\|_2^2}{\|x^{(k)} - x^{(k-1)}\|_2^2},$$

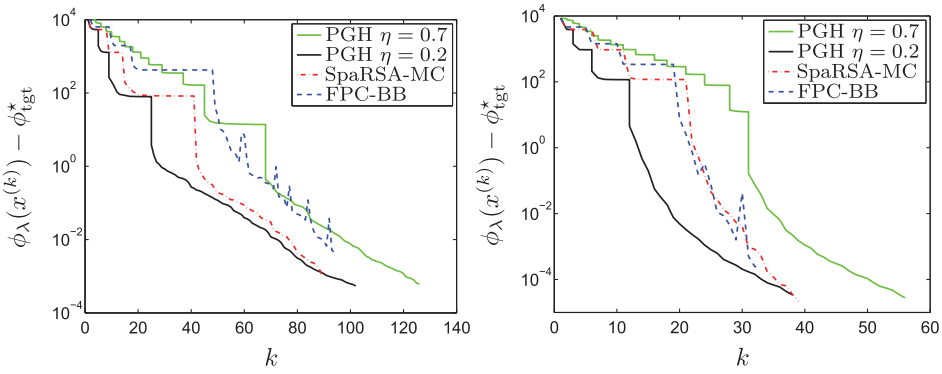
and then it is increased by a constant factor until an acceptance criterion is satisfied. When both $x^{(k)}$ and $x^{(k-1)}$ are sparse, say, $|\operatorname{supp}(x^{(k)}) \cup \operatorname{supp}(x^{(k-1)})| \le s$ for some integer s, then the above L_k satisfies

$$\rho_-(A, s) \le L_k \le \rho_+(A, s).$$

According to section 2.3, such a line search method is able to exploit the restricted strong convexity, similar to the PGH method. However, the line search acceptance criterion of SpaRSA is different from PGH, and they also have different stopping criteria for each homotopy stage. Global geometric convergence of either SpaRSA or FPC has not been established.

In our numerical experiments, we used the monotone version of SpaRSA with continuation, which we call SpaRSA-MC. For FPC, we used a more recent implementation called FPC-BB, which also employs the Barzilai–Borwein line search. Default options were used in both methods. SpaRSA-MC reduces the value of λ roughly with a factor $\eta = 0.2$, and FPC-BB has an equivalent factor $\eta = 0.25$. For meaningful comparison, we also present the results for PGH with $\eta = 0.2$, in addition to its default value $\eta = 0.7$. The same parameter $\delta = 0.2$ was used in both cases for PGH.

Figure 5.4 shows the numerical results of different algorithms on the same random instance studied in section 5.1. They demonstrate similar numerical properties, and SpaRSA-MC is especially similar to PGH with $\eta = 0.2$. The numbers of iterations at each continuation stage depend on the specific stopping criteria used in different algorithms. In Figure 5.4(b), the small number of iterations in the final stage of FPC-BB is a result of the relatively loose precision specified in its default options.



(a) Sparco 403: source separation. (b) Sparco 701: image deconvolution.

FIG. 5.5. *Comparison with SpaRSA and FPC on two image processing problems.*

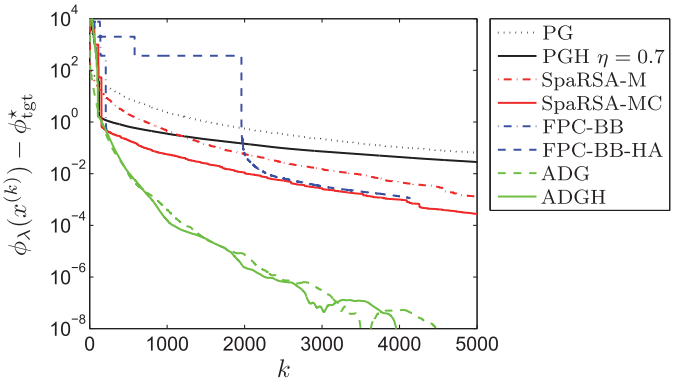


FIG. 5.6. *Comparison of different methods for solving a nonsparse random instance.*

