A Proximal-Gradient Homotopy Method for the ℓ_1-Regularized Least-Squares Problem

Lin Xiao LIN.XIAO@MICROSOFT.COM
Microsoft Research, 1 Microsoft Way, Redmond, WA 98052 USA

Tong Zhang TZHANG@STAT.RUTGERS.EDU
Department of Statistics, Rutgers University, Piscataway, NJ 08854 USA

.

Abstract

We consider the ℓ_1-regularized least-squares problem for sparse recovery and compressed sensing. Since the objective function is not strongly convex, standard proximal gradient methods only achieve sublinear convergence. We propose a homotopy continuation strategy, which employs a proximal gradient method to solve the problem with a sequence of decreasing regularization parameters. It is shown that under common assumptions in compressed sensing, the proposed method ensures that all iterates along the homotopy solution path are sparse, and the objective function is effectively strongly convex along the solution path. This observation allows us to obtain a global geometric convergence rate for the procedure. Empirical results are presented to support our theoretical analysis.

1. Introduction

This paper proposes and analyzes 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.

The ℓ_1-LS problem has important applications in machine learning, signal processing, and statistics. It has received significant attention in recent years due to the emergence of *compressed sensing* theory, which builds upon the fundamental idea that a finite-dimensional signal having a sparse or compressible representation can be recovered from a small set of linear, nonadaptive measurements. 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 consider 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. 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 entries in $\bar{x}$. For example, see (Meinshausen & Bühlmann, 2006; Zhang & Huang, 2008; Zhang, 2009; Bickel et al., 2009; Koltchinskii, 2009; van de Geer & Bühlmann, 2009). 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 (Zhang & Huang, 2008).

In this paper, we develop an efficient numerical method for solving the ℓ_1-LS problem in the context of sparse

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

recovery described above. In particular, we focus on the case when $m < n$ (i.e., the linear system $Ax = b$ is underdetermined) and the solution $x^\star(\lambda)$ is sparse (which requires the parameter λ to be sufficiently large). Under such assumptions, our method has provable lower complexity than previous algorithms.

1.1. Previous algorithms

There have been extensive research on numerical methods for solving (1) and its several variations. Here we briefly summarize the computational complexities of several methods that are most relevant to our approach, in terms of finding an ϵ-optimal solution.

Proximal gradient methods take the following basic form at each iteration $k = 0, 1, \ldots$

$$x^{(k+1)} = \arg\min_y \Big\{ f(x^{(k)}) + \nabla f(x^{(k)})^T (y - x^{(k)}) + \frac{L_k}{2}\|y - x^{(k)}\|_2^2 + \lambda\|y\|_1 \Big\}, \quad (3)$$

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 (3) has a closed-form solution

$$x^{(k+1)} = \mathrm{softh}\left(x^{(k)} - \frac{1}{L_k}\nabla f(x^{(k)}) \,,\ \frac{\lambda}{L_k} \right), \quad (4)$$

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

$$(\mathrm{softh}(x, \alpha))_i = \mathrm{sgn}(x_i) \max\{|x_i| - \alpha, 0\}\,,\ i = 1, \ldots, n.$$

Iterative methods that use the update rule (4) include Daubechies et al. (2004); Nesterov (2007); Hale et al. (2008); Wright et al. (2009). 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 (3) has an iteration complexity $O(1/\epsilon)$.

Variations and extensions of the proximal gradient method have been proposed to speed up the convergence in practice (Bioucas-Dias & Figueiredo, 2007; Wright et al., 2009; Wen et al., 2010). Nesterov's optimal gradient methods for minimizing smooth convex functions (Nesterov, 2004; 2005) have also been extended to minimize composite objective functions such as in the ℓ_1-LS problem (Nesterov, 2007; Tseng, 2008; Becker et al., 2011). 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 (Osborne et al., 2000; Efron et al., 2004). These methods exploit the piecewise linearity of the solution as a function of λ, and identify the next breakpoint along the solution path by examining the optimality conditions (also called *active set* or *pivoting* method in optimization). With efficient numerical implementations (using updating or downdating of submatrix factorizations), the computational cost at each breakpoint is $O(mn + ms^2)$, where s is the number of nonzeros in the solution. 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.

