A Proximal-Gradient Homotopy Method for the Sparse Least-Squares Problem

Lin Xiao* Tong Zhang†

March 15, 2012

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.

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

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

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 problem (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., [Tib96, CDS98, BDE09]. It received revived interests in recent years due to

*Machine Learning Department, Microsoft Research, Redmond, WA 98052. Email: lin.xiao@microsoft.com
†Department of Statistics, Rutgers University, Piscataway, NJ, 08854. Email: tzhang@stat.rutgers.edu

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 [CRT06, CT06, Don06]. 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) are related by a linear model

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

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 the regularization parameter λ. This scenario is of great modern interest, and various properties of the solution $x^\star(\lambda)$ have been investigated [CT05, DET06, MB06, Tro06, ZY06, CT07, ZH08, Zha09, BRT09, Kol09, vdGB09, Wai09]. In particular, it is known that under suitable conditions on A such as the *restricted isometry property* (RIP), and as long as $\lambda \geq c\|A^T z\|_\infty$ (for some universal constant c), one can obtain a recovery bound of the optimal form

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

where $\|\bar{x}\|_0$ denotes the number of nonzero elements in $\bar{x}$. The constant in $O(\cdot)$ depends only on the so-called RIP condition that we will discuss later on, and this bound achieves the optimal order of recovery. Moreover, it is known that in this situation, the solution $x^\star(\lambda)$ is sparse [ZH08], 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.

The ℓ_1-LS problem (1) is closely related to the following two constrained convex optimization problems:

$$\underset{x}{\text{minimize}} \quad \|Ax - b\|_2^2 \quad \text{subject to} \quad \|x\|_1 \leq \Delta, \tag{3}$$

known as the *least absolute shrinkage and selection operator* (LASSO) [Tib96], and

$$\underset{x}{\text{minimize}} \quad \|x\|_1 \quad \text{subject to} \quad \|Ax - b\|_2^2 \leq \varepsilon, \tag{4}$$

where Δ and ε are two nonnegative real parameters. These problems have the same solution as (1) for appropriate choices of the parameters λ, Δ and ϵ. However, other than in some special cases, the exact correspondence between these parameters are not known a priori. Therefore, algorithms that are specific for solving one formulation may not be used directly for solving others. Nevertheless, our method can be adapted to solve (3) and (4) efficiently, either by using an augmented Lagrangian approach [YOGD08], or by using a root-finding procedure similar as the one given in [vdBF08].

1.1 Previous algorithms

There have been extensive research on numerical methods for solving the problems (1), (3) and (4). A nice survey of major practical algorithms for sparse approximation appeared in [TW10], and performance comparisons of various algorithms can be found in, e.g., [WNF09, WYGZ10, BBC11].

Here we briefly summarize the computational complexities of several methods that are most relevant for solving the ℓ_1-LS problem (1), in terms of finding an ϵ-optimal solution (i.e., obtaining an objective value within ϵ of the global minimum).

Interior-point methods were among the first approaches used for solving the ℓ_1-LS problem [CDS98, TVW05, KKL$^+$07]. The theoretical bound on their iteration complexity is $O\left(\sqrt{n}\log(1/\epsilon)\right)$, although their practical performance demonstrate 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 $O(m^2n)$ flops, which is prohibitive for large-scale applications. Therefore all customized solvers [CDS98, TVW05, KKL$^+$07] use iterative methods (such as conjugate gradients) for solving the linear equations. These methods only require matrix-vector multiplications involving A and A^T, and the computational cost per iteration can be $O(mn)$. The cost can be further reduced if the matrix-vector multiplication can be conducted more efficiently, e.g., $O(n\log n)$ if A is a partial Fourier matrix.

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

$$x^{(k+1)} = \arg\min_y \left\{ f(x^{(k)}) + \nabla f(x^{(k)})^T(y - x^{(k)}) + \frac{L_k}{2}\|y - x^{(k)}\|_2^2 + \lambda\|y\|_1 \right\}, \tag{5}$$

where we used the shorthand $f(x) = (1/2)\|Ax-b\|_2^2$, and L_k is a parameter chosen at each iteration (e.g., using a line-search procedure). The minimization problem in (5) has a closed-form solution

$$x^{(k+1)} = \text{soft}\left(x^{(k)} - \frac{1}{L_k}\nabla f(x^{(k)}),\ \frac{\lambda}{L_k}\right), \tag{6}$$

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

$$(\text{soft}(x, \alpha))_i = \text{sgn}(x_i)\max\{|x_i| - \alpha,\ 0\}, \quad i = 1, \ldots, n. \tag{7}$$

Iterative methods that use the update rule (6) include [DDM04, CW05, Nes07, HYZ08, WNF09]. 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 appropriate choices of the parameters L_k, the proximal-gradient method (5) has an iteration complexity $O(1/\epsilon)$.

Indeed, the iteration complexity $O(\log(1/\epsilon))$ can be established for (5) if $m \geq n$ and A has full column rank, since in this case the objective function in (1) is strongly convex [Nes07]. 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 [LT92, HYZ08], and fast convergence in the final stage of the algorithm has also been observed [Nes07, HYZ08, WNF09].

Variations and extensions of the proximal gradient method have been proposed to speed up the convergence in practice; see, e.g., [BDF07, WNF09, WYGZ10]. Nesterov's optimal gradient methods for minimizing smooth convex functions [Nes83, Nes04, Nes05] have also been extended to minimize composite objective functions such as in the ℓ_1-LS problem [Nes07, Tse08, BT09, BBC11]. 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 LASSO path when varying the regularization parameter λ from large to small [OPT00a, OPT00b, EHJT04]. These methods exploit the piece-wise 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 break point 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 bounding the number of breakpoints for this class of methods (for some special cases, the number of breakpoints is the same as the number of nonzeros in the solution [DT08]).

Greedy algorithms such as *orthogonal matching pursuit (OMP)* are also very popular for sparse recovery applications (e.g., [DMA97, Tro04, NT09]). However, they are not designed to solve any of the optimization problems (1), (3) or (4). Their connections with exact homotopy methods are analyzed in [DT08].

1.2 Proposed approach and contributions

We consider an *approximate* homotopy continuation method, where the key idea is to solve (1) with a large regularization parameter λ first, and then gradually decreases λ until the target regularization is reached. For each fixed λ, we employ a proximal gradient method of the form (5) to solve (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 *proximal-gradient homotopy* (PGH) method.

proximal mapping

This is not a new idea. Approximate homotopy continuation methods that use proximal gradient methods for solving each stage (with a fixed value of λ) have been studied in, e.g., [HYZ08, WNF09, WYGZ10], and superior empirical performance have 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) 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) 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 [Nes07] to choose the parameters L_k in the proximal gradient method (5).

Under the assumptions that the target value of λ is sufficiently large (such that the final solution is sparse) 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, it is sufficient to choose $\lambda \geq c\|A^T z\|_\infty$ (for some

universal constant c), which implies that the solution satisfies a recovery bound of the optimal form (2). Since each iteration of the proximal gradient method cost $O(mn)$ flops, the overall computational complexity is $O(mn\log(1/\epsilon))$, implying global geometric rate of convergence.

The low iteration complexity of our PGH method is achieved by actively exploiting the fast local linear convergence of the standard proximal gradient method when the solution $x^\star(\lambda)$ is sparse [LT92, HYZ08]. Using the homotopy continuation strategy, the proximal gradient method at each stage always starts with a point that is close to its solution. Moreover, by choosing appropriate parameters η and δ in our method, we ensure that all iterates along the solution path (i.e., not only the final points) at each stage are sufficiently sparse. Under a RIP-like assumption on A, this implies that along the homotopy path, the objective function in (1) is effectively strongly convex, and hence global geometric rate can be established using Nesterov's analysis [Nes07].

The advantage of our method over the exact homotopy path-following approach ([OPT00a, OPT00b, EHJT04]) 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. This phenomenon is predicted by our low iteration complexity, and also confirmed in our empirical studies.

Compared with interior-point methods (IPMs), our methods 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., [Nes96]), in the sense that the least-squares problem becomes better conditioned near the regularization path (cf. *central path* in IPMs). However, our results only hold for problems with provable sparse solutions, and the parameters η and δ depends on the problem data A and the regularization parameter λ. In contrast, the performance of interior-point methods 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

$$\text{minimize} \quad \|x\|_1 \quad \text{subject to} \quad Ax = b, \tag{8}$$

which is a special case of (4) with $\varepsilon = 0$. Its solution can be obtained by running our PGH method on the ℓ_1-LS problem (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) developed in this paper.

It is also worth mentioning that variants of the proximal gradient method (5) can be directly applied to the constrained LASSO formulation (3). Moreover, under suitable conditions and when the parameter Δ is set to nearly equal to $\|\bar{x}\|_1$, geometric convergence *away* from the optimal solution can be established [ANW11]. However, for sparse recovery applications, such a result is less satisfactory than the homotopy approach we analyze in this paper due to the requirement of estimating $\|\bar{x}\|_1$ — which is extremely difficult to determine efficiently in practice even for the simple noise-free case of basis pursuit. The proof techniques are also different, and the analysis of geometric convergence for PGH is more difficult than that of [ANW11], because we have to demonstrate sparsity of all the intermediate solutions in the proximal gradient steps along the homotopy path. A significantly simpler argument can be used in [ANW11], if the extra knowledge of $\|\bar{x}\|_1$ is known a priori.

