Ridge-Adding Homotopy Approach for $l_1$-norm Minimization Problems*

**Haoran LI**†, Student Member, Binyu WANG†, Jisheng DAI†, and Tianhong PAN††, Nonmembers

**SUMMARY** Homotopy algorithm provides a very powerful approach to select the best regularization term for the $l_1$-norm minimization problem, but it is lack of provision for handling singularities. The singularity problem might be frequently encountered in practical implementations if the measurement matrix contains duplicate columns, approximate columns or columns with linear dependent in kernel space. The existing method for handling Homotopy singularities introduces a high-dimensional random ridge term into the measurement matrix, which has at least two shortcomings: 1) it is very difficult to choose a proper ridge term that applies to several different measurement matrices; and 2) the high-dimensional ridge term may accumulatively degrade the recovery performance for large-scale applications. To get around these shortcomings, a modified ridge-adding method is proposed to deal with the singularity problem, which introduces a low-dimensional random ridge vector into the $l_1$-norm minimization problem directly. Our method provides a much simpler implementation, and it can alleviate the degradation caused by the ridge term because the dimension of ridge term in the proposed method is much smaller than the original one. Moreover, the proposed method can be further extended to handle the SVMpath initialization singularities. Theoretical analysis and experimental results validate the performance of the proposed method.

**key words:** $l_1$-norm minimization, homotopy algorithm, singularity, support vector machine (SVM)

1. Introduction

With the rapid development of the compressed sensing theory [1], [2], $l_1$-norm minimization has become a popular tool for underdetermined linear regression in signal processing and machine learning communities [3], [4]. For a typical underdetermined linear regression problem, the measurement vector $y \in \mathbb{R}^{P \times 1}$ is generated from an unknown signal $x_0 \in \mathbb{R}^{N \times 1}$ by a specific linear transformation:

$$y = Ax_0 + n,$$  \hspace{1cm} (1)

where $A \in \mathbb{R}^{P \times N} (P < N)$ is a known measurement matrix and $n \in \mathbb{R}^{P \times 1}$ is a noise vector. Generally, the underdetermined problem (1) has numberless solutions. However, if $x_0$ is sparse, a unique solution can be obtained to approximate the unknown sparse signal $x_0$ by solving the following $l_1$-norm minimization problem:

$$\min_x ||x||_1 \text{ s.t. } ||y - Ax||_2^2 \leq \epsilon,$$  \hspace{1cm} (2)

where $\epsilon$ is a nonnegative constant. The $l_1$-norm minimization problem is exactly equal to the so-called LASSO problem [5]:

$$\min_x ||y - Ax||_2^2 \text{ s.t. } ||x||_1 \leq t,$$  \hspace{1cm} (3)

if the parameter $t$ is appropriately chosen [3]. Moreover, the alternative unconstrained $l_1$-penalised least-square problem instead of (2) or (3) is as follows:

$$\min_x \frac{1}{2} ||y - Ax||_2^2 + \lambda||x||_1,$$  \hspace{1cm} (4)

where $\lambda \geq 0$ is the regularization term [6]. The sparsity encouraging property of $l_1$-norm minimization has been well established in the literature, which has been successfully applied to many fields, e.g., feature selection [5], face recognition [4], array signal processing [7], and channel estimation [8], [9]. There exist many efficient and robust algorithms for solving the $l_1$-norm minimization problem, where a default value for the regularization term (or $\epsilon$ or $t$) is commonly used though it is not the optimal choice. Actually, the regularization term is critical to the performance of $l_1$-norm minimization. However, it is intractable to exhaustively solve the $l_1$-norm minimization problem for every possible regularization terms. Fortunately, the Homotopy algorithm [10] provides a very powerful and interesting method to select the best regularization term. It sets a large $\lambda$ and then calculates the solution for every value of $\lambda$ in a piecewise linear manner, so as to obtain the whole solution path of (4) with respect to $\lambda$. If the true signal is sufficiently sparse, the optimal solution to (4) can be obtained with a few iterations.

The Homotopy algorithm is an exact active-set method for the regularization term selection, which has also been extended to other common problems, such as support vector regression (SVR) [11], support vector machine (SVM) [12]–[14] and so on. Nevertheless, the active-set-typed methods are lack of provision for handling singularities. When the measurement matrix contains duplicate columns, approximate columns or columns with linear dependent in kernel space, the singularity problem might occur, which results in the failure of the Homotopy algorithm. To the best of

---

*Manuscript received October 12, 2019.
Manuscript revised January 29, 2020.
Manuscript publicized March 10, 2020.
†The authors are with the School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China.
††The authors are with School of Electrical Engineering and Automation, Anhui University, Hefei, Anhui 230061, China.
This work is supported by the National Key R&D Program under Grant of China (Grant number 2017YFF0211301), Key R&D Program of Jiangsu Province, China (Grant number BE2018370), and Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX17_1779).
a) E-mail: haoranli.ujs@foxmail.com
DOI: 10.1587/transinf.2019EDP7277

Copyright © 2020 The Institute of Electronics, Information and Communication Engineers
our knowledge, the Homotopy singularity problem was first addressed in [6]. Following the ridge-adding method for SVMpath singularities in [14], Dai et al. [6] introduced a high-dimensional random ridge term into the measurement matrix in the same way to ensure that only one index is added into or removed from the active set in each iteration. Such ridge-adding method is quite simple, which requires no additional operations. However, as mentioned in [15], [16], the traditional ridge-adding method has at least two shortcomings: 1) it is very difficult to choose a proper random ridge term that applies to several different measurement matrices; and 2) the high-dimensional ridge term may accumulatively degrade the recovery performance for large-scale applications.