We also compare these algorithms on two image processing problems generated by the software package Sparco [5], and the results are shown in Figure 5.5. More specifically, problem 403 is a source separation problem, in which we need to recover the well-known Cameraman (photographer) and Lena images from their mixtures with randomly blurred spike arrays. Problem 701 is an image deconvolution problem on the Cameraman image. Detailed descriptions of the images and problem setups can be found in the Sparco package; see also [19, 20] and other papers. For both problems, we used the regularization parameter $\lambda_{\rm tgt} = 0.1$ for all three algorithms.

Figure 5.5 show that these three methods have quite similar or comparable performance on the two image processing problems. As noted in [19], these problems are ill-conditioned. In particular, Figure 5.5(a) still demonstrates liner convergence in the final stage, but with a rather flat slope; in Figure 5.5(b), there is no longer linear convergence. For such ill-conditioned problems, SpaRSA and FPC often demonstrate faster local convergence because of their more sophisticated step-size rules based on the Barzilai–Borwein spectral approach. Experiments on other problems in the Sparco package reveal similar performance comparisons.

Finally, we conducted experiments on problems where the vector $\bar{x}$ is not sufficiently sparse. Figure 5.6 shows the objective gap of different methods when solving a random instance generated similarly to the one in section 5.1, but here the vector

$\bar{x}$ has 500 nonzero elements. In this case, all methods demonstrate sublinear convergence. SpaRSA-M is the monotone version of SpaRSA without continuation. FPC-BB terminated prematurely because of the default low accuracy in its stopping criterion, and FPC-BB-HA is the result after we set a much higher accuracy. We see that the algorithms with homotopy continuation still perform better than their single-stage counterparts, but the improvements are less impressive. Instead, the accelerated gradient methods ADG and ADGH outperform other methods by a big margin.

6. Conclusion and discussions. This paper studied a PGH method for solving the ℓ_1-regularized least-squares problems, focusing on its important application in sparse recovery. For such applications, the objective function is not strongly convex; hence the standard single-stage PG methods can obtain only a relatively slow convergence rate. However, we have shown that under suitable conditions for sparse recovery, all iterates of the PGH method along the solution path are sparse. With this extra sparsity structure, the objective function becomes effectively strongly convex along the solution path, and thus a geometric rate of convergence can be achieved using the homotopy approach. Our theoretical analysis is supported by several numerical experiments.

In our convergence analysis, the conditions (Assumption 3.1) that guarantee geometric convergence are rather strong, especially compared with those established in the compressed sensing literature. This is expected, since our analysis is based on keeping all the intermediate iterates sparse, rather than only for the optimal solution. Moreover, our conditions depend on not only the measurement matrix A but also the algorithmic parameters (η and δ) that control how fast the homotopy parameter is reduced and how accurately each intermediate stage needs to be solved. This again reflects the "dynamic" nature of our conditions.

In practice, it is often very hard to choose the parameters η and δ that exactly satisfy our conditions. (It is hard even for testing "static" sparse recovery conditions, such as estimating the restricted eigenvalues.) Nevertheless, our theory provides support and insight for two very effective rules used in approximate homotopy methods: reduce the regularization parameter geometrically and solve each intermediate stage to a loose relative precision. On the other hand, our experiments show that the numerical performance is not very sensitive to the choices of δ and η in a certain range, and their best values may not satisfy our conditions for global geometric convergence. In fact, with a good warm-start point and a very loose stopping criterion, each intermediate stage requires only a very small number of iterations, even with a sublinear convergence rate. The overall performance of the method hinges on rapidly getting to the linear convergence zone in the final stage, where a significant number of iterations are performed to reach a final high precision. From a theoretical perspective, this hints at the possibility for developing less restrictive conditions (than requiring all intermediate stages to have sparse iterates) that guarantee a fast global convergence rate.