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 proximal-gradient homotopy (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 conclude in Section 6 with some further discussions.

2 Preliminaries and notations

In this section, we first introduce composite gradient mapping and some of its key properties developed in [Nes07]. Then we describe Nesterov's proximal gradient method with adaptive line search, which we will use to solve the ℓ_1-LS problem at each stage of our PGH method. Finally we discuss the restricted eigenvalue conditions that allow us to show the local linear convergence of Nesterov's algorithm.

2.1 Composite gradient mapping

Consider the following optimization problem with *composite* objective function:

$$\underset{x}{\text{minimize}} \quad \Big\{\phi(x) \triangleq f(x) + \Psi(x)\Big\}, \tag{9}$$

where the function f is convex and differentiable, and Ψ is closed and convex on $\mathbb{R}^n$. The optimality condition of (9) 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., [Roc70, Section 27]). Therefore, a good measure of accuracy for any x as an approximate solution is the quantity

$$\omega(x) \triangleq \min_{\xi \in \partial\Psi(x)} \|\nabla f(x) + \xi\|_\infty. \tag{10}$$

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

Composite gradient mapping was introduced by Nesterov in [Nes07]. For any fixed point y and a given constant $L > 0$, we define a local model of $\phi(x)$ around y using a 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

$$T_L(y) = \arg\min_x \ \psi_L(y; x). \tag{11}$$

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 [Nes07, Theorem 2]:

Lemma 1. *For any $L > 0$,*

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

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

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

A direct consequence of having Lipschitz continuous gradient is the following inequality (see, e.g., [Nes04, Theorem 2.1.5]):

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_f}{2}\|y - x\|_2^2, \quad \forall\, x, y \in \mathbb{R}^n. \tag{12}$$

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

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

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

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

Proof. Let $D\phi(x)[u]$ denote the directional derivative of ϕ at x along the direction u, i.e,

$$D\phi(x)[u] = \lim_{\alpha \downarrow 0} \frac{1}{\alpha}\big(\phi(x + \alpha u) - \phi(x)\big).$$

Corollary 1 in [Nes07] states that for any $u \in \mathbb{R}^n$ with $\|u\|_2 = 1$, the following inequality holds:

$$D\phi(T_L(y))[u] \geq -\left(1 + \frac{S_L(y)}{L}\right)\|g_L(y)\|_2.$$

In addition, it is shown in [Nes07] that for any $x \in \mathbb{R}^n$,

$$\min_{\xi \in \partial\Psi(x)} \|\nabla f(x) + \xi\|_2 = -\min_{\|u\|_2 = 1} D\phi(x)[u].$$

(See [Nes07, Section 2].) Therefore, we have

$$\omega(T_L(y)) \leq \min_{\xi \in \partial\Psi(T_L(y))} \|\nabla f(T_L(y)) + \xi\|_2 \leq \left(1 + \frac{S_L(y)}{L}\right)\|g_L(y)\|_2.$$

The last desired inequality follows from the fact $S_L(y) \leq L_f$. □

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 $\{\hat{x}, \hat{M}\}$

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

$$\begin{aligned} f(x) &= \frac{1}{2}\|Ax - b\|_2^2 \\ \phi_\lambda(x) &= f(x) + \lambda\|x\|_1. \end{aligned}$$

Correspondingly, we add the subscript λ in specifying the composite gradient mapping:

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

We call the process of computing $T_L(y)$ a proximal gradient step. For the ℓ_1-LS problem, $T_{\lambda,L}(x)$ has the closed-form solution given in (6). 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 proximal gradient methods in [Nes07]. We use the non-accelerated primal-gradient version described in

Algorithms 1 and 2, which correspond to (3.1) and (3.2) in [Nes07], 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 $\gamma_{\text{inc}} > 1$. A key feature of this algorithm is the adaptive line search: it always tries to use a smaller Lipschitz constant first at each iteration.

Each iteration of the proximal gradient method generates the next iterate in 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 for line search ensures

$$\begin{aligned} \phi_\lambda(x^{(k+1)}) &\leq \psi_{\lambda,M_k}\left(x^{(k)}, x^{(k+1)}\right) = \psi_{\lambda,M_k}\left(x^{(k)}, T_{\lambda,M_k}(x^{(k)})\right) \\ &\leq \phi_\lambda(x^{(k)}) - \frac{1}{2M_k}\left\|g_{\lambda,M_k}(x^{(k)})\right\|_2^2, \end{aligned} \tag{13}$$

where the last inequality follows from Lemma 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, $x^{(k+1)}$ is an optimal solution.

The only difference between Algorithm 2 and Nesterov's gradient method [Nes07, (3.2)] is that Algorithm 2 has an explicit stopping criterion. This stopping criterion is based on the optimality residue $\omega_\lambda(x^{(k+1)})$ being small. For the ℓ_1-LS problem, it can be computed with additional $O(n)$ flops given the gradient $\nabla f(x)$. For other problems, depending on the form of Ψ, this residue may be hard to compute. But we can always use the alternative stopping criterion

$$\left\|g_{\lambda,M_k}(x^{(k)})\right\|_2 \leq \hat{\epsilon}.$$

According to Lemma 2, these two measures may differ by a factor $(1 + S_{M_k}(x^{(k+1)})/M_k)$. So the precision $\hat{\epsilon}$ may need to be reduced by a similar factor.

Since f has Lipschitz constant L_f, the inequality (12) implies that the line search procedure is guaranteed to terminate if $L \geq L_f$. Therefore, we have

$$L_{\min} \leq L_k \leq M_k < \gamma_{\text{inc}} L_f. \tag{14}$$

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 [Nes07] 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

$$N_k \leq 2(k+1) + \log_2 \frac{L_f}{L_{\min}}. \tag{15}$$

Nesterov established the following iteration complexities of Algorithm 2 for finding an ϵ-optimal solution of the problem (9):

- If ϕ_λ is convex but not strongly convex, then the convergence is sublinear, with an iteration complexity $O(1/\epsilon)$ [Nes07, Theorem 4];
- If ϕ_λ is strongly convex, then the convergence is geometric, with an iteration complexity $O(\log(1/\epsilon))$ [Nes07, 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 interested case $m < n$, the objective function in Problem (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 [Nes07] when the solution is sparse). Nevertheless, as explained in the introduction, we can use a homotopy continuation strategy 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 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 notations for sparse recovery. For a vector $x \in \mathbb{R}^n$, let

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

Throughout the paper, we denote $\mathrm{supp}(\bar{x})$ by $\bar{S}$, and use $\bar{S}^c$ for its complement. We use the notations $x_{\bar{S}}$ 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 *restricted isometry property* (RIP) introduced in [CT05]. 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 1. *Given an integer $s > 0$, we say that A satisfies the restricted eigenvalue condition at sparsity level s if there exists positive constants $\rho_-(A, s)$ and $\rho_+(A, s)$ such that*

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

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), 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 (12), could use a much smaller restricted Lipschitz constant instead of the global constant $L_f = \rho_+(A, n)$. More precisely, we have the following lemma.

Lemma 3. *Let $f(x) = (1/2)\|Ax - b\|_2^2$. Suppose x and y are two sparse vectors such that*

$$|\mathrm{supp}(x) \cup \mathrm{supp}(y)| \leq s$$

for some integer $s < m$. Then the following two inequalities hold:

$$f(y) \;\leq\; f(x) + \langle \nabla f(x), y - x\rangle + \frac{\rho_+(A,s)}{2}\|y - x\|_2^2, \tag{16}$$

$$f(y) \;\geq\; f(x) + \langle \nabla f(x), y - x\rangle + \frac{\rho_-(A,s)}{2}\|y - x\|_2^2. \tag{17}$$

Proof. For any $x, y \in \mathbb{R}^n$, it is straightforward to verify that if $f(x) = (1/2)\|Ax - b\|_2^2$, then

$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle \;=\; \frac{1}{2}\|A(y - x)\|_2^2.$$

Since the assumption $|\mathrm{supp}(x) \cup \mathrm{supp}(y)| \leq s$ implies $\|y - x\|_0 \leq s$, we use the definition of restricted eigenvalues to conclude

$$\rho_-(A,s)\|y - x\|_2^2 \;\leq\; \|A(y - x)\|_2^2 \;\leq\; \rho_+(A,s)\|y - x\|_2^2.$$

These lead to the two desired inequalities. □

The inequality (16) represents *restricted smoothness*, and (17) represents *restricted strong convexity*. A key feature of our PGH method is that sparsity along the whole solution path can be enforced. Therefore the objective function in (1) becomes strongly convex along the solution path if the sparse eigenvalues in Definition 1 are well behaved (i.e., they grow slowly when s is increased). In such a situation, the PGH method exhibits geometric convergence along the homotopy path, and the convergence rate depends on a *restricted condition number*, defined as

$$\kappa(A,s) = \frac{\rho_+(A,s)}{\rho_-(A,s)}. \tag{18}$$

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

3 A proximal-gradient homotopy method

The key idea of the proximal-gradient homotopy (PGH) method is to solve (1) with a large regularization parameter λ_0 first, and then gradually decreases λ until the target regularization is reached. For each fixed λ, we employ Nesterov's proximal-gradient method described in Algorithms 1 and 2, to solve problem (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 listed as Algorithm 3. To make the presentation more clear, 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:

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 $\hat{x}^{(\mathrm{tgt})}$

- 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) with a proportional precision $\delta\lambda_K$. For the last stage with λ_{tgt}, we solve to 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 RIP or restricted eigenvalue condition. It turns out that such conditions are also sufficient for fast convergence of our PGH method. More precisely, we have the following assumption:

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

$$\lambda_{\mathrm{tgt}} \;\geq\; \max\left\{4,\; \frac{\gamma+1}{(1-\delta')\gamma - (1+\delta')}\right\} \|A^T z\|_\infty. \tag{19}$$

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

$$\tilde{s} \;>\; \frac{16\big(\gamma_{\mathrm{inc}}\rho_+(A, \bar{s}+2\tilde{s}) + 2\rho_+(A,\tilde{s})\big)}{\rho_-(A, \bar{s}+\tilde{s})}(1+\gamma)\bar{s}. \tag{20}$$

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

According to [ZH08], the above assumption implies that the solution $x^\star(\lambda)$ of (1) is sparse whenever $\lambda \geq \lambda_{\mathrm{tgt}}$; more specifically, $\|x^\star(\lambda)_{\bar{S}^c}\|_0 \leq \tilde{s}$ (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 a similar argument as that of [ZH08]. Before stating the main convergence results, we make some further remarks on Assumption 1.

- The condition (19) states that the λ must be sufficiently large to dominate the noise. Such a condition is adequate for sparse recovery applications because recovery performance given

in (2) achieves optimal error bound under stochastic noise model by picking λ of the order $\|A^T z\|_\infty$ [CT07, ZH08, Zha09, BRT09, Kol09, vdGB09, Wai09]. 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 the conditions (20) is necessary and standard in sparse recovery analysis. This is closely related to the RIP condition of [CT05] 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 $\nu = 0.2$ at $s = 193(1+\gamma)\bar{s}$, then we may take $\gamma_{\text{inc}} = 2$ and $\tilde{s} = 96(1+\gamma)\bar{s}$ so that the condition (20) is satisfied. To see this, let $s = \bar{s} + 2\tilde{s}$ and note that

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

  Therefore we have

$$\tilde{s} = 96(1+\gamma)\bar{s} = 64\frac{1+\nu}{1-\nu}(1+\gamma)\bar{s} > 16\frac{2\rho_+(A, \bar{s} + 2\tilde{s}) + 2\rho_+(A, \tilde{s})}{\rho_-(A, \bar{s} + \tilde{s})}(1+\gamma)\bar{s}.$$

  Although for practical purpose these constants are rather large, it is worth mentioning that our analysis focuses on the high level message, without paying special attention to optimizing the constants.

- If $L_{\min} > \gamma_{\text{inc}}\rho_+(A, \bar{s} + 2\tilde{s})$, then we may simply replace $\gamma_{\text{inc}}\rho_+(A, \bar{s} + 2\tilde{s})$ by $L_{\min}$ in the assumption, and all theorem statements hold with $\gamma_{\text{inc}}\rho_+(A, \bar{s} + 2\tilde{s})$ replaced by $L_{\min}$. Nevertheless in practice, it is natural to simply pick

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

  where A_i is the i-th column of A. It automatically satisfies the condition $L_{\min} \leq \rho_+(A, \bar{s}+2\tilde{s})$.

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 along the solution path are sparse, and Algorithm 2 has geometric convergence. To simplify the presentation, we use a single symbol κ to denote the restricted condition number

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

Theorem 1. *Suppose Assumption 1 holds. If the initial point $x^{(0)}$ in Algorithm 2 satisfies*

$$\left\|x^{(0)}_{\bar{S}^c}\right\|_0 \leq \tilde{s}, \qquad \omega_\lambda(x^{(0)}) \leq \delta'\lambda, \tag{21}$$

then for all $k \geq 0$, we have

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

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 in Algorithm 3. 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 2. *Suppose Assumption 1 holds with $\lambda_{\rm tgt} \le \lambda_0$ and the parameters δ and η in Algorithm 3 are chosen such that*

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

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

1. *The condition (21) holds for each call of Algorithm 2. For $K = 0, \ldots, N-1$, the number of proximal-gradient steps in each call of Algorithm 2 is no more than*

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

   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 $\hat{x}^{(K)}$ satisfies*

$$\phi_{\lambda_{\rm tgt}}(\hat{x}^{(K)}) - \phi^\star_{\lambda_{\rm tgt}} \le \eta^{2(K+1)} \frac{4.5(1+\gamma)\lambda_0^2\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}, \tag{22}$$

   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 proximal-gradient steps is no more than*

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

   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

$$K_0 = \left\lceil \frac{1}{2} \ln\left(\frac{4.5(1+\gamma)\lambda_0^2\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})\epsilon_0}\right) \Bigg/ \ln \eta^{-1} \right\rceil - 1.$$

  From the inequality (22), we see that if $0 \le K_0 \le N-1$, then for all $K \ge K_0$,

$$\phi_{\lambda_{\rm tgt}}(\hat{x}^{(K)}) - \phi^\star_{\lambda_{\rm tgt}} \le \epsilon_0.$$

  If we let $\epsilon_0 \to 0$ and run the PGH method forever, then the number of proximal-gradient iterations is no more than $O(\ln(\lambda_0/\epsilon_0))$ to achieve an ϵ_0 accuracy both on the gap of objective value and on 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 can use 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 $O((\ln(\lambda_0/\epsilon))^2)$.

Theorem 2 plus restricted strong convexity immediately implies that the approximate solutions $\hat{x}^{(K)}$ (and the last step solution $\hat{x}^{(\text{tgt})}$) also converge to $x^\star(\lambda_{\text{tgt}})$ at a globally geometric rate. A particularly interesting case is noise-free compressed sensing using the BP formulation (8), which has the optimal solution $\bar{x}$. For this problem, we can simply run Algorithm 3 with $\lambda_{\text{tgt}} = 0$ to solve (8). While the convergence metrics such as objective value gap or optimality residue are no longer informative in this case, Theorem 2 implies geometric convergence of the recovery error $\|\hat{x}^{(K)} - \bar{x}\|_2$. More precisely, we have:

Corollary 1. *Suppose $b = A\bar{x}$ and the assumptions stated in Theorem 2 hold. We can choose an arbitrarily small $\lambda_{\text{tgt}} > 0$ in Algorithm 3, and after K outer iterations, we have*

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

Note that part 1 of Theorem 2 implies that K outer iterations of Algorithm 3 requires no more than $O(K)$ proximal-gradient steps. This result can be interpreted as a global geometric rate of convergence for solving the BP problem.

4 Proofs of convergence results

The proofs of our convergence results are divided into the following subsections. In Section 4.1, we show that under Assumption 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 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 the assumption (19):

$$\begin{aligned}
(1-\delta')\lambda - \|A^T z\|_\infty &> 0 && (23)\\
(1+\delta')\lambda + \|A^T z\|_\infty &\leq 2\lambda && (24)\\
\lambda + \|A^T z\|_\infty &\leq (2-\delta')\lambda && (25)\\
\frac{(1+\delta')\lambda + \|A^T z\|_\infty}{(1-\delta')\lambda - \|A^T z\|_\infty} &\leq \gamma. && (26)
\end{aligned}$$

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. *Suppose Assumption 1 holds, and $\lambda \geq \lambda_{\rm tgt}$. If x is sparse, i.e., $\|x_{\bar{S}^c}\|_0 \leq \tilde{s}$, and it satisfies the approximate optimality condition*

$$\min_{\xi \in \partial \|x\|_1} \left\| A^T(Ax-b) + \lambda \xi \right\|_\infty \leq \delta' \lambda, \tag{27}$$

then we have

$$\|(x-\bar{x})_{\bar{S}^c}\|_1 \leq \gamma \|(x-\bar{x})_{\bar{S}}\|_1 \tag{28}$$

and

$$\|x-\bar{x}\|_2 \leq \frac{2\lambda\sqrt{\bar{s}}}{\rho_-(A, \bar{s}+\tilde{s})} \tag{29}$$

and

$$\phi_\lambda(x) \leq \phi_\lambda(\bar{x}) + \frac{2\delta'(1+\gamma)\lambda^2 \bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}. \tag{30}$$

Proof. Let $\xi \in \partial\|x\|_1$ be a subgradient that achieves the minimum on the left-hand side of (27). Then the approximate optimality condition leads to

$$\begin{array}{rcl} (x-\bar{x})^T \left(A^T(Ax-b) + \lambda\xi \right) & \leq & \|x-\bar{x}\|_1 \left\| A^T(Ax-b) + \lambda\xi \right\|_\infty \\ & \leq & \delta'\lambda \|x-\bar{x}\|_1. \end{array}$$

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

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

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 $\|\xi\|_\infty \leq 1$)

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

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

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

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

$$\|A(x-\bar{x})\|_2^2 + \left((1-\delta')\lambda - \|A^T z\|_\infty \right) \|(x-\bar{x})_{\bar{S}^c}\|_1 \;\leq\; \left((1+\delta')\lambda + \|A^T z\|_\infty \right) \|(x-\bar{x})_{\bar{S}}\|_1. \tag{31}$$

By further using the inequalities (23) and (26), we obtain

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

which is the first desired result in (28).