Inspired by the work presented in [16], a modified ridge-adding method is proposed to overcome the above shortcomings in this paper. Specifically, instead of introducing $P \times N$ ridges into the measurement matrix directly as in [6], a random ridge vector of size $N \times 1$ is introduced into the $l_1$-norm minimization problem to ensure that only one index is added into or removed from the active set at the same time. The proposed method provides a much simpler implementation, and it can alleviate the degradation caused by the ridge term. Because the dimension of the ridge term in our method is much smaller than that of the original one in [6]. Moreover, the presented method can be further extended to handle the SVMpath initialization singularities [17]. Theoretical analysis and experimental results validate the performance improvement of the proposed method.

The rest part of this paper is organized as follows. In Sect. 2, the traditional Homotopy algorithm is reviewed briefly. In Sect. 3, the modified ridge-adding method is proposed to handle the singularity problem. In Sect. 4, the proposed method is extended to handle the SVMpath initialization singularities. In Sect. 5, the simulation results are given to demonstrate the effectiveness of the proposed method. Finally, the conclusion is presented in Sect. 6.

2. Review of the Homotopy Algorithm

The basic principle of the Homotopy algorithm is to fit the entire solution path of (4) by tracking the Karush-Kuhn-Tucker (KKT) optimality conditions with respect to $\lambda$. The active set $\mathcal{E}$ and its complementary set $\mathcal{R}$ are denoted as follows:

$$\mathcal{E} = \{i : x_i \neq 0\},$$  
(5)

$$\mathcal{R} = \{i : x_i = 0\},$$  
(6)

where $x_i$ denotes the $i$-th element of $x$. The Homotopy algorithm starts out at $x^{(0)} = 0$ with the initial sets $\mathcal{E}^{(0)} = \emptyset$ and $\mathcal{R}^{(0)} = \{1, 2, \ldots, N\}$, where $\| \cdot \|_{\infty}$ denotes the infinite norm of the vector, and $(\cdot)^{(0)}$ denotes the initialization of the corresponding variable. Let $f_\lambda(x)$ be the objective function of (4), and then its derivative, with respect to $x$, gives:

$$\frac{\partial f_\lambda(x)}{\partial x} = -A^T(y - Ax) + \lambda u = 0,$$  
(7)

where $u$ denotes the subgradient of $\|x\|_1$, whose the $i$-th element is defined as [3]:

$$\begin{cases} 
    u_i = \text{sgn}(x_i), & \text{if } i \in \mathcal{E}, \\
    u_i \in [-1, 1], & \text{if } i \in \mathcal{R}.
\end{cases}$$

The KKT optimality conditions (7) is rewritten as:

$$A_E^T(y - Ax) = \lambda \cdot \text{sgn}(x_E),$$  
(8)

$$|A_E^T(y - Ax)| \leq \lambda \cdot 1,$$  
(9)

where $A_E$ (or $A_R$) corresponds to the portion of $A$ with columns indexed by $\mathcal{E}$ (or $\mathcal{R}$) and $| \cdot |$ is applied elementwise. Differentiating both sides of (8), with respect to $\lambda$, yields:

$$-A_E^T A x' = \text{sgn}(x_E),$$  
(10)

where $x' = dx/d\lambda$. Each index $i \in \mathcal{R}$ remains in its original set with a sufficiently small change of $\lambda$. Hence, $x'_i = 0, \forall i \in \mathcal{R}$, and then (10) is rewritten as:

$$-A_E^T A x_E' = \text{sgn}(x_E).$$  
(12)

Note that when $A_E^T A_E$ is of full rank, the search direction will be obtained by

$$x_E' = -(A_E^T A_E)^{-1} \text{sgn}(x_E).$$  
(13)

Therefore, at the $l$-th iteration, the corresponding $\lambda$ and its optimal solution $x$ can be updated as follows:

$$\lambda^{(l+1)} = \lambda^{(l)} + \Delta \lambda,$$  
(14)

$$x^{(l+1)} = x^{(l)} + \Delta \lambda \cdot x',$$  
(15)

where $(\cdot)^{(l)}$ denotes the value of the corresponding variable at the $l$-th iteration, $\Delta \lambda$ is the step-size along the search direction, and $x'$ is composed of $x_R'$ and $x_E'$. With (14) and (15), we monitor any occurrence of the following events:

- **Dropping event**: An index $i \in \mathcal{E}$ moves to $\mathcal{R}$, and the corresponding $x_i$ reaches 0; thus for each $i$, the $\Delta \lambda$ can be calculated by

$$\Delta \lambda = -\frac{x_i^{(l)}}{x_i'}. $$  
(16)