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 (3) 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.

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., Hale et al. (2008); Wright et al. (2009); Wen et al. (2010), 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.

In this paper, we present a PGH method that has provable low iteration complexity. Under the assumptions that the target value of λ is sufficiently large (so that the final solution is sparse) and the matrix A satisfies a RIP-like condition, our PGH method exhibits a global geometric rate of convergence, with an overall computational complexity of $O(mn\log(1/\epsilon))$ in order to achieve accuracy ϵ. Moreover, it is sufficient to choose $\lambda \geq c\|A^T z\|_\infty$ (for some universal constant c), which implies that the final solution satisfies a recovery bound of the optimal form (2).

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. Under a RIP-like assumption on A, the sparsity implies that along the homotopy path, the objective function in (1) is effectively strongly convex, and hence global geometric rate can be established using the analysis of proximal gradient methods by Nesterov (2007).

2. Preliminaries and notations

In this section, we first introduce composite gradient mapping and some of its key properties developed by Nesterov (2007). We then 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 geometric convergence of Nesterov's algorithm.

2.1. Composite gradient mapping

To simplify presentation, let $f(x) = (1/2)\|Ax - b\|_2^2$. Then the ℓ_1-LS problem can be written as

$$\underset{x}{\text{minimize}} \quad \left\{ \phi_\lambda(x) \triangleq f(x) + \lambda\|x\|_1 \right\}, \qquad (5)$$

The optimality condition of (5) states that $x^\star$ is a solution if and only if there exists $\xi \in \partial\|x^\star\|_1)$ such that $\nabla f(x^\star) + \lambda\xi = 0$ (see, e.g., Rockafellar, 1970, Section 27). Therefore, a good measure of the quality of an approximate solution is the quantity

$$\omega_\lambda(x) \triangleq \min_{\xi\in\partial\|x\|_1} \|\nabla f(x) + \lambda\xi\|_\infty. \qquad (6)$$

We call $\omega_\lambda(x)$ the *optimality residue* of x. Given the gradient $\nabla f(x)$, it can be computed with $O(n)$ flops. We will use it in the stopping criterion of our method.

Composite gradient mapping was introduced by Nesterov (2007). For any fixed point y and a given constant $L > 0$, we define a local model of $\phi_\lambda(x)$ around y using a quadratic approximation of f:

$$\psi_{\lambda,L}(y;x) = f(y) + \nabla f(y)^T(x-y) + \frac{L}{2}\|x-y\|_2^2 + \lambda\|x\|_1,$$

and let

$$T_{\lambda,L}(y) = \arg\min_x \ \psi_{\lambda,L}(y;x). \qquad (7)$$

We note that $T_{\lambda,L}(x)$ can be computed using (4). The *composite gradient mapping* of ϕ_λ at y is defined as

$$g_{\lambda,L}(y) = L(y - T_{\lambda,L}(y)).$$

The following property of composite gradient mapping was shown in Nesterov (2007, Theorem 2):

Lemma 1. *For any $y \in \mathbb{R}^n$ and any $L > 0$,*

$$\psi_{\lambda,L}(y; T_{\lambda,L}(y)) \le \phi_\lambda(y) - \frac{1}{2L}\|g_{\lambda,L}(y)\|_2^2.$$

Algorithm 1 $\{x^+, M\} \leftarrow$ `LineSearch`(λ, 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 $\{\hat{x}, \hat{M}\} \leftarrow$ `ProxGrad`$(\lambda, \hat{\epsilon}, x^{(0)}, L_0)$

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

2.2. Nesterov's proximal gradient method with adaptive line-search

With the machinery of composite gradient mapping, Nesterov developed several variants of proximal gradient methods (2007). We use the non-accelerated primal-gradient version described in Algorithms 1 and 2, which correspond to (3.1) and (3.2) in Nesterov (2007), respectively. To use this algorithm, we need to first choose an initial optimistic estimate $L_{\min}$ for the Lipschitz constant L_f (i.e., $0 < L_{\min} \le L_f$) and two adjustment parameters $\gamma_{\text{dec}} \ge 1$ and $\gamma_{\text{inc}} > 1$.