Since by assumption $\|x_{\bar{S}^c}\|_0 \le \tilde{s}$, we can use the restricted eigenvalue condition to obtain

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

where the second inequality is a result of (31), the third inequality follows from (24), and the fourth inequality holds because $|\bar{S}| = \bar{s}$. This proves the second desired bound in (29).

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

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

From the inequality in (28), we have

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

Therefore,

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

which, together with (29), leads to the third desired result. □

The following result 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.

Lemma 5. *Suppose Assumption 1 holds, and $\lambda \ge \lambda_{\mathrm{tgt}}$. Consider x such that*

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

then

$$\max\left\{\frac{1}{2\lambda}\|A(x-\bar{x})\|_2^2,\; \|x-\bar{x}\|_1\right\} \le \frac{4(1+\gamma)\lambda\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}.$$

In fact, similar results holds under the condition $\omega_\lambda(x) \le \delta'\lambda$, and are already proved in Lemma 4. However, in the proximal gradient 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 iterates along the solution path, we need to show them when the objective function is sufficiently small, which is more involved.

Proof. For notational convenience, let

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

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

$$\frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1 \leq \frac{1}{2}\|A\bar{x}-b\|_2^2 + \lambda\|\bar{x}\|_1 + \Delta. \tag{32}$$

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

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

Plugging the above inequality into (32), 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

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

By 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

$$\frac{1}{2}\|A(x-\bar{x})\|_2^2 + \left(\lambda - \|A^Tz\|_\infty\right)\|(x-\bar{x})_{\bar{S}^c}\|_1 \;\leq\; \left(\lambda + \|A^Tz\|_\infty\right)\|(x-\bar{x})_{\bar{S}}\|_1 + \Delta. \tag{33}$$

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

$$\|x-\bar{x}\|_1 \;\leq\; \frac{\Delta}{\delta'\lambda} \;=\; \frac{2(1+\gamma)\lambda\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}. \tag{34}$$

From (23), 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 (33) to obtain

$$\begin{aligned}
\frac{1}{2}\|A(x-\bar{x})\|_2^2 &\leq \left(\lambda + \|A^Tz\|_\infty\right)\|(x-\bar{x})_{\bar{S}}\|_1 + \Delta \\
&\leq (2\lambda - \delta'\lambda)\|(x-\bar{x})_{\bar{S}}\|_1 + \Delta \\
&\leq (2\lambda - \delta'\lambda)\frac{2(1+\gamma)\lambda\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})} + \frac{2\delta'(1+\gamma)\lambda^2\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})} \\
&= \frac{4\lambda(1+\gamma)\lambda\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})},
\end{aligned}$$

where in the second inequality we used (25), and in the third inequality we used (34). This means the claim holds.

In the second case, the assumption in (34) does not hold. Then $\Delta < \delta'\lambda\|x-\bar{x}\|_1$ and (33) implies

$$\frac{1}{2}\|A(x-\bar{x})\|_2^2 + \left(\lambda - \|A^T z\|_\infty\right)\|(x-\bar{x})_{\bar{S}^c}\|_1 \le \left(\lambda + \|A^T z\|_\infty\right)\|(x-\bar{x})_{\bar{S}}\|_1 + \delta'\lambda\|x-\bar{x}\|_1.$$

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

$$\frac{1}{2}\|A(x-\bar{x})\|_2^2 + \left((1-\delta')\lambda - \|A^T z\|_\infty\right)\|(x-\bar{x})_{\bar{S}^c}\|_1 \;\le\; \left((1+\delta')\lambda + \|A^T z\|_\infty\right)\|(x-\bar{x})_{\bar{S}}\|_1. \quad (35)$$

By further using the inequalities (23) and (26), we get

$$\|(x-\bar{x})_{\bar{S}^c}\|_1 \;\le\; \frac{(1+\delta')\lambda + \|A^T z\|_\infty}{(1-\delta')\lambda - \|A^T z\|_\infty}\|(x-\bar{x})_{\bar{S}}\|_1 \;\le\; \gamma\|(x-\bar{x})_{\bar{S}}\|_1. \quad (36)$$

Moreover, we can use the restricted eigenvalue condition and the assumption $\|x_{\bar{S}^C}\|_0 \le \tilde{s}$ to obtain

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

where the second inequality follows from (35), the third inequality follows from (24), and the forth inequality holds because $|\bar{S}| = \bar{s}$. Hence

$$\|x-\bar{x}\|_2 \le \frac{4\lambda\sqrt{\bar{s}}}{\rho_-(A, \bar{s}+\tilde{s})}.$$

The above arguments also imply

$$\frac{1}{2}\|A(x-\bar{x})\|_2^2 \;\le\; 2\lambda\sqrt{\bar{s}}\,\|x-\bar{x}\|_2 \;\le\; \frac{8\lambda^2\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})} \;\le\; \frac{4(1+\gamma)\lambda^2\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})},$$

where the last inequality holds because $\gamma > 1$. Finally, using (36), we get

$$\|x-\bar{x}\|_1 \;\le\; (1+\gamma)\|(x-\bar{x})_{\bar{S}}\|_1 \;\le\; (1+\gamma)\sqrt{\bar{s}}\,\|(x-\bar{x})_{\bar{S}}\|_2 \;\le\; \frac{4(1+\gamma)\lambda\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}.$$

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 6. *Suppose Assumption 1 holds, and $\lambda \ge \lambda_{\mathrm{tgt}}$. Suppose x satisfies*

$$\|x_{\bar{S}^c}\|_0 \le \tilde{s}, \qquad \phi_\lambda(x) \le \phi_\lambda(\bar{x}) + \frac{2\delta'(1+\gamma)\lambda^2\bar{s}}{\rho_-(A, \bar{s}+\tilde{s})}, \quad (37)$$

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 as in (6). That is,

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

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:

- $\lambda/4L$ on elements of $x_{\bar{S}^c}$,
- $\lambda/4L$ on elements of $(1/L)A^Tz$, and
- $\lambda/2L$ on elements of $(1/L)A^TA(x-\bar{x})$.

Since by assumption $\|A^Tz\|_\infty \le \lambda/4$, we have $\big|\{j : ((1/L)A^Tz)_j > \lambda/4L\}\big| = 0$. Therefore,

$$\big\|\big(T_L(x)\big)_{\bar{S}^c}\big\|_0 \ \le\ \big|\{j \in \bar{S}^c : |x_j| > \lambda/4L\}\big| + \big|\{j : \big|\big(A^TA(x-\bar{x})\big)_j\big| \ge \lambda/2\}\big|.$$

Note that

$$\begin{aligned}
\big|\{j \in \bar{S}^c : |x_j| \ge \lambda/4L\}\big| &= \big|\{j \in \bar{S}^c : |(x-\bar{x})_j| \ge \lambda/4L\}\big| \\
&\le \big|\{j : |(x-\bar{x})_j| \ge \lambda/4L\}\big| \\
&\le 4L\lambda^{-1}\|x-\bar{x}\|_1 \\
&\le \frac{16L(1+\gamma)\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})},
\end{aligned} \tag{38}$$

where the last inequality follows from Lemma 5.

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

$$S' \subset \{j : |(A^TA(x-\bar{x}))_j| \ge \lambda/2\}.$$

Then there exists u such that $\|u\|_\infty = 1$ and $\|u\|_0 = s'$, and $s'\lambda/2 \le u^TA^TA(x-\bar{x})$. Moreover,

$$s'\lambda/2 \ \le\ u^TA^TA(x-\bar{x}) \ \le\ \|Au\|_2\|A(x-\bar{x})\|_2 \ \le\ \sqrt{\rho_+(A,s')}\sqrt{s'}\sqrt{\frac{8(1+\gamma)\lambda^2\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})}},$$

where the last inequality again follows from Lemma 5. Taking squares of both sides of the above inequality gives

$$s' \le \frac{32\,\rho_+(A,s')(1+\gamma)\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})} \le \frac{32\,\rho_+(A,\tilde{s})(1+\gamma)\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})} < \tilde{s},$$

where the last inequality is due to (20). Since $s' = |S'|$ achieves the maximum possible value such that $s' \le \tilde{s}$ for any subset S' of $\{j : |(A^TA(x^{(k)}-\bar{x}))_j| \ge \lambda/2\}$, and the above inequality shows that $s' < \tilde{s}$, we must have

$$S' = \{j : |(A^TA(x^{(k)}-\bar{x}))_j| \ge \lambda/2\},$$

and thus

$$\left|\{j : |(A^TA(x^{(k)} - \bar{x}))_j| \geq \lambda/2\}\right| = s' \leq \left\lfloor \frac{32\,\rho_+(A,\tilde{s})(1+\gamma)\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})} \right\rfloor.$$

Finally, combining the above bound with the bound in (38) gives

$$\left\|\left(T_{\lambda,L}(x)\right)_{\bar{S}^c}\right\|_0 \;\leq\; \frac{16\,(L + 2\rho_+(A,\tilde{s}))}{\rho_-(A,\bar{s}+\tilde{s})}(1+\gamma)\bar{s}.$$

Under the assumption $L < \gamma_{\text{inc}}\rho_+(A, \bar{s}+2\tilde{s})$ and (20), 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 (13), the objective value $\phi_\lambda(x^{(k)})$ is monotone decreasing. So if $x^{(0)}$ satisfies the condition (37), every iterate $x^{(k)}$ satisfies the same condition. 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 procedure (Algorithm 1) always terminates with

$$M_k \leq \gamma_{\text{inc}}\rho_+(A, \bar{s}+2\tilde{s}). \tag{39}$$

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 6 implies that $\left\|\left(T_{\lambda,L}(x)\right)_{\bar{S}^c}\right\|_0 < \tilde{s}$ and the restricted smoothness property (16) implies the termination of line-search.