We commented in the numerical experiments that accelerated gradient methods cannot automatically exploit restricted strong convexity. As discussed in [30, section 2.2] and [32], they need to explicitly use the strong convexity parameter, or a nontrivial lower bound of it, to obtain geometric convergence. In order to exploit restricted strong convexity in the ℓ_1-LS problem with $m < n$, accelerated gradient methods need an extra facility to come up with an explicit estimate of the restricted convexity parameter on the fly. Nesterov gave some suggestions along this direction in [32], and strategies such as periodic restart have been studied recently [21, 3]. However, an in-depth investigation on this matter is beyond the scope of this paper.

Acknowledgment. Tong Zhang is partially supported by NSF grants DMS-1007527, IIS-1016061, and IIS-1250985.

REFERENCES

[1] J. Barzilai and J. Borwein, *Two point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148.

[2] S. R. Becker, J. Bobin, and E. J. Candès, *NESTA: A fast and accurate first-order method for sparse recovery*, SIAM J. Imaging Sci., 4 (2011), pp. 1–39.

[3] S. R. Becker, E. J. Candès, and M. C. Grant, *Templates for convex cone problems with applications to sparse signal recovery*, Math. Program. Comput., 3 (2011), pp. 165–218.

[4] A. Beck and M. Teboulle, *A fast iterative shrinkage-threshold algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[5] E. van den Berg, M. P. Friedlander, G. Hennenfent, F. Herrmann, R. Saab, and Ö. Yilmaz, *Sparco: A Testing Framework for Sparse Reconstruction*, Tech. report TR-2007-20, Department of Computer Science, University of British Columbia, Vancouver, 2007.

[6] P. Bickel, Y. Ritov, and A. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, Ann. Statist., 37 (2009), pp. 1705–1732.

[7] J. M. Bioucas-Dias and M. A. T. Figueiredo, *A new TwIST: Two-step iterative shrinking/thresholding algorithms for image restoration*, IEEE Trans. Image Process., 16 (2007), pp. 2992–3004.

[8] A. M. Bruckstein, D. L. Donoho, and M. Elad, *From sparse solutions of systems of equations to sparse modeling of signals and images*, SIAM Rev., 51 (2009), pp. 34–81.

[9] E. J. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.

[10] E. J. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory, 51 (2005), pp. 4203–4215.

[11] E. J. Candès and T. Tao, *Near-optimal signal recovery from random projections: Universal encoding strategies?*, IEEE Trans. Inform. Theory, 52 (2006), pp. 5406–5425.

[12] E. J. Candes and T. Tao, *The Dantzig selector: Statistical estimation when p is much larger than n (with discussion)*, Ann. Statist., 35 (2007), pp. 2313–2404.

[13] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61.

[14] P. Combettes and V. Wajs, *Signal recovery by proximal forward-backward splitting*, SIAM J. Multiscale Model. Simul., 4 (2005), pp. 1168–1200.

[15] I. Daubechies, M. Defriese, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math., 57 (2004), pp. 1413–1457.

[16] D. L. Donoho, M. Elad, and V. Temlyakov, *Stable recovery of sparse overcomplete representations in the presence of noise*, IEEE Trans. Inform. Theory, 52 (2006), pp. 6–18.

[17] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.

[18] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression (with discussion)*, Ann. Statist., 32 (2004), pp. 407–499.

[19] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, *Gradient projection for sparse reconstruction: Applications to compressed sensing and other inverse problems*, IEEE J. Selected Topics in Signal Processing, 1 (2007), pp. 586–597.

[20] M. A. T. Figueiredo and R. D. Nowak, *An EM algorithm for wavelet-based image restoration*, IEEE Trans. Image Process., 12 (2003), pp. 906–916.

[21] M. Gu, L.-H. Lim, and C. J. Wu, *ParNes: A Rapidly Convergent Algorithm for Accurate Recovery of Sparse and Approximately Sparse Signals*, preprint. arXiv:0911.0492, 2009.