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

$$\begin{aligned}\phi_\lambda(x^{(k+1)}) &\le \psi_{\lambda,M_k}\left(x^{(k)}, x^{(k+1)}\right) \\ &\le \phi_\lambda(x^{(k)}) - \frac{1}{2M_k}\left\|g_{\lambda,M_k}(x^{(k)})\right\|_2^2, \end{aligned} \qquad (8)$$

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, $x^{(k+1)}$ is an optimal solution.

Nesterov established the following iteration complexities of Algorithm 2 for finding an ϵ-optimal solution: if ϕ_λ is convex but not strongly convex, then the convergence is sublinear, with an iteration complexity $O(1/\epsilon)$ (Nesterov, 2007, Theorem 4); if ϕ_λ is strongly convex, then the convergence is geometric, with an iteration complexity $O(\log(1/\epsilon))$ (Nesterov, 2007, 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.

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 $O(1/\epsilon)$ sublinear convergence. Nevertheless, 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 (which we explain next), 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.

2.3. Restricted eigenvalue conditions

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

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

Throughout the paper, we denote $\operatorname{supp}(\bar{x})$ by $\bar{S}$, and use $\bar{S}^c$ for its complement. We use the 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 Candès & Tao (2005). 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*

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

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 for
$\{\hat{x}^{(\mathrm{tgt})}, \hat{M}_{\mathrm{tgt}}\} \leftarrow \texttt{ProxGrad}\big(\lambda_{\mathrm{tgt}}, \epsilon, \hat{x}^{(N)}, \hat{M}_N\big)$
return $\hat{x}^{(\mathrm{tgt})}$
```

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.

The convergence rate of our PGH method depends on a *restricted condition number*, defined as

$$\kappa(A,s) = \frac{\rho_+(A,s)}{\rho_-(A,s)}. \tag{9}$$

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 a 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 PG 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 $\lambda_0 = \|A^T b\|_\infty$, which 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 additional parameters $\eta \in (0,1)$ and $\delta \in (0,1)$. They control the algorithm as follows:

- The sequence of values for the regularization parameter is determined as $\lambda_K = \eta^K \lambda_0$ for $K = 1, 2, \ldots$, until the target value $\lambda_{\rm tgt}$ is reached.

- For each λ_K except $\lambda_{\rm tgt}$, we solve problem (1) with a proportional precision $\delta\lambda_K$. For the last stage $\lambda_{\rm tgt}$, 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*

$$\lambda_{\rm tgt} \geq \max\left\{4,\ \frac{\gamma+1}{(1-\delta')\gamma-(1+\delta')}\right\}\|A^T z\|_\infty. \quad (10)$$

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

$$\tilde{s} > \frac{16\big(\gamma_{\rm inc}\rho_+(A,\bar{s}+2\tilde{s})+2\rho_+(A,\tilde{s})\big)}{\rho_-(A,\bar{s}+\tilde{s})}(1+\gamma)\bar{s}. \quad (11)$$

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

In this paper, we show that by choosing the parameters η and δ in Algorithm 3 appropriately, these conditions imply that all iterates along the solution path are sparse. Our proof employs a similar argument as that of Zhang & Huang (2008). Before stating the main convergence results, we make some further remarks on Assumption 1.