4.2 Proof of Theorem 1

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

First, since the sub-level set $\{x : \phi_\lambda(x) \leq \phi_\lambda(x^{(0)})\}$ is bounded and $\phi_\lambda(x^{(k)})$ is monotone decreasing, the sequence $\left\{x^{(k)}\right\}_{k=0}^{\infty}$ is bounded. By the Bolzano-Weierstrass theorem, it has a convergent subsequence and a corresponding accumulation point. Moreover, from the inequality (13) 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, this implies that any accumulation point of the sequence $\left\{x^{(k)}\right\}_{k=0}^{\infty}$ satisfies the optimality condition, therefore is a minimizer of ϕ_λ.

Let $x^\star(\lambda)$ denote an accumulation point of the sequence $\left\{x^{(k)}\right\}_{k=0}^{\infty}$. As a consequence of Lemma 6, 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 (17), we have

$$f(x) \geq f(x^\star) + \langle \nabla f(x^\star(\lambda)), x - x^\star(\lambda)\rangle + \frac{\rho_-(A, \bar{s}+2\tilde{s})}{2}\|x - x^\star(\lambda)\|_2^2. \tag{40}$$

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

$$\nabla f(x^\star(\lambda)) + \lambda\xi = 0. \tag{41}$$

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

$$\lambda\|x\|_1 \geq \lambda\|x^\star(\lambda)\|_1 + \langle \lambda\xi, x - x^\star(\lambda)\rangle. \tag{42}$$

Adding the two inequalities (40) and (42) and using (41), we get

$$\phi_\lambda(x) - \phi_\lambda(x^\star(\lambda)) \geq \frac{\rho_-(A, \bar{s} + 2\tilde{s})}{2}\|x - x^\star(\lambda)\|_2^2, \qquad \forall\, x : \|x_{\bar{S}^c}\|_0 \leq \tilde{s}. \tag{43}$$

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 $\left\{x^{(k)}\right\}_{k=0}^{\infty}$.

Next we show that under the assumptions in Lemma 6, especially with $x^{(0)}$ satisfying (37), 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

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

Since the conclusion of Lemma 6 implies that $\|x_{\bar{S}^c}^{(k)}\|_0 \leq \tilde{s}$ for all $k \geq 0$, we can use the "restricted" strong convexity property (43) to obtain

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

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

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

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

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

where the second inequality follows from (39). Therefore, we have

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

where

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

is a restricted condition number. Note that the above convergence rate does not depend on λ.

4.3 Proof of Theorem 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 proximal gradient 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 7. *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

$$\frac{1+\delta}{1+\delta'} \le \eta < 1. \tag{44}$$

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

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

Since the condition (44) implies $\delta\lambda_K + (1-\eta)\lambda_K \le \delta'\lambda_{K+1}$, we have the desired result. □

Lemma 8. *Assume that for some x and $\lambda \ge \lambda_{\rm tgt}$,*

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

Then for all $\lambda' \in [\lambda_{\rm tgt}, \lambda]$, 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

$$\begin{aligned} \phi_{\lambda'}(x) - \phi_{\lambda'}(x^\star(\lambda')) &\leq \langle \nabla f(x) + \lambda' \xi(\lambda), x - x^\star(\lambda') \rangle \\ &\leq (\|\nabla f(x) + \lambda \xi(\lambda)\|_\infty + \lambda - \lambda') \|x - x^\star(\lambda')\|_1 \\ &= (\omega_\lambda(x) + \lambda - \lambda') \|x - x^\star(\lambda')\|_1. \end{aligned} \tag{45}$$

By Lemma 4, we have

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

and

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

Therefore, we have

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

Now we obtain from (45) that

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

This proves the desired result. □

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

Using Lemma 2, we can upper bound the measure for approximate optimality as

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

where the second inequality follows from

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

which are direct consequences of the line-search termination criterion, the restricted smoothness property (16) and the restricted strong convexity property (17).

In order to bound the norm of $g_{\lambda, M_k}(x^{(k)})$, we use the inequality (13) and Theorem 1 to obtain

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

where $\phi_\lambda^\star = \phi_\lambda(x^\star(\lambda)) = \min_x \phi_\lambda(x)$. Therefore, in order to satisfy the stopping criteria

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

it suffices to ensure

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

which requires

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

We still need to bound the gap $\phi_\lambda(x^{(0)}) - \phi_\lambda^\star$. Since Lemma 7 implies that $\omega_\lambda(x^{(0)}) \le \delta'\lambda$, we can obtain directly from Lemma 8 the following inequality 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

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

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

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

This proves Part 1 of Theorem 2. We note that this bound is independent of λ.

In the homotopy method (Algorithm 3), after K outer iterations for $K \le N-1$, we have from Lemma 7 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 and $\lambda_{K+1} = \eta^{K+1}\lambda_0$. Moreover, from Lemma 8 with $\lambda' = \lambda_{\rm tgt}$, $\lambda = \lambda_{K+1}$, and $x = \hat{x}^{(K)}$, we obtain

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

This proves Part 2 of Theorem 2.

In Algorithm 3, the number of outer iterations, excluding the last one for λ_{tgt}, is

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

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

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

Finally, when the PGH method terminates, we have $\omega_{\lambda_{\text{tgt}}}(\hat{x}^{(\text{tgt})}) \leq \epsilon$. Therefore we can apply Lemma 8 with $\lambda = \lambda' = \lambda_{\text{tgt}}$ and $x = \hat{x}^{(\text{tgt})}$ to obtain the last desired bound in Part 3.

5 Numerical experiments

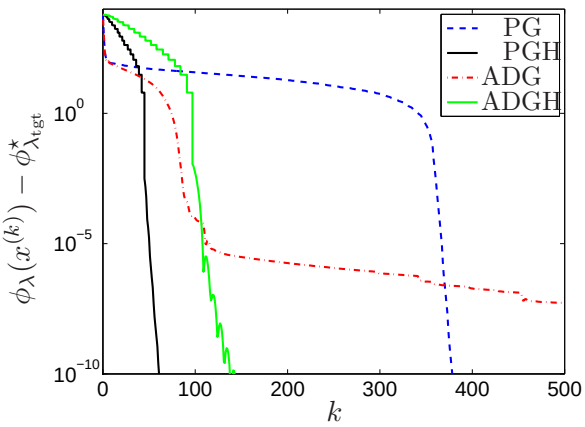
In this section, we present numerical experiments to supports our theoretical analysis. First, we illustrate the numerical properties of the PGH method by comparing it with several other methods. More specifically, we implemented the following methods for solving the ℓ_1-LS problem:

- PG: Nesterov's proximal gradient 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., Algorithm (4.9) in [Nes07].
- ADGH: the PGH method in Algorithm 3, but with PG replaced by ADG.

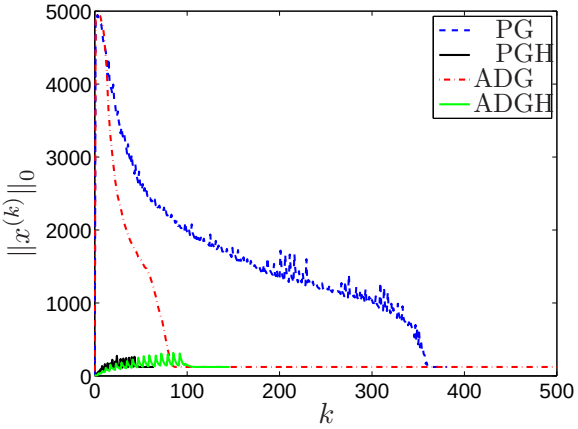
We generated a random instance of (1) with dimensions $m = 1000$ and $n = 5000$. The entries of the matrix $A \in \mathbb{R}^{m\times n}$ are generated independently with the uniform distribution over the interval $[-1, +1]$. The vector $\bar{x} \in \mathbb{R}^n$ was generated with the same distribution at 100 randomly chosen coordinates (i.e., $\bar{s} = |\text{supp}(\bar{x})| = 100$). 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 first experiment, we set $\sigma = 0.01$ and choose $\lambda_{\text{tgt}} = 1$. For this particular 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 1 illustrates various numerical properties of the four different methods for solving this random instance. We used the parameters $\gamma_{\text{inc}} = 2$ and $\gamma_{\text{dec}} = 2$ 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 the first four subfigures (a)-(d), the horizontal axes show the cumulative count of inner iterations (total number of proximal-gradient steps). For the two homotopy methods, the vertical line segments in the subfigures (a), (c) and (f) indicate switchings of homotopy stages (when the value of λ is reduced by the factor η) — they reflect the change of objective function or the optimality residue for the same vector $x^{(k)}$.