- **Adding event**: An index $i \in \mathcal{R}$ moves to $\mathcal{E}$, and the index $i$ reaches the KKT condition that $A_i^T(y - Ax_i^{(l+1)}) = \pm \lambda^{(l+1)}$, where $A_i$ is the $i$-th column of $A$. Substituting (14) and (15) into the above equality, we obtain

$$\Delta \lambda = \frac{A_i^T(y - Ax_i^{(l)}) \mp \lambda^{(l)}}{A_i^T A x_i' \mp 1}. $$  
(17)

Recall that the regularization path evolves $\lambda$ from a large initial value to 0. In order to construct a descending sequence of $\lambda$, all the candidates of $\Delta \lambda$ must be negative. We select the maximum $\Delta \lambda$ of these negative candidates to determine
the next breakpoint, and let \( A^{(l+1)} \) be
\[
A^{(l+1)} = A^{(l)} + \Delta \lambda^*.
\] (18)

Using (18), we update the sets \( \mathcal{E}, \mathcal{R} \) respectively, and then go back to calculate (13). The process continues in an iterative way until a termination (i.e., \(|y - Ax|^2 = \epsilon \) or \(|x|_r = \tau \).

With the decreasing of \( \lambda \), the adding event is more likely to occur than the dropping event. However, it does not mean that the dropping event will not occur. The maximum \( \Delta \lambda^* \) determines which event occurs. Note that the least angle regression (LARS) [18] provides an approximation to the Homotopy algorithm, which considers the adding event only.

Not that the singularity problem can be a major obstacle for the application of the Homotopy algorithm. As mentioned before, the Homotopy algorithm calculates (13) to obtain the search direction with the key assumption that the matrix \( A_E^TA_E \) is of full rank. However, if \( A_E \) contains duplicate columns, approximate columns or columns with linear dependent in kernel space, such assumption will no longer be valid [6]. The presence of such matrices is quite common and inevitable in real-world data, especially in large-scale applications. Inspired by the ridge-adding method of [14] in the SVMpath algorithm, [6] introduced a novel method to handle the singularity problem for the Homotopy algorithm, which similarly added a high-dimensional random ridge term to \( A \). Since the maximal value of \( \Delta \lambda \) is unique almost surely (with probability one), only one index is added into or removed from the active set at the same time. According to Lemma 1 in [6], we conclude that the singularity problem can be effectively avoided. However, the ridge-adding method [6] has at least two limitations. Firstly, it is very difficult to choose a proper ridge term that applies to several different measurement matrices. Adding a very small ridge term might fail to work in some rare cases, while adding a relatively large value may cause significant errors. Secondly, the high-dimensional ridge term may accumulatively degrade the recovery performance for large-scale applications. To get around these limitations, we try to propose a modified ridge-adding method to deal with the singularity problem.

3. New Ridge-Adding Approach

Instead of introducing a ridge terms of size \( P \times N \) into the measurement matrix directly as in [6], a random ridge vector of size \( N \times 1 \) is introduced into the \( l_1 \)-norm minimization problem as
\[
\min_x \frac{1}{2} ||y - Ax||^2 + \lambda(1 + \tau) \circ |x|_1,
\] (19)

where \( \tau \) is the ridge vector which has i.i.d zero-mean Gaussian entries with a sufficient small variance, and \( \circ \) stands for the Hadamard product. Let \( \eta = 1 + \tau \), and then the partial derivatives of the objective function \( f_1(x) \), with respect to \( x \), gives:
\[
\frac{\partial f_1(x)}{\partial x} = -A^T(y - Ax) + \lambda \cdot (\eta \circ u) = 0.
\] (20)

Then, the KKT optimal conditions (8) and (9) are rewritten as:
\[
A_E^T(y - Ax) = \lambda \cdot [\eta_E \circ \text{sgn}(x_E)],
\] (21)
\[
|A_E^T(y - Ax)| \leq \lambda \cdot \eta_R.
\] (22)

Obviously, the proposed method is equivalent to introducing a small random ridge scalar into the KKT optimality condition for each index. In this case, (12) becomes:
\[
-A_E^TA_Ex'_E = \eta_E \circ \text{sgn}(x_E).
\] (23)

When \( A_E^TA_E \) is of full rank, the search direction is obtained as follows:
\[
x'_E = -(A_E^TA_E)^{-1}[\eta_E \circ \text{sgn}(x_E)].
\] (24)

When the \( l \)-th event occurs, the candidates of \( \Delta \lambda \) are calculated by
\[
\Delta \lambda_1 = \max_{\Delta \lambda} \left\{ \Delta \lambda | \Delta \lambda = \frac{-y_i^l}{A_E^T x_i^l}, \Delta \lambda \leq \lambda, i \in \mathcal{E} \right\},
\] (25)
\[
\Delta \lambda_2 = \max_{\Delta \lambda} \left\{ \Delta \lambda | \Delta \lambda = \frac{A_E^T (y - Ax^l) \pm \lambda^l \eta_i}{A_E^T A_E \pm \eta_i}, \Delta \lambda \leq \lambda, i \in \mathcal{R} \right\},
\] (26)

where (26) follows the KKT condition that \( A_E^T(y - Ax_l^{(l+1)}) = \pm A_E \eta_i \).

