International Journal of Pure and Applied Mathematics
Volume 112 No. 2 2017, 333-340
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
doi: 10.12732/ijpam.v112i2.10

HOMOTOPY ALGORITHM FOR BOX-CONSTRAINED LASSO AND ITS CONVERGENCE

Xijun Liang[1] §, Yongxiang Wang[2]
[1,2]College of Science
China University of Petroleum
Qingdao, P.R. CHINA

Abstract: The least absolute shrinkage and selection operator (Lasso) has received extensive attentions when sparse solutions are required. The homotopy method has become a significant approach for Lasso models because it can produce an entire regularization path. Although convergence has been studied on some special design matrices, the convergence analysis of the general homotopy method remains a problem.

In this work, we present a homotopy method for the generalized Lasso model with box constraints. We propose mild convergence assumptions and prove that the algorithm terminates at an optimal solution under convergence assumptions.

AMS Subject Classification: 90C25, 90C20
Key Words: LASSO, homotopy, box constraints, regularization path

1. Introduction

The Lasso model [1] and its variations have been widely studied, and applied in model selection, compressive sensing and logistic regression. Lasso is closely related to Dantzig selector. They have been compared and discussed theoretically and numerically, e.g., [2]. It was emphasized in [2] that computing Dantzig selector or Lasso for a single value of the penalty parameter does not work in practice, and the entire solution path is needed to select a meaningful model with good predictive performance.

Received: September 13, 2016
Revised: November 14, 2016
Published: February 1, 2017

§Correspondence author

The homotopy method [3, 4] can generate the entire solution path in one run and it has shown encouraging numerical results [5] . While numerical results of existing homotopy algorithms are promising, limit attention has been given to the convergence analysis. For special design matrices, D. L. Donoho and Y. Tsaig [6] proved that the classical homotopy algorithm has k-step solution property. However, the assumed condition does not hold on many practical datasets and the sequence of regularization parameters may not strictly decrease. If recursion occurs, the algorithm fails to find an optimal solution.

In this work, we consider a general Lasso model for recommender systems [7], in which box constraints are added to avoid a predominant variable. We propose a generalized homotopy algorithm to seek for sparse solutions. Under convergence assumptions, we prove that the proposed homotopy algorithm terminates at an optimal solution. These assumptions, especially the non-degeneration index assumption, ensures that the degeneration case will not occur and the sequence of regularization parameters strictly decreases.

2. The homotopy Algorithm for the Box-Constrained Lasso

The box-constrained Lasso model can be recast as follows:

$$\begin{array}{ll} \min_x & \frac{1}{2}\|Ax-y\|_2^2+\lambda\|x\|_1 \\ \text{s.t.} & \|Wx\|_\infty \le \kappa. \end{array} \tag{1}$$

where design matrix $A := [a_1 \ \dots \ a_n] \in \mathbb{R}^{m,n}$, $m \le n$, $y \in \mathbb{R}^m$, regularization parameter $\lambda > 0$ and $\kappa > 0$ are pre-determined parameters, W is a diagonal matrix $\mathrm{Diag}(w_1, \dots, w_n)$, $w_i \ge 0$. Model (1) degenerates to the classical Lasso when $W = 0$.

We generalize the classical homotopy algorithm [3, 4] for solving model (1). As model (1) is a convex, a vector x is optimal if and only if there exists $u \in \mathbb{R}$ such that (x, u) satisfies the KKT conditions:

$$\begin{cases} 0 \in A^T(Ax-y)+\lambda\cdot\partial\|x\|_1+u\cdot\partial\|Wx\|_\infty, & (2) \\ \|Wx\|_\infty \le \kappa, & (3) \\ u \ge 0, \quad u(\|Wx\|_\infty-\kappa)=0. & (4) \end{cases}$$

Denote $I = \{i \,|\, x_i \neq 0\}$ the active index set. Let

$$I_1 = \{i \,|\, |w_i x_i| = \kappa\}, \quad I_2 = I \setminus I_1, \quad I_3 = \{1, \dots, n\} \setminus I. \tag{5}$$

For any index set J, denote A_J the submatrix consisting of the columns with the indices in J, and $x(J)$ the vector consisting of the elements of the

vector x with the indices in J. Let $\mathrm{co}\{\cdot\}$ be convex hull of a give set. Condition (2) can be written as

$$\begin{cases} -A_{I_1}^T(Ax-y) \in \lambda \cdot \mathrm{sgn}(x(I_1)) + u \cdot \mathrm{co}_{i\in I_1}\{\mathrm{sgn}(x_i)w_i e_i(I_1)\} & (6) \\ -A_{I_2}^T(Ax-y) = \lambda \cdot \mathrm{sgn}(x(I_2)), & (7) \\ |A_{I_3}^T(Ax-y)| \le \lambda, & (8) \end{cases}$$

where e_i is a vector of all 0's except 1 at the i-th position.