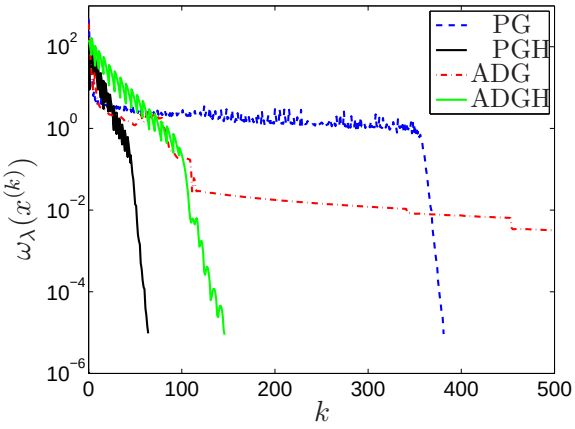
Figure 1(a) shows the objective gap $\phi_\lambda(x^{(k)}) - \phi^\star_{\lambda_{\text{tgt}}}$ versus the total number of iterations k. The PG method solves the problem with the target regularization parameter λ_{tgt} directly. For the first 350 or so iterations, it demonstrated a slow sublinear convergence rate (theoretically $O(1/k)$), but converged rapidly for the last 30 iterations with a linear rate. Referring to Figure 1(b), we see that



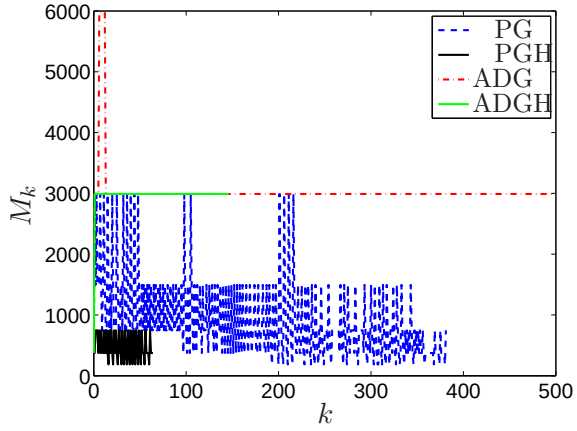
(a) Objective gap.



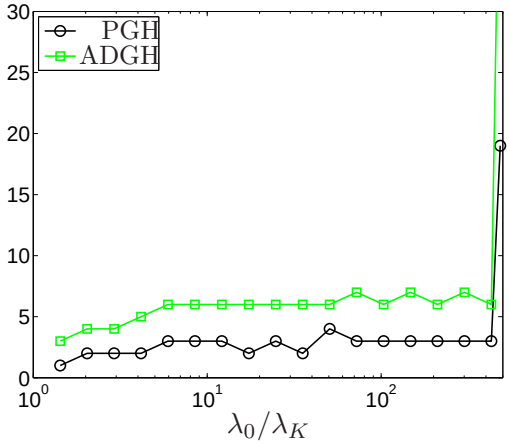
(b) Sparsity along solution path.



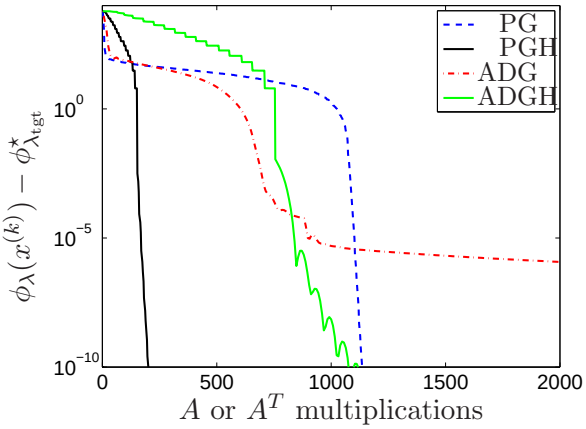
(c) Optimality residues.



(d) Line search results.



(e) Number of iterations for each λ_K.



(f) Number of matrix-vector multiplications.

Figure 1: Solving a random instance of the ℓ_1-LS problem. Problem sizes: $m = 1000$, $n = 5000$, $\bar{s} = 100$, and $\lambda_{\rm tgt} = 1$. Entries of $A \in \mathbb{R}^{m\times n}$ were generated with independent uniform distributions over $[-1, +1]$, and $\|z\|_\infty = 0.01$. Algorithmic parameters: $\gamma_{\rm inc} = 2$, $\gamma_{\rm dec} = 2$, $\eta = 0.7$, and $\delta = 0.2$.

the slow convergence phase of PG is associated with relatively dense iterates (with $\|x^{(k)}\|_0$ ranging from 5,000 to several hundreds), while the fast linear convergence in the end coincides with sparse iterates with $\|x^{(k)}\|_0$ around one hundred. In contrast, the PGH method maintains sparse iterates (always less than 300) along the whole solution path, and demonstrates geometric convergence at each stage of homotopy continuation.

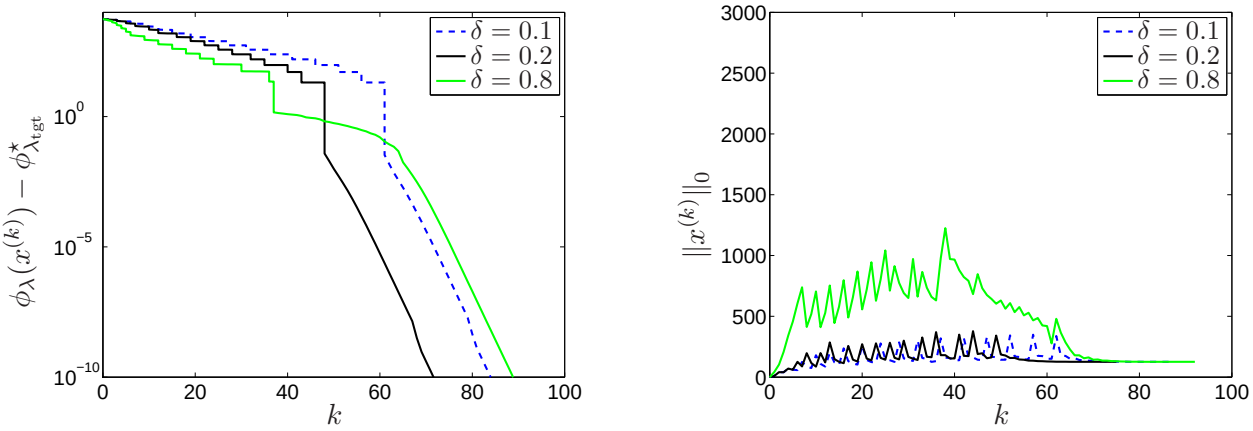
Figure 1(c) shows the optimality residues of different methods versus the number of iterations k. They demonstrate similar trends as the objective function gap, but clearly they oscillate along the solution path and do not decrease monotonically. Figure 1(d) plots the local Lipschitz constants returned by the line search procedure at each iteration. We see that the adaptive line-search method settles with much smaller M_k when the iterates are sparse. There is a striking similarity between the final stages of the PG method and the PGH method. However, the PGH method avoids the slow sublinear convergence by maintaining sparse iterates along its whole solution path.

Also plotted in Figure 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 [Nes07]). In the method ADGH, we combine the homotopy continuation strategy with the ADG method. It improves a lot compared with ADG, but still does not have linear convergence and thus is much slower than the PGH method.

Figure 1(e) shows the number of proximal-gradient steps performed at each homotopy 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 $\epsilon = 10^{-5}$, 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 2. The ADGH method, which employs the ADG method for solving each stage, took more number of inner iterations at each stage. This again reflects its lack of capability of exploiting the restricted strong convexity.

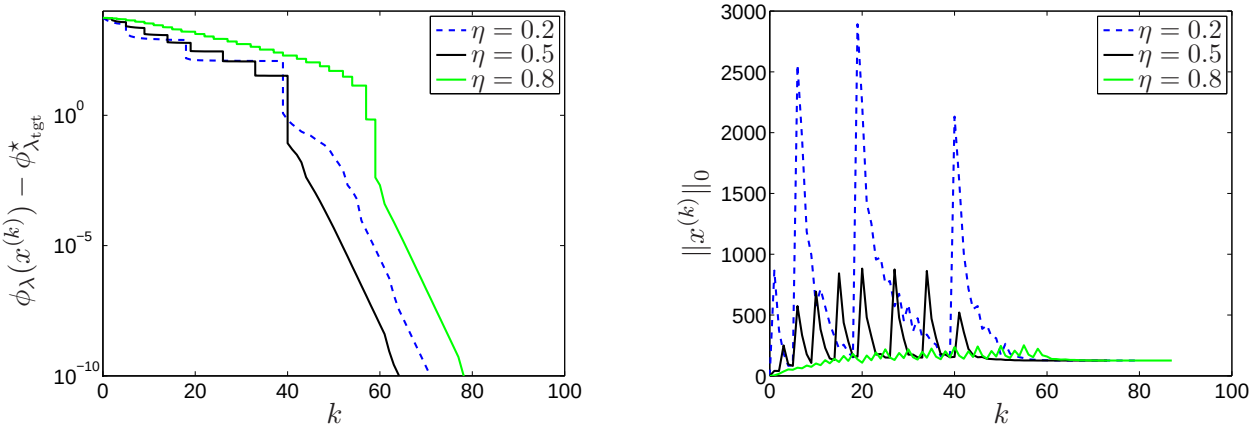
The number of inner iterations is not the whole story for evaluating the performance of the algorithms. Figure 1(f) shows the objective gap versus the total number of matrix-vector multiplications with either A or A^T. Evaluating the objective function $f(x^{(k)})$ costs one matrix-vector multiplication, and evaluating the gradient $\nabla f(x^{(k)})$ costs an additional multiplication. The estimate in (15) states that each proximal-gradient 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 [Nes07]. These factors are confirmed by comparing the horizontal scales of the Figures 1(a) and 1(f). 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 2 shows the objective gap and sparsity of the iterates along the solution path for different δ while keeping $\eta = 0.7$. 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



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

Figure 2: Performance of the PGH method by varying δ while keeping $\eta = 0.7$.