Due to the randomness of \( \eta_i \), the maximal value of \( \Delta \lambda \) is unique (with measure 1), which means that only one corresponding index will be chosen to added into or removed from the active set. To show the effectiveness of the proposed method, what remains is to prove the following theorem.

**Theorem 1.** If \( A_{E_l} \) is of full column rank, \( A_{E_l(l+1)} \) will remain to be of full column rank for arbitrarily small random ridges.

**Proof.** See Appendix. \( \square \)

From Theorem 1, it can be seen that when the ridge term tends to 0, the modified approach always works well and effectively avoids the singularity problem. Compared with the previous ridge-adding method in [6], the proposed method can overcome its drawbacks as follows:

1. The proposed method just adds a random ridge vector into a constant vector \( \mathbf{1} \) as in (19) to guarantee the validity of the sufficient condition, rather than modifying the measurement matrix directly. This indicates that it is easy to choose proper ridges using the presented method. The added ridges should be sufficiently small within a reasonable accuracy, such as with \( O(10^{-n}) \), where \( 6 \leq n \leq 10 \). In this case, the distance between the optimal solutions to (19) and (4) is of order \( O(10^{-n}) \) for every value of the regularization parameter \( \lambda \).
2. The proposed method modifies the vector \( \mathbf{1} \) with a total of \( N \) ridges, while \( PN \) ridges are required in the previous ridge-adding method [6]. Thus, the proposed
approach reduces the influence of the added ridges on the solution path for high-dimension measurement matrices in practical application.

Although the above findings coincide with the one in [16], the method proposed in [16] is designed for SVMpath singularities only, which cannot be applied to the SVMpath initialization problem directly. It should be pointed that the proof of Theorem 1 is by no means a trivial task because the derivations are quite distinct from the one in [16]. Moreover, it is worth noting that solving the KKT system (23) might be the most expensive process. Fortunately, we can adopt null-space method [19] with the Cholesky factorization update formula in [20] to significantly reduce the computational complexity.

4. Extension to the SVMpath Initialization

In this section, we further extend our proposed method to handle the SVMpath initialization singularities [12], [17]. The SVMpath initialization problem is an exact \( l_1 \)-norm minimization problem with some additional linear constraints. Consider a typical binary SVM classification problem, given a set of sample data \( D = \{ (s_i, t_i) \} \), \( i = 1, 2, \ldots, N \), where \( s_i \in \mathbb{R}^n \) represents the feature vector of the \( i \)-th observation and \( t_i \in \{-1, +1\} \) is its label. Assume that \( N_+ > N_- \), where \( N_+ \) and \( N_- \) are the numbers of positive and negative samples, respectively. The SVMpath initialization problem is written as:

\[
\min_\alpha \frac{1}{2} \alpha^T D^T D \alpha \quad \text{s.t.} \quad \sum_{i \in I} \alpha_i = N_-, \quad \alpha_i = 1, \ i \in I, \quad 0 \leq \alpha_i \leq 1, \ i \in I_+,
\]

where \( \alpha \in \mathbb{R}^{N \times 1} \) is the Lagrangian multiplier vector, and \( \alpha_i \) represents the \( i \)-th element of \( \alpha \), \( I_- \) and \( I_+ \) are the sets of indices of positive and negative samples, respectively. \( D^T D \) is an arbitrary symmetric decomposition of the matrix with the \( (i, j) \)-th element being \( d_{ij} \) and \( K(s_i, s_j) \) stands for the kernel function. Letting \( y = -D^T \alpha \), \( A = D_0 \), and \( x = \alpha_k \), the problem (27) can be transformed into

\[
\min_x \frac{1}{2} \| y - Ax \|^2_2 + \lambda \cdot \sum_{i=1}^{N_+} x_i \\
\text{s.t.} \quad 0 \leq x_i \leq 1, \ i = 1, 2, \ldots, N_+,
\]

Then, the random ridge vector \( \tau \) is added similarly as in (19), i.e.,

\[
\min_x \frac{1}{2} \| y - Ax \|^2_2 + \lambda \cdot \eta^T x \\
\text{s.t.} \quad 0 \leq x_i \leq 1, \ i = 1, 2, \ldots, N_+.
\]

To solve the constrained \( l_1 \)-norm minimization problem (28), its Lagrangian function is constructed as:

\[
L(x, \lambda, a, b) = \frac{1}{2} \| y - Ax \|^2_2 + \lambda \cdot \eta^T x \\
- \sum_{i=1}^{N_+} a_i x_i - \sum_{i=1}^{N_-} b_i (1 - x_i),
\]

where \( a = [a_1, a_2, \ldots, a_{N_-}]^T \) and \( b = [b_1, b_2, \ldots, b_{N_+}]^T \) are the Lagrangian multipliers. Differentiate (29) with respect to \( x \) and we can obtain the following:

\[
\frac{\partial L}{\partial x} = -A^T (y - Ax) + \lambda \eta - a + b = 0,
\]

\[
0 \leq x_i \leq 1, \quad a_i \geq 0, \quad b_i \geq 0.
\]