- The condition (10) 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 $\|A^T z\|_\infty$. 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 (11) is necessary and standard in sparse recovery analysis. This is closely related to the RIP condition of Candès & Tao (2005) 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 $\gamma_{\rm inc} = 2$ and $\tilde{s} = 96(1+\gamma)\bar{s}$ so that the condition (11) is satisfied. 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.

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*

$$\big\|x^{(0)}_{\bar{S}^c}\big\|_0 \leq \tilde{s}, \qquad \omega_\lambda(x^{(0)}) \leq \delta'\lambda, \quad (12)$$

then for all $k \geq 0$, we have $\big\|x^{(k)}_{\bar{S}^c}\big\|_0 \leq \tilde{s}$, and

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

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

Due to space limit, all proofs for results presented in this paper are given in the supplementary material.

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 $\lambda_{\rm tgt} \leq \lambda_0$ and the parameters δ and η are chosen such that $(1+\delta)/(1+\delta') \leq \eta < 1$. Let $N = \big\lfloor \ln(\lambda_0/\lambda_{\rm tgt}) / \ln \eta^{-1} \big\rfloor$, then*

1. *The condition (12) 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*

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

   where $C = 8\gamma_{\rm inc}(1+\kappa)^2(1+\gamma)\kappa\bar{s}$.

2. *For $K = 0, \ldots, N-1$, the outer-loop iterates $\hat{x}^{(K)}$ satisfies*

   $$\phi_{\lambda_{\rm tgt}}(\hat{x}^{(K)}) - \phi^\star_{\lambda_{\rm tgt}} \leq \eta^{2(K+1)}\frac{4.5(1+\gamma)\lambda_0^2\bar{s}}{\rho_-(A,\bar{s}+\tilde{s})}, \quad (13)$$

and we have the sparse recovery bound

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

3. *When Algorithm 3 terminates, the total number of proximal-gradient steps is no more than*

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

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

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

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

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

  From the inequality (13), we see that if $0 \leq K_0 \leq N-1$, then for all $K \geq K_0$,

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

  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 $\omega_\lambda(\cdot) \leq \epsilon_0$. 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

$$\ln\left(1-\frac{1}{4\gamma_{\rm inc}\kappa}\right)^{-1} \approx \frac{1}{4\gamma_{\rm inc}\kappa}.$$

  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 $\hat{x}^{(\rm tgt)}$) also converge to $x^\star(\lambda_{\rm tgt})$ at a globally geometric rate. A particularly interesting case is noise-free compressed sensing using the *basis pursuit* formulation

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

which has the optimal solution $\bar{x}$. For this problem, we can simply run Algorithm 3 with $\lambda_{\rm tgt} = 0$. While the convergence metrics such as objective value gap or optimality residue are no longer informative in this case, Theorem 2 implies the following numerical convergence result after K iterations:

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

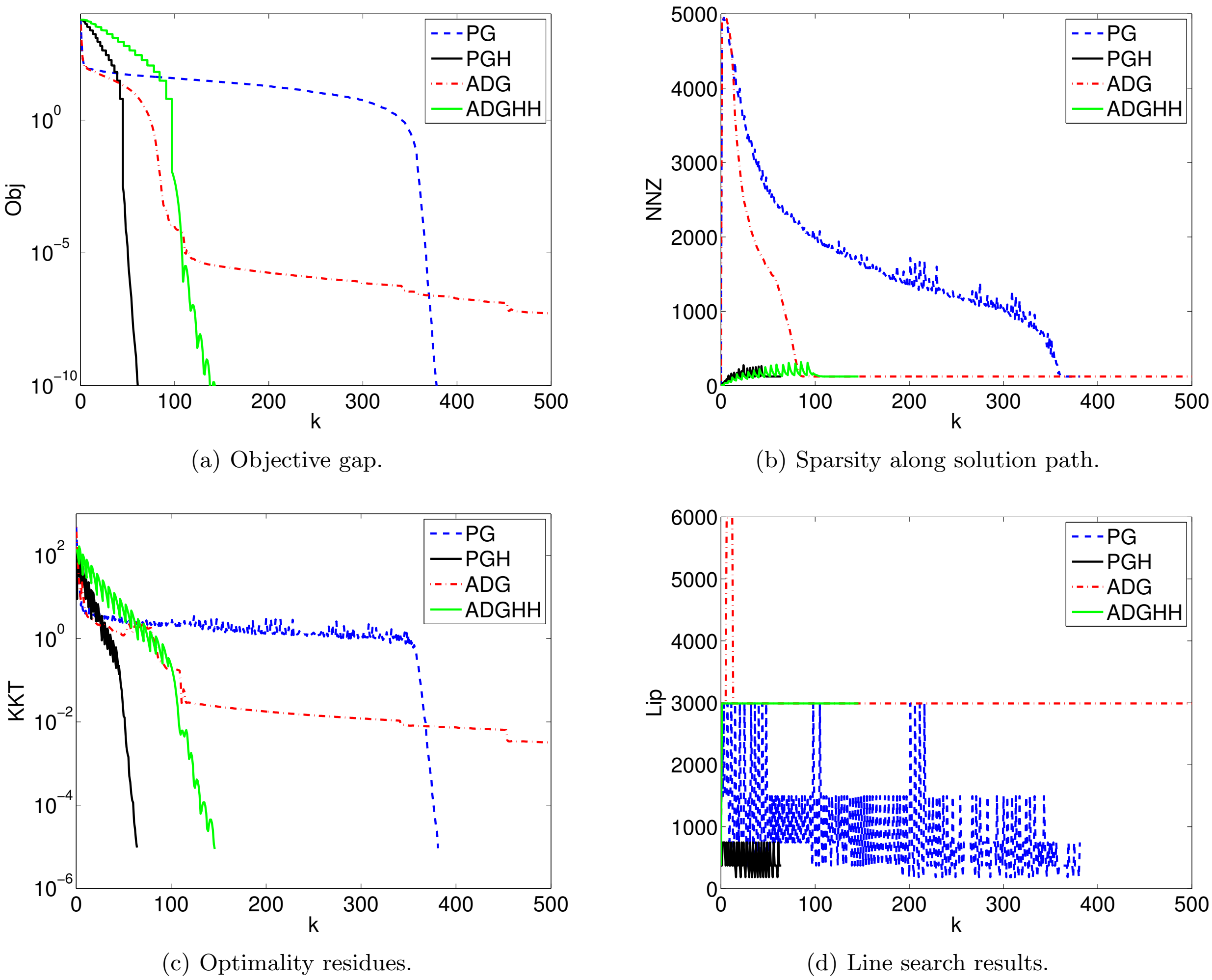
This requires no more than $O(K)$ proximal-gradient steps in the PGH method. This result can be interpreted as a global geometric rate of convergence for solving the BP problem.

