On LARS/Homotopy Equivalence Conditions for Over-Determined LASSO

Junbo Duan, Charles Soussen, David Brie, Jérôme Idier, and Yu-Ping Wang

***Abstract*—We revisit the positive cone condition given by Efron *et al.* [1] for the over-determined least absolute shrinkage and selection operator (LASSO). It is a sufficient condition ensuring that the number of nonzero entries in the solution vector keeps increasing when the penalty parameter decreases, based on which the least angle regression (LARS) [1] and homotopy [2] algorithms yield the same iterates. We show that the positive cone condition is equivalent to the diagonal dominance of the Gram matrix inverse, leading to a simpler way to check the positive cone condition in practice. Moreover, we elaborate on a connection between the positive cone condition and the mutual coherence condition given by Donoho and Tsaig [3], ensuring the exact recovery of any k-sparse representation using both LARS and homotopy.**

***Index Terms*—LASSO, homotopy, LARS, ℓ_1-norm, diagonally dominant, k-step solution property, and positive cone condition.**

I. Introduction

FOR a given signal $\boldsymbol{y} \in \mathbb{R}^m$, we want to estimate the approximation $\boldsymbol{y} \approx \boldsymbol{A}\boldsymbol{u}$, or representation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{u}$, in a given matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. This signal approximation or restoration problem often suffers from instabilities when $\boldsymbol{A}$ is ill-conditioned. To alleviate the instability in the signal restoration problem [4], some constraints must be imposed. In the signal processing community, the following penalized optimization problem has received widespread attention.

$$\boldsymbol{u}^*(\lambda) = \arg\min_{\boldsymbol{u}} \left\{ E(\boldsymbol{u},\lambda) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{u}\|^2 + \lambda\|\boldsymbol{u}\|_1 \right\}, \quad (1)$$

where the penalty parameter λ controls the tradeoff between the approximation error and the model complexity; $\|\cdot\|$ and $\|\cdot\|_1$ stand for the ℓ_2- and ℓ_1-norm respectively. The constrained form of (1) is well known in the literature as the least absolute shrinkage and selection operator (LASSO) [5].

The solution path of optimization problem (1) is defined as the set of all the optimizers w.r.t. the penalty parameter λ: $\{\boldsymbol{u}^*(\lambda) \,|\, \lambda \in (0,\infty)\}$. As a consequence of the piecewise linear property of the solution path [2], efficient algorithms such as the homotopy [2], [6] and the least angle regression (LARS) [1] were developed. The homotopy algorithm starts with $\lambda_0 = +\infty$ (or $\lambda_1 = \|\boldsymbol{A}^T\boldsymbol{y}\|_\infty$ [6], where $\|\cdot\|_\infty$ is the uniform norm) and decreases λ gradually. During each iteration, a new critical value of λ (i.e., λ_p) and the corresponding $\boldsymbol{u}^*(\lambda_p)$ are calculated from the previous values. We note that at each iteration of the homotopy algorithm, the active set $\mathcal{I}(\boldsymbol{u}) = \{i \,|\, u_i \neq 0\}$ is maintained and the nonzero entries of $\boldsymbol{u}$ are updated. When the iteration is completed, either some u_i changes from zero to nonzero (i is appended to $\mathcal{I}$) or is removed on the contrary (i is removed from $\mathcal{I}$). In that respect, the homotopy algorithm is a forward-backward algorithm. On the contrary, the LARS[1] is just a forward algorithm since only insertions into the active set are allowed.