All the positive samples \( 1 \leq i \leq N_+ \) are divided into three auxiliary subsets \( E, L, R \):

\[
0 \leq x_i \leq 1, \quad i \in E, \quad x_i = 1, \quad i \in L, \quad x_i = 0, \quad i \in R.
\]

With these auxiliary subsets, the KKT conditions are rewritten as follows:

\[
A_E^T (y - Ax) = \lambda \eta_E, \quad (35)
\]

\[
A_L^T (y - Ax) \geq \lambda \eta_L, \quad (36)
\]

\[
A_R^T (y - Ax) \leq \lambda \eta_R. \quad (37)
\]

When \( A_E^T A_E \) is of full rank, the search direction is obtained as follows:

\[
\begin{cases}
\begin{align*}
x_E' &= - (A_E^T A_E)^{-1} \cdot \eta_E \\
x_{E,L,R}' &= 0
\end{align*}
\end{cases}
\]

At the \( t \)-th iteration, the candidates of \( \Delta \lambda \) is calculated when one of the following events occurs:

- Adding event: An index \( i \in L \cup R \) enters \( E \), satisfying the KKT condition \( A_E^T (y - Ax^{(t+1)}) = \lambda^{(t+1)} \eta_i \). Similarly, we obtain

\[
\Delta \lambda = \frac{A_E^T (y - Ax^{(t)}) - \lambda^{(t)} \eta_i}{A_E^T A_E' + \eta_i}. \quad (39)
\]

- Dropping event: An index \( i \in E \) enters \( L \cup R \), and the corresponding \( x_i \) reaches 0 or 1. Thus, the corresponding \( \Delta \lambda \) is calculated by

\[
\Delta \lambda = - \frac{x_i^{(t)}}{x_i'}, \quad (40)
\]

or,

\[
\Delta \lambda = \frac{1 - x_i^{(t)}}{x_i'}. \quad (41)
\]

Then, we choose the maximum of these negative candidates. Update \( \lambda, x, \) and the auxiliary subsets \( E, L, R, \) and go back to calculate (38). The process terminates if \( 1^T x = N_+ \) is satisfied, and then the final optimal solution \( x^* \) can be obtained.
5. Simulation Results

In this section, several simulation results are presented to illustrate the effectiveness of the proposed method. The experiments are performed on an Intel Core i7-6700HQ CPU with 16 GB RAM under the Windows 10 operating system using MATLAB R2019a.

In Simulation 1, we test the validity of handling the singularity problem by two types of measurement matrices with duplicate columns and linearly dependent columns, respectively. The first type of measurement matrix is in the form of
\[
A_1 = \begin{bmatrix}
-1.3 & 0.4 & -1.3 & 1.6 & -0.3 & 1.6 & 0.7 & -0.2 & -1.6 \\
-0.1 & -1.0 & 0.3 & 0.3 & 0.7 & 0.3 & -0.3 & 0.6 & 1.0 \\
-0.8 & -0.3 & -0.8 & -2.0 & 1.6 & -2.0 & 0.1 & 1.2 & 0.2 \\
0.8 & -0.1 & -0.3 & 2.4 & 0.0 & 2.4 & -0.1 & -0.3 & -0.9 \\
-0.2 & 0.9 & 0.3 & 0.1 & 1.4 & 0.1 & 0.0 & 0.8 & 0.9 
\end{bmatrix},
\]
where column 4 and 6 are duplicates of each other, and $y_1 = [-0.7, 0.1, -1.1, -1.9, 0.5]^T$. The other tested measurement matrix is in the form of
\[
A_2 = \begin{bmatrix}
-1.0 & 0.2 & -0.5 & -0.5 & 0.2 & 0.8 \\
0.0 & 0.4 & 0.0 & -0.2 & -0.3 & 0.0 \\
0.5 & 0.4 & -1.0 & -0.6 & -0.6 & 0.1 
\end{bmatrix},
\]
where column 1, 3 and 6 are linearly dependent, and $y_2 = [1.0, -1.5, 1.0]^T$. The whole regularization path with the candidates of $\lambda$ obtained by our proposed method in the case of two types of measurement matrices, respectively. From the tables, it can be seen that only one index moves to another set in each iteration with the changing of $\lambda$, and the singularity problems are totally avoided. We also find that the objective cost (defined as $\frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1$) for our proposed method is almost consistent with that obtained by CVX package [21], [22]. Obviously, our proposed method can maintain a reasonable accuracy along the whole regularization path.

In Simulation 2, we investigate the performance of the proposed method for large-scale applications. Assume that both $A$ and $y$ have i.i.d. zero-mean Gaussian entries with unit variance. Figure 1-a depicts the cost value comparisons among our method, the traditional method [17] and the CVX method, where $P = 50$ and $N = 200$; Figure 1-b depicts the cost value comparisons among our method, the traditional method [17] and the CVX method, where $P = 200$ and $N = 2000$. It can be seen that: 1) for a moderate problem (Fig. 1-a), all the methods achieve the same performance; 2) for a large-scale problem (Fig. 1-b), our method retains a good performance (compared with the CVX method); while the traditional method [17] has a notable performance degradation.