[22] E. T. Hale, W. Yin, and Y. Zhang, *Fixed-point continuation for ℓ_1-minimization: Methodology and convergence*, SIAM J. Optim., 19 (2008), pp. 1107–1130.

[23] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, *An interior-point method for large-scale ℓ_1-regularized least squares*, IEEE J. Selected Topics in Signal Processing, 1 (2007), pp. 606–617.

[24] V. Koltchinskii, *The Dantzig selector and sparsity oracle inequalities*, Bernoulli, 15 (2009), pp. 799–828.

[25] S. Li and Q. Mo, *New bounds on the restricted isometry constant δ_{2k}*, Appl. Comput. Harmon. Anal., 31 (2011), pp. 460–468.

[26] Z.-Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM J. Control Optim., 30 (1992), pp. 408–425.
[27] N. Meinshausen and P. Bühlmann, *High dimensional graphs and variable selection with the lasso*, Ann. Statist., 34 (2006), pp. 1436–1462.
[28] Y. Nesterov, *A method for solving a convex programming problem with convergence rate* $O(1/k^2)$, Soviet Math. Dokl., 27 (1983), pp. 372–376.
[29] Y. Nesterov, *Long-step strategies in interior-point primal-dual methods*, Math. Program., 76 (1996), pp. 47–94.
[30] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer, Boston, 2004.
[31] Y. Nesterov, *Smooth minimization of nonsmooth functions*, Math. Program., 103 (2005), pp. 127–152.
[32] Y. Nesterov, *Gradient Methods for Minimizing Composite Objective Function*, CORE discussion paper 2007/76, Center for Operations Research and Econometrics, Catholic University of Louvain, Belgium, 2007.
[33] M. Osborne, B. Presnell, and B. Turlach, *A new approach to variable selection in least squares problems*, IMA J. Numer. Anal., 20 (2000), pp. 389–404.
[34] M. Osborne, B. Presnell, and B. Turlach, *On the lasso and its dual*, J. Comput. Graph. Statist., 9 (2000), pp. 319–337.
[35] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[36] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B Stat. Methodol., 58 (1996), pp. 267–288.
[37] J. A. Tropp and S. J. Wright, *Computational methods for sparse solution of linear inverse problems*, Proc. IEEE, 98 (2010), pp. 948–958.
[38] J. A. Tropp, *Just relax: Convex programming methods for identifying sparse signals in noise*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1030–1051.
[39] P. Tseng and S. Yun, *A coordinate gradient descent method for the nonsmooth separable minimization*, Math. Program. Ser. B, 117 (2009), pp. 387–423.
[40] P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization.* unpublished manuscript, 2008.
[41] B. A. Turlach, W. N. Venables, and S. J. Wright, *Simultaneous variable selection*, Technometrics, 47 (2005), pp. 349–363.
[42] S. van de Geer and P. Bühlmann, *On the conditions used to prove oracle results for the lasso*, Electron. J. Statist., 3 (2009), pp. 1360–1392.
[43] M. J. Wainwright, *Sharp thresholds for noisy and high-dimensional recovery of sparsity using* ℓ_1*-constrained quadratic programming (lasso)*, IEEE Trans. Inform. Theory, 55 (2009), pp. 2183–2202.
[44] Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang, *A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization and continuation*, SIAM J. Sci. Comput., 32 (2010), pp. 1832–1857.
[45] S. J. Wright, R. D. Nowad, and M. A. T. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Trans. Signal Process., 57 (2009), pp. 2479–2493.
[46] C.-H. Zhang and J. Huang, *The sparsity and bias of the lasso selection in high-dimensional linear regression*, Ann. Statist., 36 (2008), pp. 1567–1594.
[47] T. Zhang, *Some sharp performance bounds for least squares regression with* l_1 *regularization*, Ann. Statist., 37 (2009), pp. 2109–2144.
[48] P. Zhao and B. Yu, *On model selection consistency of lasso*, J. Mach. Learn. Res., 7 (2006), pp. 2541–2567.