For over-determined systems ($m > n$), the positive cone condition (PCC) introduced by Efron *et al.* implies the monotonic increase of the active set cardinality when λ decreases ([1], Theorem 4). Meinshausen [7] showed a stronger result: if the PCC is fulfilled, not only the cardinality $\mathrm{Card}[\mathcal{I}]$ (number of elements in the active set $\mathcal{I}$) increases monotonically, but also the amplitudes $|u_i(\lambda)|$ increase monotonically when λ decreases. The recent work of Tibshirani *et al.* [8], [9] shows that if the inverse of the Gram matrix of $\boldsymbol{A}$, i.e., $(\boldsymbol{A}^T\boldsymbol{A})^{-1}$, is diagonally dominant, the so-called 'slope bound' holds, which is similar to the monotonic increasing property. The relation between the PCC and diagonally dominant condition (DDC) was already noticed by Meinshausen and Yu [10]; however, a clear characterization of their relations is still lacking. In this paper, we show that the PCC is equivalent to the strictly diagonally dominant condition (SDDC).

For under-determined systems ($m < n$), Donoho and Tsaig [3] derived the mutual coherence condition (MCC) on $\boldsymbol{A}$ and $k = \|\boldsymbol{u}\|_0$. Under this condition the LARS and homotopy have the so-called k-step solution property, i.e., any k-sparse representation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{u}$ with $\|\boldsymbol{u}\|_0 = k$ can be exactly recovered in k steps, that is, by performing k insertions. The k-step solution property implies the monotonic increase of the cardinality $\mathrm{Card}[\mathcal{I}]$ when λ decreases. Donoho and Tsaig's results [3] were essentially dedicated to under-determined systems. We note that the notion of the k-step solution property and MCC, can be naturally extended to the over-determined systems, and therefore can be connected with the DDC and PCC.

This paper is organized as follows: In Section II, we show that the PCC and SDDC are equivalent. In Section III, we establish a connection with the MCC. We conclude the paper in Section IV.

Manuscript received August 14, 2012; accepted September 25, 2012. Date of publication October 02, 2012; date of current version October 24, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jared W. Tanner.

J. Duan and Y.-P. Wang are with the Department of Biomedical Engineering, Tulane University, New Orleans, LA 70118 USA (e-mail: jduan@tulane.edu; wyp@tulane.edu).

C. Soussen and D. Brie are with the Université de Lorraine and the CNRS at the CRAN (Centre de Recherche en Automatique de Nancy), UMR 7039, Campus Sciences, BP 70239 Vandoeuvre-lès-Nancy Cedex 54506, France (e-mail: charles.soussen@univ-lorraine.fr; david.brie@univ-lorraine.fr).

J. Idier is with Institut de Recherche en Communication et Cybernétique de Nantes, Nantes, France (e-mail: jerome.idier@irccyn.ec-nantes.fr).

Digital Object Identifier 10.1109/LSP.2012.2221712

[1]Here we refer to the original version of LARS. The modified version of LARS [1] enabling the removal of indices from the active set $\mathcal{I}$, is equivalent to homotopy.

II. The Positive Cone and Diagonally Dominant Conditions

The Positive Cone Condition (PCC) [1]: We say that $\boldsymbol{A}$ obeys the PCC if, for any diagonal matrix $\boldsymbol{B}$ whose diagonal elements are ± 1, the sum of each row of the inverse matrix of any principal minor of $\boldsymbol{B}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{B}$ is positive.

Theorem 1 ([7], Lemma 12): Let $\boldsymbol{A}$ be a full rank matrix of size $m \times n$, with $m \geqslant n$. Under the PCC[2], for any $\boldsymbol{y} \in \mathbb{R}^m$, the absolute value of each entry in $\boldsymbol{u}^*(\lambda)$ monotonically increases when λ decreases.

Definition 1: A matrix $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ is called:

- (row) diagonally dominant (DD)[3] if $h_{ii} \geqslant \sum_{j\neq i} |h_{ij}| (i = 1, \ldots, n)$;
- strictly diagonally dominant (SDD) if $h_{ii} > \sum_{j\neq i} |h_{ij}| (i = 1, \ldots, n)$.

Tibshirani *et al.* [8] showed that if $(\boldsymbol{A}^T\boldsymbol{A})^{-1}$ is DD, the so-called 'slope bound' holds, which is similar to the monotonic increasing property. Their result can be stated as the diagonally dominant condition (DDC) in Theorem 2. Their proof was derived from the work of Tibshirani and Taylor [9], in which an alternative cost function was used. In the Appendix of this paper, we present an alternative but simpler proof by considering the formulation (1) directly.