In Simulation 3, we test the performance of our method by solving SVMpath initialization with several practical datasets from the University of California Irvine (UCI) repository [23]. The kernel functions used in the experiment are as follows: 1) linear kernel with $K(s_i, s_j) = s_i^Ts_j$; and 2) Radial Basis Function (RBF) kernel with $K(s_i, s_j) = \exp(-\gamma\|s_i - s_j\|_2^2)$, where $\gamma = 0.1$. Comparison between the proposed method and the traditional algorithm in [17] for linear and RBF kernel has been implemented (shown in Tables 3 and 4). The quantities $l_{\text{max}}$, $|E|_{\text{max}}$ and $T$ refer to the maximal number of events along the entire path, and the maximal cardinality of the set $E$, and the total runtime, respectively. As shown in the tables, the proposed method

| Table 1 | The whole regularization path with the candidates of $\lambda$ obtained by our method in the case of $A_1$. |
|---------|------------------------------------------------|
| $l$-th | $\lambda^{(l)}$ | Index | Event | Our cost | CVX cost |
| 0 | 3.4000000000 | 4 | $R \rightarrow E$ | 2.785000000 | 2.785000002 |
| 1 | 3.012377026 | 9 | $R \rightarrow E$ | 2.778951225 | 2.778951241 |
| 2 | 1.974606034 | 3 | $R \rightarrow E$ | 2.642821264 | 2.642821274 |
| 3 | 1.89859410 | 5 | $R \rightarrow E$ | 2.320065592 | 2.320065603 |
| 4 | 0.973698385 | 2 | $R \rightarrow E$ | 2.167699406 | 2.167699415 |
| 5 | 0.957605859 | 3 | $E \rightarrow R$ | 2.095138689 | 2.095138699 |
| 6 | 0.707761235 | 1 | $R \rightarrow E$ | 1.867526999 | 1.867527010 |
| 7 | 0.123819973 | 4 | $E \rightarrow R$ | 0.455530722 | 0.455530725 |
| 8 | 0.045344180 | 3 | $R \rightarrow E$ | 0.173755660 | 0.173755662 |
| 9 | 0.000000000 | 4 | $R \rightarrow E$ | 0.000000000 | 0.000000000 |

| Table 2 | The whole regularization path with the candidates of $\lambda$ obtained by our method in the case of $A_2$. |
|---------|------------------------------------------------|
| $l$-th | $\lambda^{(l)}$ | Index | Event | Our cost | CVX cost |
| 0 | 1.5000000000 | 3 | $R \rightarrow E$ | 2.125000000 | 2.125000012 |
| 1 | 0.500001782 | 6 | $R \rightarrow E$ | 1.725001426 | 1.725001432 |
| 2 | 0.499999375 | 5 | $R \rightarrow E$ | 1.682999450 | 1.682999458 |
| 3 | 0.439023236 | 6 | $E \rightarrow R$ | 1.672574677 | 1.672574679 |
| 4 | 0.403361060 | 2 | $R \rightarrow E$ | 1.631409844 | 1.631409852 |
| 5 | 0.000000000 | 6 | $R \rightarrow E$ | 0.000000000 | 0.000000000 |
works quite well with the practical datasets, and it even outperforms the original method [17] in terms of training time, especially for RBF kernel.

6. Conclusion

A modified ridge-adding approach for handling singularity problem in the Homotopy algorithm has been proposed in this paper. Inspired by the idea of adding ridges in [16], a random ridge vector is introduced into the $l_1$-norm minimization problem to ensure that only one index is added into or removed from the active set. The singularity problem can always be avoided by using the proposed even when the ridge term tends to 0. In addition, the proposed method is extended to handle the SVMpath initialization singularities. The proposed method provides a much simpler implementation and reduces the influence of the added ridges on the solution path. Theoretical analysis and experimental results show that the proposed method has a great advantage over the existing methods.