4. Numerical experiments

In this section, we illustrate the numerical properties of our 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 in Algorithm 3.
- ADG: Nesterov's accelerated dual gradient method Nesterov (2007, Algorithm (4.9)).
- ADGH: the method outlined in Algorithm 3, but with PG replaced by ADG for the inner loop.

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 (all other coordinates of $\bar{x}$ was set to zero). The noise $z \in \mathbb{R}^m$ is a dense vector with independent random entries with the uniform distribution over the interval $[-0.01, 0.01]$. Finally the vector b was obtained as $b = A\bar{x} + z$. In our experiment, we choose $\lambda_{\rm tgt} = 1$. For this particular instance we have roughly $\|A^T z\|_\infty \approx 0.411$. The PGH method started with $\lambda_0 = \|A^T b\|_\infty = 483.5$.



(a) Objective gap.

(b) Sparsity along solution path.

(c) Optimality residues.

(d) Line search results.

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

Figure 1 illustrates various numerical properties of the four different methods for solving this random instance. We used the parameters $\gamma_{\mathrm{inc}} = 2$ and $\gamma_{\mathrm{dec}} = 2$ in all four methods. For the two homotopy methods (whose acronyms end with the letter H), we used the parameters $\eta = 0.7$ and $\delta = 0.2$. In all four subfigures, 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 Figures 1(a) and 1(c) indicate switchings of homotopy stages (when the value of λ is reduced by the factor η) — they reflect the change of objective function for the same vector $x^{(k)}$.

Figure 1(a) shows the objective gap $\phi_\lambda(x^{(k)}) - \phi^\star_{\lambda_{\mathrm{tgt}}}$ versus the 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 the slow convergence phase of PG is associated with relatively dense iterates (with NNZs ranging from 5,000 to a few hundreds), while the fast linear convergence in the end coincides with sparse iterates with $\|x^{(k)}\|_0$ around 100. 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 much faster 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 lags behind when the solution is very sparse. 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 geometric convergence and thus is much slower than the PGH method.

References

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

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

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

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

Daubechies, I., Defriese, M., and Mol, C. De. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

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

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

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

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

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

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

Nesterov, Yu. 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.

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

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

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

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

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

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

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

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