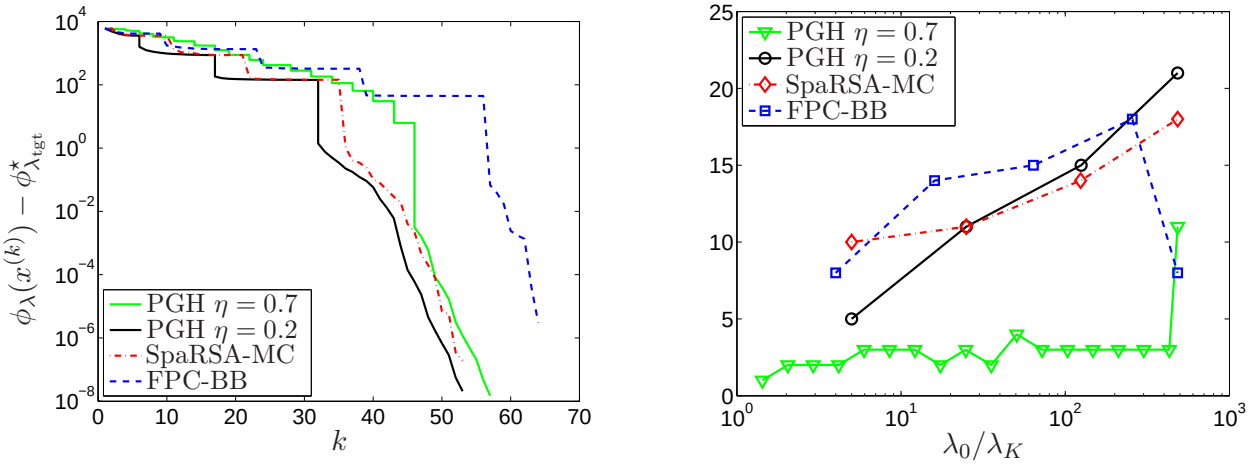


(a) Objective gap. (b) Sparsity along solution path.

Figure 3: Performance of the PGH method by varying η while keeping $\delta = 0.2$.

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 3 shows the numerical behaviors of the PGH method by 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. It is interesting to note that the sufficient conditions $(1 + \delta)/(1 + \delta') \le \eta < 1$ (in Theorem 2) and $0 < \delta < \delta' < 1$ implies $\eta > 0.5$. But we see that a more aggressive $\eta = 0.2$ still works well for this instance.



(a) Objective gap.

(b) Number of inner iterations for each λ_K.

Figure 4: Comparison with SpaRSA and FPC.

5.1 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 two most relevant ones: sparse reconstruction by separable approximation (SpaRSA) [WNF09], and fixed point continuation (FPC) [HYZ08]. In particular, the same proximal gradient method (5) is used in each iteration of both SpaRSA and FPC. Their continuation strategies are both based on reducing λ by a constant factor at each stage.

SpaRSA uses Barzilai-Borwein (spectral) method for choosing L_k at each step. More specifically, at each iteration the parameter L_k is initialized as

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

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 $|\mathrm{supp}(x^{(k)}) \cup \mathrm{supp}(x^{(k-1)})| \le s$ 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 as 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 by the authors of [HYZ08] that also employs Barzilai-Borwein line search, which is called FPC-BB. In fact FPC-BB solves the equivalent problem

$$\underset{x}{\text{minimize}} \quad \|x\|_1 + \frac{\mu}{2}\|Ax - b\|_2^2$$

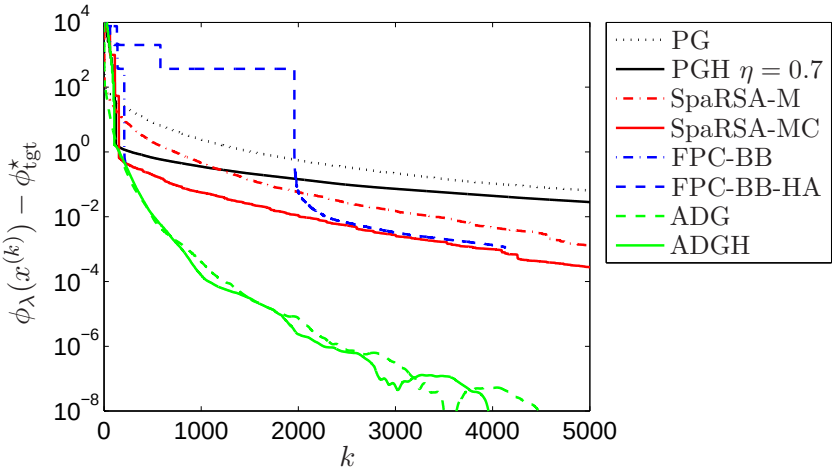
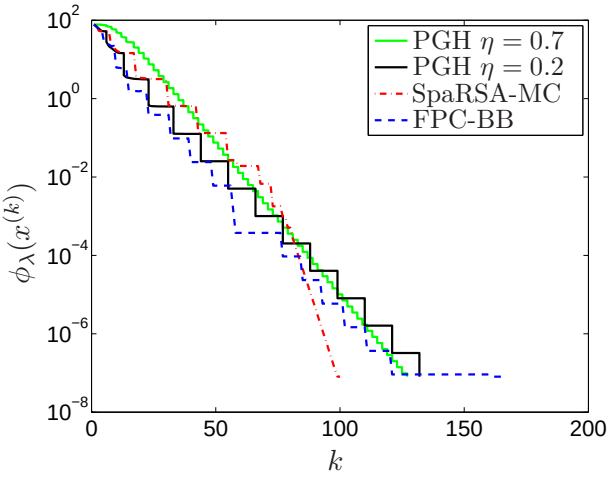


Figure 5: Comparison of different methods for solving a non-sparse random instance.

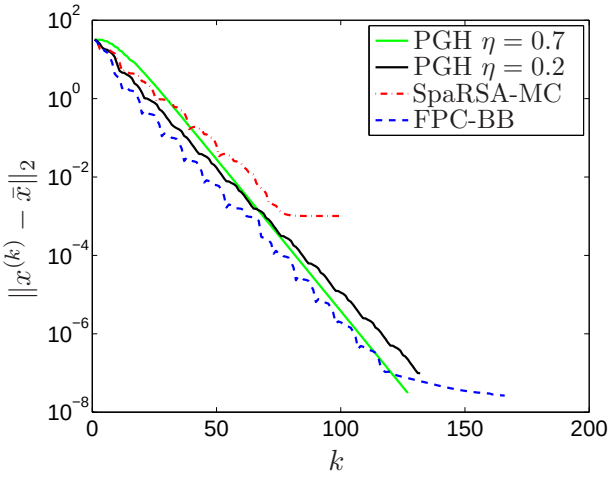
where $\mu = 1/\lambda$. Moreover, it further scales the matrix A so that the maximum singular value is at most 1. In Figures 4 and 5, the results of FPC-BB are plotted after we reversed the scalings in order to compare with other methods. Default options were used in both methods. SpaRSA-MC reduces the value of λ roughly with an 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 relative precision $\delta = 0.2$ was used in both cases for PGH.

Figure 4 shows the numerical results of different algorithms on the same random instance studied in Figure 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 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, which is also reflected in Figure 4(a). According to Figure 3(b), the aggressive decreasing factors η used in SpaRSA and FPC can lead to less sparse iterates along the solution path, thus relatively slower convergence at the intermediate stages. But their overall iteration counts are comparable to PGH with $\eta = 0.7$.