Let $c = -A^T(Ax-y)$. It follows by (6), (7) that $\mathrm{sgn}(c^k(I_1^k)) = \mathrm{sgn}(x^k(I_1^k))$, and $\mathrm{sgn}(c^k(I_2^k)) = \mathrm{sgn}(x^k(I_2^k))$. Note that if conditions

$$\begin{cases} -A_{I_1}^T(Ax-y) \in \lambda \cdot \mathrm{sgn}(x(I_1)) + u \cdot \mathrm{ri}\,\{\mathrm{co}_{i\in I_1}\{\mathrm{sgn}(x_i)\cdot w_i e_i(I_1)\}\}, \\ |A_{I_3}^T(Ax-y)| < \lambda, \end{cases}$$

hold for an x, then they also hold in a small neighborhood of x. Taking derivative of x with respect to λ on both side of (7), we have that $-A_{I_2}^T A_{I_2}\frac{dx(I_2)}{d\lambda} = \mathrm{sgn}(x(I_2))$. Then there exists an interval of λ such that the solution x is a linear mapping with λ and the optimality conditions are satisfied pointwise. The key of the algorithm is to determine the length of the interval.

The details of the homotopy algorithm are described as follows. Initially, we choose

$$\lambda^0 = \|A^T y\|_\infty, \tag{9}$$

so that $x^0 = 0$ is an optimal solution of model (1) at $\lambda := \lambda^0$. At the kth iteration, the algorithm updates x^k and λ^k by

$$x^{k+1} = x^k + \gamma^k d^k, \tag{10}$$

and

$$\lambda^{k+1} = \lambda^k - \gamma^k, \tag{11}$$

where d^k is the direction along which x^k is updated, and γ^k is the step size.

1. Determine descent direction d^k.

Set $d^k(I_1^k) = 0$, $d^k(I_3^k) = 0$, and solve $d^k(I_2^k)$ such that the magnitudes of the coordinates of the right-hand of equation (7) decrease equally, i.e., $-A_{I_2^k}^T[A(x^k + \gamma d^k) - y] = (\lambda - \gamma)\cdot \mathrm{sgn}(x^k(I_2^k))$. Since $\mathrm{sgn}(x^k(I_2^k)) = \mathrm{sgn}(c^k(I_2^k))$, we have

$$A_{I_2^k}^T A_{I_2^k} d^k(I_2^k) = \mathrm{sgn}(c^k(I_2^k)). \tag{12}$$

2. Calculate step size γ^k.

Let $x^k(\gamma) := x^k + \gamma d^k$ with $\gamma \geq 0$. Then the homotopy algorithm computes the step size such that $x^k(\gamma)$ satisfies (7) and reaches a breakpoint if either of the conditions (3), (6)–(8) is violated. There are four cases. First, an element of $-A_{I_3^k}^T(Ax^k(\gamma) - y)$ increases in magnitude beyond $\lambda^k - \gamma$, violating (8). This occurs when

$$\gamma_+^k = \min_{i \in I_3^k}^+ \left\{ \frac{\lambda^k - c_i^k}{1 - a_i^T v^k} \,\middle|\, a_i^T v^k \neq 1 \right\} \cup \left\{ \frac{\lambda^k + c_i^k}{1 + a_i^T v^k} \,\middle|\, a_i^T v^k \neq -1 \right\}, \tag{13}$$

where $v^k = Ad^k = A_{I_2^k} d^k(I_2^k)$, $\min^+$ denotes minimum over nonnegative values and i_+^k denotes the minimizing index. The second case occurs when an element of $x^k(I_2^k) + \gamma d^k(I_2^k)$ crosses zero, violating (7). This occurs when

$$\gamma_-^k = \min{}^+\{-x_i^k/d_i^k \mid i \in I_2^k, d_i^k \neq 0,\ \operatorname{sgn}(d_i^k) = -\operatorname{sgn}(c_i^k)\}, \tag{14}$$

where i_-^k denotes the minimizing index. The third case occurs when a coordinate of $x^k(I_2^k) + \gamma d^k(I_2^k)$ breaks the box constraint $w_i|x_i^k + \gamma d_i^k| \leq \kappa$, violating (3). This occurs when