Theorem 2 ([8], Theorem 1): For full rank matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n} (m \geqslant n)$, in optimization problem (1), if $(\boldsymbol{A}^T\boldsymbol{A})^{-1}$ is DD, then for any $\boldsymbol{y} \in \mathbb{R}^m$, $\mathrm{Card}[\mathcal{I}(\boldsymbol{u}^*(\lambda))]$ is monotonically increasing[4] when λ decreases.

In Theorem 3, we show that the PCC is equivalent to the strictly diagonally dominant condition (SDDC) on $(\boldsymbol{A}^T\boldsymbol{A})^{-1}$.

Definition 2 (Notations): $\boldsymbol{0}_{k\times n} \in \mathbb{R}^{k\times n}$ is the null matrix; $\boldsymbol{I}_n \in \mathbb{R}^{n\times n}$ is the identity matrix; $\boldsymbol{J}_{k\times n} = [\boldsymbol{I}_k, \boldsymbol{0}_{k\times(n-k)}] \in \mathbb{R}^{k\times n}$ is a dimension reduction (from n to k) matrix; $\boldsymbol{P}$ is a square permutation matrix whose size depends on the context; and $\boldsymbol{P}^T$ is the transpose of $\boldsymbol{P}$.

Theorem 3: The PCC is equivalent to the SDDC on $\boldsymbol{H} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}$.

Proof:

- PCC $\Rightarrow$ SDDC on $(\boldsymbol{A}^T\boldsymbol{A})^{-1}$
  Each principal minor of $\boldsymbol{B}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{B}$ can be written as $\boldsymbol{M} = \boldsymbol{J}_{k\times n}\boldsymbol{P}\boldsymbol{B}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{B}\boldsymbol{P}^T\boldsymbol{J}_{k\times n}^T$, the PCC demands that for any $\boldsymbol{P}, \boldsymbol{B}$ and for all $k = 1, \ldots, n$, the sum of each row of $\boldsymbol{M}^{-1}$ is positive. For the configuration where $\boldsymbol{P}$ is the identity matrix and $k = n$, the sum of the i-th row of $(\boldsymbol{B}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^T\boldsymbol{H}\boldsymbol{B}$ can be written as $\sum_{j=1}^n b_{ii}b_{jj}h_{ij} = h_{ii} + \sum_{j\neq i} b_{ii}b_{jj}h_{ij}$; the PCC reads $h_{ii} + \sum_{j\neq i} b_{ii}b_{jj}h_{ij} > 0$. Because b_{ii} and b_{jj} can be either $+1$ or -1, proper choice of b_{ii} and b_{jj} yields $h_{ii} > \sum_{j\neq i} |h_{ij}|$, i.e., $\boldsymbol{H} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}$ is SDD.
- SDDC on $(\boldsymbol{A}^T\boldsymbol{A})^{-1} \Rightarrow$ PCC
  $\boldsymbol{H} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}$ being SDD yields $h_{ii} > \sum_{j\neq i} |h_{ij}| \geqslant \sum_{j\neq i} b_{ii}b_{jj}h_{ij}$ for any configuration of $\boldsymbol{B}$. So the PCC is true for $k = n$. From Lemma 3 in the Appendix, which can be extended to SDD straightforward, the inverse matrix of each principal minor of $\boldsymbol{A}^T\boldsymbol{A}$ is also SDD. So the PCC is true for $k < n$.

∎

[2]Although Meinshausen used term 'restricted PCC' in ([7], Lemma 12), we note that it is equivalent to the PCC for the over-determined systems.

[3]The typical definition of a diagonally dominant matrix relies on the absolute value of h_{ii}; however, in this paper, $\boldsymbol{H}$ is always a positive definite matrix, thus the positiveness of h_{ii} is implied.

[4]Here the 'monotonically increasing' has to be understood as 'monotonically non-decreasing'.