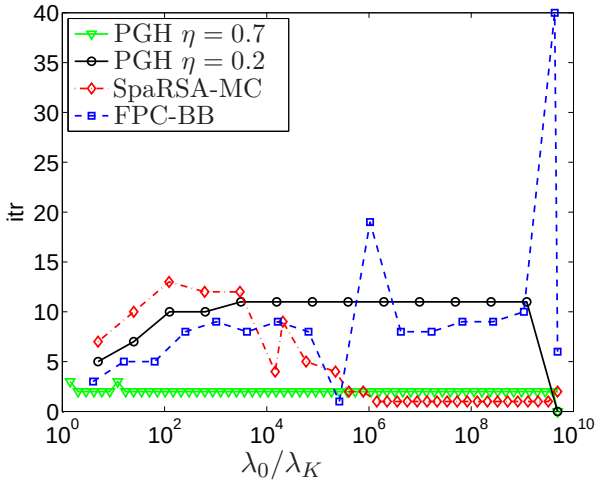
We also conducted experiments with random problem instances where the vector $\bar{x}$ is not sufficiently sparse. Figure 5 shows the objective gap of different methods when solving a random problem instance generated similarly as the one studied in Figure 1. The only difference is that here the vector $\bar{x}$ has 500 nonzero elements. In this case, all methods demonstrate sub-linear convergence. SpaRSA-M is the monotone version of SpaRSA without continuation. FPC-BB terminated prematurely because its default accuracy for its stopping criterion is too low. FPC-BB-HA is the result after we set a much higher accuracy in calling the FPC-BB method. It looks that the same higher accuracy is used in all the homotopy stages, so the number of inner iterations increased for each stage. 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.



(a) Objective value. Note that $\phi_\lambda^\star \to 0$ as $\lambda \to 0$.



(b) Recovery error.



(c) Number of inner iterations

Figure 6: Basis pursuit via homotopy continuation: an example with partial FFT matrix.

5.2 Basis pursuit

Finally we present an experiment of solving the basis pursuit (BP) problem (8) using PGH, and compare it with FPC and SpaRSA. In this experiment, the matrix A is a partial FFT matrix. More specifically, we choose $m = 10,000$ rows at random from the $n \times n$ FFT matrix with $n = 2^{16} = 65536$. The vector $\bar{x} \in \mathbb{R}^n$ has nonzero entries at only $\bar{s} = 1000$ randomly chosen coordinates, and they were generated independently from the normal distribution with zero mean and unit variance. Then we set $b = A\bar{x}$ in the BP problem (i.e., this is the noise-free case with $z = 0$).

In this case, since A is a matrix with complex numbers, we need to replace all the real transpose in the algorithms with Hermitian transpose, and replace the soft-thresholding operator in (7) with

$$\mathrm{soft}(x_i, \alpha) = \frac{\max\{|x_i| - \alpha, 0\}}{\max\{|x_i| - \alpha, 0\} + \alpha},$$

where $|x_i|$ denotes the modulus of the complex number x_i [WNF09].

The solution to the BP problem (8) can be obtained by letting $\lambda \to 0$ in the ℓ_1-LS problem (1). In order to use the PGH method, we set $\lambda_{\rm tgt} = 10^{-10}$. The same parameter was also used in calling SpaRSA-MC and FPC-BB. Figure 6 shows the numerical results. Again we observe remarkable resemblance between these methods in Figure 6(a). However, in Figure 6(b), we see the recovery error of SpaRSA-MC stayed at the level 10^{-3} while its objective function in Figure 6(a) converged to zero faster than other methods. The reason is that SpaRSA has a fixed accuracy requirement for all continuation stages except for the last one. As shown in Figure 6(c), when λ_K becomes very small, this constant accuracy is always reached within one iteration, and such a low accuracy is too loose for the algorithm to track the homotopy path closely. Therefore, even though the objective function converges to zero quickly, the recovery error stayed large. This is also confirmed through the denser continuation stages in the second half of SpaRSA-MC, as shown in Figure 6(c). To see this, we note that the adaptive continuation used in SpaRSA is

$$\lambda_{K+1} = \max\left\{\eta\|A^T(A\hat{x}^{(K)} - b)\|_\infty,\ \lambda_{\rm tgt}\right\}.$$

If $\hat{x}^{(K)}$ is an accurate solution for the stage λ_K, then we have $\|A^T(A\hat{x}^{(K)} - b)\|_\infty \approx \lambda_K$ and thus $\lambda_{K+1} \approx \eta\lambda_K$ with $\eta = 0.2$ as the default value. When this is not the case, then $\|A^T(A\hat{x}^{(K)} - b)\|_\infty$ can be notably larger than λ_K, and thus the regularization parameter reduces at a much slower pace. Similar as PGH, FPC-BB sets the accuracy for each continuation stage to be proportional to the regularization parameter, but for a different stopping criterion. With our choice of stopping criterion, $\omega_\lambda(\hat{x}^{(K)}) \le \delta\lambda_K$, the number of inner iterations for each continuation stage stayed roughly constant along the homotopy path.

This example also demonstrates the advantage of PGH and other approximate homotopy continuation methods over the exact homotopy path-following methods [OPT00a, OPT00b, EHJT04]. Figure 6(b) shows that high-precision recovery can be obtained by PGH in less than 150 iterations (which corresponds to roughly 450 matrix-vector multiplications). This is much more efficient than using the exact homotopy path-following methods, which need to track at least 1000 breakpoints. In addition, their computational cost at each break point is much higher than a matrix-vector multiplication.

6 Conclusion and discussions

This paper studied a proximal-gradient homotopy 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 proximal gradient methods can only obtain relatively slow convergence rate. However, we have shown that under suitable conditions for sparse recovery, all iterates of the proximal-gradient homotopy 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 are supported by several numerical experiments.

We commented in the numerical experiments that accelerated gradient methods cannot automatically exploit restricted strong convexity. As discussed in [Nes04, Section 2.2] and [Nes07], they need to explicitly use the strong convexity parameter, or a non-trivial 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 [Nes07], and strategies such as periodic restart have been studied recently [GLW09, BCG11]. However, an in-depth investigation on this matter is beyond the scope of this paper.

References

[ANW11] A. Agarwal, S. N. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. Technical Report arXiv:1104.4824v1, arXiv, 2011.

[BBC11] S. R. Becker, J. Bobin, and E. J. Candès. NESTA: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

[BCG11] S. R. Becker, E. J. Candès, and M. C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011.

[BDE09] A. M. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, 2009.

[BDF07] J. M. Bioucas-Dias and M. A. T. Figueiredo. A new TwIST: Two-step iterative shrinking/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing*, 16(12):2992–3004, 2007.

[BRT09] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

[BT09] A. Beck and M. Teboulle. A fast iterative shrinkage-threshold algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[CDS98] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[CRT06] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, February 2006.

[CT05] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, December 2005.

[CT06] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, December 2006.

[CT07] E. J. Candes and T. Tao. The dantzig selector: statistical estimation when p is much larger than n (with discussion). *Annals of Statistics*, 35:2313–2404, 2007.

[CW05] P. Combettes and V. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Journal on Multiscale Modeling and Simulation*, 4(4):1168–1200, 2005.

[DDM04] I. Daubechies, M. Defriese, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

[DET06] D. L. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, January 2006.

[DMA97] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Journal of Constructive Approximation*, 13(1):57–98, 1997.

[Don06] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, April 2006.

[DT08] D. L. Donoho and Y. Tsaig. Fast solution of ℓ_1-norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, 54(11):4789–4812, November 2008.

[EHJT04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression (with discussion). *Annals of Statistics*, 32:407–499, 2004.

[GLW09] 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.

[HYZ08] E. T. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.

[KKL+07] 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 Journal on Selected Topics in Signal Processing*, 1(4):606–617, December 2007.

[Kol09] V. Koltchinskii. The dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15:799–828, 2009.

[LT92] Z.-Q. Luo and P. Tseng. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30(2):408–425, 1992.

[MB06] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[Nes83] Yu. Nesterov. A method for solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics - Doklady*, 27(2):372–376, 1983.

[Nes96] Yu. Nesterov. Long-step strategies in interior-point primal-dual methods. *Mathematical Programming*, 76:47–94, 1996.

[Nes04] Yu. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, Boston, 2004.

[Nes05] Yu. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[Nes07] Yu. Nesterov. Gradient methods for minimizing composite objective function. CORE discussion paper 2007/76, Center for Operations Research and Econometrics, Catholic University of Louvain, Belgium, September 2007.

[NT09] D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[OPT00a] M. Osborne, B. Presnell, and B. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20:389–404, 2000.

[OPT00b] M. Osborne, B. Presnell, and B. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

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

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58:267–288, 1996.

[Tro04] J. A. Tropp. Greedy is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[Tro06] J. A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52:1030–1051, 2006.

[Tse08] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Manuscript submitted to *SIAM Journal on Optimization*, 2008.

[TVW05] B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47:349–363, 2005.

[TW10] J. A. Tropp and S. J. Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, 2010.

[vdBF08] E. van den Berg and M. P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):980–912, 2008.

[vdGB09] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[Wai09] M. J. Wainwright. Sharp thresholds for noisy and high–dimensional recovery of sparsity using ℓ_1–constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

[WNF09] S. J. Wright, R. D. Nowad, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, July 2009.

[WYGZ10] Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang. A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization and continuation. *SIAM Journal on Scientific Computing*, 32(4):1832–1857, 2010.

[YOGD08] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Science*, 1(1):143–168, 2008.

[ZH08] C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high–dimensional linear regression. *Annals of Statistics*, 36:1567–1594, 2008.

[Zha09] T. Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *Annals of Statistics*, 37:2109–2144, 2009.

[ZY06] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.