$$\gamma_{--}^k = \min{}^+ \left\{ \frac{\pm\kappa/w_i - x_i^k}{d_i^k} \,\middle|\, i \in I_2^k, d_i^k x_i^k \geq 0,\ d_i^k \neq 0, w_i > 0 \right\}, \tag{15}$$

where i_{--}^k denotes the minimizing index. The fourth case occurs when the magnitude of a coordinate of $-A_{I_1^k}^T(A(x^k + \gamma d^k) - y)$ decreases to $\lambda^k - \gamma$, violating (6). This occurs when

$$\begin{aligned} \gamma_{++}^k = \min_{i \in I_1^k}^+ \quad & \left\{ \frac{\lambda^k - c_i^k}{1 - a_i^T v^k} \,\middle|\, c_i^k > 0,\ a_i^T v^k > 1 \right\} \\ & \cup \left\{ \frac{\lambda^k + c_i^k}{1 + a_i^T v^k}, \,\middle|\, c_i^k < 0,\ a_i^T v^k < -1 \right\}, \end{aligned} \tag{16}$$

where i_{++}^k denotes the minimizing index. In summary, γ^k is determined by

$$\gamma^k = \min\{\gamma_+^k, \gamma_-^k, \gamma_{--}^k, \gamma_{++}^k, \lambda^k - \lambda\}, \tag{17}$$

and index set I^k, I_1^k, I_2^k and I_3^k are updated by

$$\begin{cases} I_1^k \xrightarrow{i_{++}^k} I_2^k, & \text{if } \gamma^k = \gamma_{++}^k, \\ I_1^k \xleftarrow{i_{--}^k} I_2^k, & \text{if } \gamma^k = \gamma_{--}^k, \\ I_2^k \xrightarrow{i_-^k} I_3^k,\ I^{k+1} = I^k \setminus \{i_-^k\}, & \text{if } \gamma^k = \gamma_-^k, \\ I_2^k \xleftarrow{i_+^k} I_3^k, I^{k+1} = I^k \cup \{i_+^k\}, & \text{if } \gamma^k = \gamma_+^k. \end{cases} \tag{18}$$

The updated index sets are denoted by I^{k+1}, I_1^{k+1}, I_2^{k+1} and I_3^{k+1}, resp. The homotopy algorithm is summarized in Algorithm 1.

Algorithm 1. homotopy Algorithm for the Box-constrained Lasso

Step 1. Set $k = 0$, $x^0 = 0$. Initialize λ^0 by (9) and I_i^0, $i = 1, 2, 3$.

Step 2. Update direction d^k by solving equation (12).

Step 3. Determine step size γ^k by solving equation (17).

Step 4. Calculate x^{k+1} by equation (10), and λ^{k+1} by equation (11). Update I_1^k, I_2^k, I_3^k and I^k by equation (18).

Step 5. Set $k := k + 1$ and go to Step 2 until $\lambda^k = \lambda$.

3. Convergence Analysis

We discuss the convergence of the algorithm under the following assumptions.

Assumption 3.1. (Unique Index Assumption). At each iteration, only one step size reaches γ^k in equation (17), and the step size is made by only one index according to equation (13)–(16).

Assumption 3.2. (Non-degeneration Index Assumption). For all $k = 1, 2, \ldots$, there does not exist an index $i \in \{1, \ldots, n\}$ satisfying

$$\begin{cases} a_i^T A d^k = \operatorname{sgn}(c_i^k), & (19) \\ |c_i^k| = \lambda^k, & (20) \\ x_i^k = 0 \text{ or } |x_i^k| = \kappa / w_i, \ w_i \neq 0, & (21) \\ d_i^k = 0. & (22) \end{cases}$$

Mapping $\lambda \mapsto x^*(\lambda)$, $\mathbb{R}_+ \longrightarrow \mathbb{R}^n$, is called *a regularization path* of model (1) iff $x^*(\lambda)$ is an optimal solution with $\lambda > 0$. For a regularization path $\lambda \mapsto x^*(\lambda)$, the direction of a vector $\bar{d} \in \mathbb{R}^n$ is called *the direction of the regularization path at* $x^*(\bar{\lambda})$ iff there exists $\bar{\gamma} > 0$ such that $x^*(\bar{\lambda}) + \gamma \bar{d} = x^*(\bar{\lambda} - \gamma)$ for any $\gamma \in [0, \bar{\gamma}]$.