References

[1] D.L. Donoho, “Compressed sensing,” IEEE Trans. Inf. Theory, vol.52, no.4, pp.1289–1306, 2006.
[2] E.J. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,” IEEE Trans. Inf. Theory, vol.52, no.2, pp.489–509, 2006.
[3] D.L. Donoho and Y. Tsaig, “Fast solution of $l_1$-norm minimization problems when the solution may be sparse,” IEEE Trans. Inf. Theory, vol.54, no.11, pp.4789–4812, 2008.
[4] A.Y. Yang, S.S. Sastry, A. Ganesh, and Y. Ma, “Fast $l_1$-minimization algorithms and an application in robust face recognition: A review,” 2010 17th IEEE International Conference on Image Processing (ICIP), pp.1849–1852, IEEE, 2010.
[5] R. Tibshirani, “Regression shrinkage and selection via the lasso,” Journal of the Royal Statistical Society. Series B (Methodological), vol.58, no.1, pp.267–288, 1996.
[6] J. Dai, W. Xu, J. Zhang, and C. Chang, “Homotopy algorithm for $l_1$-norm minimisation problems,” IET Signal Processing, vol.9, no.1, pp.1–9, 2015.
[7] D. Malioutov, M. Cetin, and A.S. Willsky, “A sparse signal reconstruction perspective for source localization with sensor arrays,” IEEE Trans. Signal Process., vol.53, no.8, pp.3010–3022, 2005.
[8] J. Dai, A. Liu, and V.K. Lau, “FDD massive MIMO channel estimation with arbitrary 2D-array geometry,” IEEE Trans. Signal Process., vol.66, no.10, pp.2584–2599, 2018.
[9] X. Rao and V.K. Lau, “Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems,” IEEE Trans. Signal Process., vol.62, no.12, pp.3261–3271, 2014.
[10] S. Rosset and J. Zhu, “Piecewise linear regularized solution paths,” The Annals of Statistics, vol.35, no.3, pp.1012–1030, 2007.
[11] G. Wang, D.-Y. Yeung, and F.H. Lochovsky, “A new solution path algorithm in support vector regression,” IEEE Trans. Neural Netw., vol.19, no.10, pp.1753–1767, 2008.
[12] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, “The entire regularization path for the support vector machine,” Journal of Machine Learning Research, vol.5, no.Oct, pp.1391–1415, 2004.
[13] C.-J. Ong, S. Shao, and J. Yang, “An improved algorithm for the solution of the regularization path of support vector machine,” IEEE Trans. Neural Netw., vol.21, no.3, pp.451–462, 2010.
[14] J. Dai, C. Chang, F. Mai, D. Zhao, and W. Xu, “On the SVMpath singularity,” IEEE Trans. Neural Netw. Learn. Syst., vol.24, no.11,
Appendix: Proof of Theorem 1

When the dropping event occurs, Theorem 1 is distinctly valid. Hence, we just need to provide the proof for the adding event. Let $\mathbf{e}(l) = \{1, 2, \ldots, m\}$, and $(m + 1)$ be the new index entering the active set at $\lambda = \lambda(l)$. The remaining proof is achieved by contradiction.

Assume that there exists a nonzero vector $\nu$ such that $
abla\mathbf{A}_{1:m, \nu} = \mathbf{A}_{m+1}$ in $\mathbf{A}_{\nu}^{(l+1)}$, where $\mathbf{A}_{1:m}$ represents the sub-matrix from the first to the $m$-th column of $\mathbf{A}$. Let $x^{(l+1)}$ and $\mathbf{X}^{(l+1)}$ be the optimal solutions to the $l_1$-norm minimization problem (19) at $\lambda = \lambda(l)$ and $\lambda = \lambda(l+1)$, respectively, where $\cdot$ represents the state just after the adding event occurs and $\lambda(l+1) = \lambda(l) + \delta$. Then, $x^{(l+1)}$, can be represented as:

$$x^{(l+1)} = \left[ x_1^{(l+1)}, e_1, x_2^{(l+1)}, e_2, \ldots, x_m^{(l+1)}, e_m + e_{m+1} \right]^T,$$

where $x_i^{(l+1)}$ is the $i$-th element of $x^{(l+1)}$. From (A-1), we can obtain the following:

$$A^T_{l,m} \left( y - A x^{(l+1)} \right) = A^T_{l,m} \left( y - A (x^{(l+1)} + \mathbf{e}) \right) = A^T_{l,m} \left( y - A x^{(l+1)} \right) - A^T_{l,m} \mathbf{e} = \lambda(l+1) \left[ (1 + \tau_{l,m}) \circ \text{sgn}(x_{l,m}^{(l+1)}) \right] - A^T_{l,m} \mathbf{e}_{1:m} \mathbf{e}_{1:m} + e_{m+1},$$

where $\mathbf{e} = \left[ e_1, e_2, \ldots, e_{m+1} \right]^T$, and $\tau_{l,m}$ is the sub-vector from the first to the $m$-th element of $\tau$. As we know that the left side of (A-2) can also be written as:

$$A^T_{l,m} \left( y - A x^{(l+1)} \right) = A^T_{l,m} \left[ (1 + \tau_{l,m}) \circ \text{sgn}(x_{l,m}^{(l+1)}) \right].$$

Thus, we have

$$A^T_{l,m} \left[ (1 + \tau_{l,m}) \circ \text{sgn}(x_{l,m}^{(l+1)}) \right] = \lambda(l+1) \left[ (1 + \tau_{l,m}) \circ \text{sgn}(x_{l,m}^{(l+1)}) \right] - A^T_{l,m} \mathbf{e}_{1:m} \mathbf{e}_{1:m} + e_{m+1} \mathbf{v}.$$  

Then, we can obtain that

$$\left( A^{(l+1)} \circ \text{sgn}(x_{l,m}^{(l+1)}) \right) = \lambda(l+1) \circ \text{sgn}(x_{l,m}^{(l+1)}) - A^T_{l,m} \mathbf{e}_{1:m} \mathbf{e}_{1:m} + e_{m+1} \mathbf{v}.$$  