A characterization of Condition (6) is presented in the following lemma. We omit its proof.

Lemma 3.1. *Condition (6) holds if and only if*

$$\begin{cases} \operatorname{sgn}(-A_{I_1}^T (Ax - y)) = \operatorname{sgn}(x(I_1)), & (23) \\ |-A_{I_1}^T (Ax - y)|_j \geq \lambda, \ \forall\, j. & (24) \end{cases}$$

We show that the following recursion cases do not occur: 1) an index moved out of the base I_2 enters into the base immediately at the next iteration (Recursion case I and III); 2) an index entered into the base I_2 moves out immediately at the next iteration (Recursion case II and IV).

Recursion case I: In the calculation of γ_{++}^k by equation (16), there exists $i \in I_1^k$ such that i moves from I_1^k to I_2^k and $\gamma_{++}^k = 0$ iff

$$\begin{cases} \lambda^k = |c_i^k|, & (25) \\ \operatorname{sgn}(a_i^T v^k) = \operatorname{sgn}(c_i^k), & (26) \\ |a_i^T v^k| > 1. & (27) \end{cases}$$

If index i_{--}^{k-1} moves from I_2^{k-1} to I_1^{k-1}, and i_{--}^{k-1} satisfies (25) – (27), then i_{--}^{k-1} will move from I_1^k to I_2^k and recursion occurs. Theorem 1 indicates that recursion case I will not occur.

Theorem 1. *At $(k-1)$th iteration, let i_{--}^{k-1} be the index that makes the step size and it moves from I_2^{k-1} to I_1^{k-1}, then i_{--}^{k-1} cannot satisfy (26) and (27) at the same time.*

Similarly, it can be shown that recursion case II–IV will not occur. Theorem 2 illustrates that the homotopy algorithm generates a piecewise linear regularization path of model (1).

Theorem 2. *Under the convergence assumptions, d^k generated by the homotopy algorithm is a direction of regularization path of the box-constrained Lasso at x^k and the sequence of $\{\lambda^k\}$ decreases strictly.*

Proof. Along d^k, the algorithm ensures that $x := x^k + \gamma d^k$ satisfies KKT conditions (2)–(4) with $\lambda = \lambda^k - \gamma$, for all $\gamma \in [0, \gamma^k]$. Initially, $\gamma^0 > 0$. Given that $\gamma^{k-1} > 0$, we need to show $\gamma^k > 0$.

Let $\xi_i^k = \frac{\lambda^k - c_i^k}{1 - a_i^T v^k}$, $(a_i^T v^k \neq 1)$ and $\eta_i^k = \frac{\lambda^k + c_i^k}{1 + a_i^T v^k}$, $(a_i^T v^k \neq -1)$. Define function $\zeta(t) : \mathbb{R} \longrightarrow \mathbb{R}$, $\zeta(t) = \begin{cases} t, & t \geq 0, \\ +\infty, & t < 0, \end{cases}$ and

$$\nu_i^k = \begin{cases} \min\{\zeta(\xi_i^k), \zeta(\eta_i^k)\}, & \text{if } |a_i^T v^k| \neq 1 \\ \eta_i^k, & \text{if } a_i^T v^k = 1, \operatorname{sgn}(c_i^{k-1}) = -1 \\ \xi_i^k, & \text{if } a_i^T v^k = -1, \operatorname{sgn}(c_i^{k-1}) = 1 \end{cases}$$

then we have by (13) that $\gamma_+^k = \min_{i \in I_3^k} \nu_i^k$. Moreover, for any $i \in \{1, \ldots, n\}$,

we have by (11) and $x^k = x^{k-1} + \gamma^{k-1} d^{k-1}$ that

$$\begin{cases} c_i^{k-1} - \gamma^{k-1} a_i^T v^{k-1} = c_i^k, & (28) \\ \lambda^k = \lambda^{k-1} - \gamma^{k-1}. & (29) \end{cases}$$

Denote i^{k-1} the index that moves in or out of the set I_2^{k-1} at $(k-1)$th iteration. It can be shown that $\gamma_+^k > 0$, $\gamma_-^k > 0$, $\gamma_{++}^k > 0$ and $\gamma_{--}^k > 0$. We prove $\gamma_{++}^k > 0$, the other step sizes follow similar arguments.

Suppose for contrary that $\gamma_{++}^k = 0$, which occurs if and only if there exists $i \in I_1^k$ such that (25)–(27) hold. By Theorem 1, we have i^{k-1}, does not satisfy (25)–(27) simultaneously, if $i^{k-1} \in I_1^k$. Hence, by Assumption 3.1, $i \in I_1^k \setminus \{i^{k-1}\} \subseteq I_1^{k-1}$. Together with Lemma 3.1 and $\operatorname{sgn}(x_i^{k-1}) = \operatorname{sgn}(x_i^k) = \operatorname{sgn}(c_i^k)$, we have

$$\begin{cases} |c_i^{k-1}| \geq \lambda^{k-1}, & (30) \\ \operatorname{sgn}(c_i^{k-1}) = \operatorname{sgn}(x_i^{k-1}) = \operatorname{sgn}(c_i^k). & (31) \end{cases}$$

It follows by (25), (29), (30) and $\gamma^{k-1} > 0$ that $|c_i^{k-1}| \geq \lambda^{k-1} = \lambda^k + \gamma^{k-1} > \lambda^k = |c_i^k|$. Combining with (28) and (31) we have

$$\operatorname{sgn}(c_i^{k-1}) = \operatorname{sgn}(a_i^T v^{k-1}). \tag{32}$$

By (25), (28) and (29) we have

$$|c_i^{k-1} - \gamma^{k-1} a_i^T v^{k-1}| = \lambda^{k-1} - \gamma^{k-1}. \tag{33}$$

According to (28), (30), (32) and (33), we have that if $c_i^k > 0$, then $a_i^T v^{k-1} \geq 1$. Note that the case $a_i^T v^{k-1} = 1$, $c_i^{k-1} = \lambda^{k-1}$ does not occur; otherwise, it violates Assumption 3.2 according to $|x_i^{k-1}| = \kappa / w_i$, $d_i^{k-1} = 0$ (because $i \in I_1^k$). Hence, $a_i^T v^{k-1} > 1$. Then, (33) implies that $\gamma^{k-1} = \frac{\lambda^{k-1} - c_i^{k-1}}{1 - a_i^T v^{k-1}}$, $a_i^T v^{k-1} > 1$. Similarly, we have that if $c_i^{k-1} < 0$, $\gamma^{k-1} = \frac{\lambda^{k-1} + c_i^{k-1}}{1 + a_i^T v^{k-1}}$, $a_i^T v^{k-1} < -1$. Then, according to (16) and Assumption 3.1, i makes the step size of γ^{k-1} and should move from I_1^{k-1} to I_2^{k-1} at $(k-1)$th iteration, which contradicts the fact that $i \neq i^{k-1}$ and $i \in I_1^k$.

By equation (17), we have $\gamma^k > 0$, which concludes the proof. □

4. Concluding Remarks

Convergence assumptions are necessary in the algorithm implementation; otherwise, degeneration may occur. For an instance of $\lambda^k = \lambda^{k+1} = \ldots = \lambda^{k+q}$ with $q > 0$, if the index sets and the value of λ at iteration $k + q$ turns out to be the same as those at iteration k, then the algorithm will fall into infinite loops. In this work, we presented mild convergence assumptions and proved that the algorithm generated the entire regularization path in the procedure of iterations.

Acknowledgements

This research is partially supported by Natural Science Foundation of Shandong Province under Grant ZR2014AP004 and Fundamental Research Funds for the Central Universities under Grant 15CX02051A.

References

[1] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. Roy. Stat. Soc.: Series B*, **58** (1996), 267-288.

[2] N. Meinshausen, G. Rocha, and B. Yu, Discussion: A tale of three cousins: Lasso, L2Boosting and Dantzig, *Ann. Stat.*, **35** (2007), 2373-2384.

[3] M. R. Osborne, B. Presnell, and B. A. Turlach, A new approach to variable selection in least squares problems, *IMA J. Numer. Anal.*, **20** (2000), 389-403.

[4] B. Efron, T. Hastie, I. Johnstone, et al. Least angle regression, *Ann. Stat.*, **32** (2004), 407-499.

[5] F. Bach, R. Jenatton, and J. Mairal, et al. Optimization with sparsity-inducing penalties, *Found. Trends Mach. Learn.*, **4** (2011), 1-108.

[6] D. L. Donoho, and Y. Tsaig, Fast solution of l1-norm minimization problems when the solution may be sparse, *IEEE Trans. Inform. Theory*, **54** (2008), 4789-4812.

[7] X. J. Liang, Z. H. Xia, L. P. Pang, et al. Measure prediction capability of data for collaborative filtering. *Knowledge and Information Systems.* (2016), 1-30.