Since that $\mathbf{A}_{\nu}^{(l+1)}$ is of full column rank, from (24) and (A-5), we claim the following:

$$\epsilon_{1:m} + e_{m+1} \mathbf{v} \bigg|_{\epsilon \rightarrow 0} = \mathbf{x}_{\nu}^{(l+1)},$$

$$\left[ \lambda(l+1) \circ \text{sgn}(x_{l,m}^{(l+1)}) - A^{(l+1)} \circ \text{sgn}(x_{l,m}^{(l+1)}) \right] \bigg|_{\epsilon \rightarrow 0} = \delta \circ \text{sgn}(x_{l,m}^{(l+1)}).$$

According to these results, the contradiction can be derived as follows. Firstly, the KKT condition for the index $(m + 1)$ at $\lambda = \lambda(l+1)$, is

$$A^T_{m+1} \left( y - A x^{(l+1)} \right) = \lambda(l+1) (1 + \tau_{m+1}) \circ \text{sgn}(x_{m+1}^{(l+1)}),$$

while $A^T_{m+1} \left( y - A x^{(l+1)} \right)$ is also equal to

$$A^T_{m+1} \left( y - A x^{(l+1)} \right) = \lambda(l+1) (1 + \tau_{m+1}) \circ \text{sgn}(x_{m+1}^{(l+1)}) - A^T_{m+1} \mathbf{A}_{1:m} \mathbf{e}_{1:m} + e_{m+1} \mathbf{v}.$$  

Thus, when $\tau_{m+1}$ tends to 0, we can obtain the following:

$$\delta \circ \text{sgn}(x_{m+1}^{(l+1)}) + \delta \mathbf{A}_{m+1} \mathbf{A}_{1:m} \mathbf{X}_{\nu}^{(l+1)} = 0.$$  

Secondly, the KKT condition for the index $(m + 1)$ at $\lambda = \lambda(l+1)$ is

$$A^T_{m+1} \left( y - A x^{(l+1)} \right) \bigg|_{\epsilon \rightarrow 0} = \lambda(l+1) (1 + \tau_{m+1}) \circ \text{sgn}(x_{m+1}^{(l+1)}) - A^T_{m+1} \mathbf{A}_{1:m} \mathbf{e}_{1:m} + e_{m+1} \mathbf{v},$$

where $\lambda(l+1) = \lambda(l) - \delta$, while the left side of (A-11) can be written as

$$A^T_{m+1} \left( y - A x^{(l+1)} \right) \bigg|_{\epsilon \rightarrow 0} = A^T_{m+1} \left( y - A x^{(l+1)} \right) +$$
\[
\begin{align*}
A^T_{m+1}A_{1:m}(e_{1:m} + e_{m+1}v)\bigg|_{\tau_{m+1} \rightarrow 0} &= \lambda^{(l+1)} \cdot \text{sgn}(x_{m+1}^{(l+1)})\bigg|_{\tau_{m+1} \rightarrow 0} + \delta A^T_{m+1}A_{1:m}x_E^{(l)}, \\
&= \lambda^{(l+1)} \cdot \text{sgn}(x_{m+1}^{(l+1)})\bigg|_{\tau_{m+1} \rightarrow 0} + \delta A^T_{m+1}A_{1:m}x_E^{(l)}, \\
&= \lambda^{(l+1)} \cdot \text{sgn}(x_{m+1}^{(l+1)})\bigg|_{\tau_{m+1} \rightarrow 0} + \delta A^T_{m+1}A_{1:m}x_E^{(l)},
\end{align*}
\]
thus, we can obtain
\[
\delta \cdot \text{sgn}(x_{m+1}^{(l+1)}) + \delta A^T_{m+1}A_{1:m}x_E^{(l)} \neq 0,
\]
which contradicts with (A·10). Hence, \(A_{E^{(l+1)}}\) must be of full column rank.

Haoran Li received the B.Eng. degree in electronic information engineering from Jiangsu University, Zhenjiang, China, in 2014, where he is currently pursuing the Ph.D. degree in control science and engineering in Jiangsu University. His research interests include chemometrics, data driven modelling and optimization.

Binyu Wang received the B.Eng. degree in electronic information engineering from Jiangsu University, Zhenjiang, China, in 2016, where she is currently pursuing the master’s degree with the School of Electrical and Information Engineering, Jiangsu University. Her research interests include support vector machine and sparse signal recovery.

Jisheng Dai received the B.Eng. degree in electronic engineering from the Nanjing University of Technology, Nanjing, China, in 2005, and the Ph.D. degree in information and communication engineering from the University of Science and Technology of China, Hefei, China, in 2010. He is currently a Research Fellow with the School of Electrical and Information Engineering, Jiangsu University, Zhenjiang, China. His research interests include convex optimization theory, wireless communications, machine learning, and bioinformatics.

Tianhong Pan received the Ph.D. degree in control theory and control engineering from Shanghai Jiao Tong University in 2007. He is currently a Professor with the School of Electrical Engineering and Automation, Anhui University, Hefei, China. His research interests include multiple model approach and its application, machine learning, virtual metrology, predictive control and run-to-run control theory and practice, etc.