III. The Connections With Mutual Coherence Condition

In this section, we elaborate the connection with the mutual coherence condition. First we introduce the k-step solution property.

The k-Step Solution Property [3]: Let $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] \in \mathbb{R}^{m\times n}$ be an under-determined system and $\tilde{\boldsymbol{y}} = \boldsymbol{A}\tilde{\boldsymbol{u}}$ be a vector admitting a k-sparse representation, i.e., $\tilde{\boldsymbol{u}}$ has k nonzero entries. We say that an algorithm satisfies the k-step solution property for the given problem $(\boldsymbol{A}, \tilde{\boldsymbol{y}})$ if it can get the correct solution $\tilde{\boldsymbol{u}}$ with at most k-steps.

Here, we use the notation $\tilde{\boldsymbol{y}}$ instead of $\boldsymbol{y}$ to stress that the results in [3] are dedicated to the noise-free setting. Donoho and Tsaig [3] presented the following sufficient condition under which the LARS and homotopy algorithms satisfy the k-step solution property.

The Mutual Coherence Condition (MCC) [3]: Let $\boldsymbol{A}$ be an under-determined system. We say that a problem instance $(\boldsymbol{A}, \tilde{\boldsymbol{y}})$ satisfies the mutual coherence condition if the sparsity level k obeys

$$k \leqslant \frac{1+\mu^{-1}}{2}, \tag{2}$$

where μ is the mutual coherence of $\boldsymbol{A}$:

$$\mu = \max_{i\neq j} \frac{|\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle|}{\|\boldsymbol{a}_i\|\|\boldsymbol{a}_j\|}.$$

Theorem 4 ([3], Theorem 1 and Corollary 2): Let $\boldsymbol{A}$ be a matrix of size $m \times n$ with $m < n$. If a problem instance $(\boldsymbol{A}, \tilde{\boldsymbol{y}})$ satisfies the MCC, then the LARS and homotopy algorithms run in k steps and stop, delivering the solution $\tilde{\boldsymbol{u}}$.

Throughout this section, $\boldsymbol{a}_i$ is normalized for convenience, i.e., $\|\boldsymbol{a}_i\| = 1$.

Although the k-step solution property and MCC were originally introduced in the context of the under-determined systems, they can be obviously extended to the over-determined systems (see the proof of ([3], Theorem 1 and Corollary 2)).

Corollary 1: Let $\boldsymbol{A}$ be a matrix of size $m \times n$ with $m \geqslant n$. If a problem instance $(\boldsymbol{A}, \tilde{\boldsymbol{y}})$ satisfies the MCC, then the LARS and homotopy algorithms run in k steps and stop, delivering the solution $\tilde{\boldsymbol{u}}$.

Therefore, in the remainder of this paper, the MCC and k-step solution property are also considered in the context of the over-determined systems.

In the case of homotopy algorithm, we remark that the k-step solution property states that homotopy finds the support of $\tilde{\boldsymbol{u}}$ in k iterations, thus homotopy never removes entries from the active set. In other words, the number of nonzero entries of the solution vector monotonically increases at each iteration. Herein, the MCC, the PCC and the SDDC/DDC are connected *via* the k-step solution property, as illustrated by Fig. 1.

The MCC reflects that low correlated matrices $\boldsymbol{A}$ (small μ's) enable the recovery of vectors $\boldsymbol{u}$ whose number of nonzero

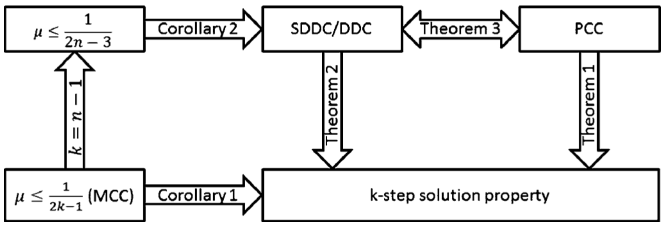

Fig. 1. The connections between mutual coherence condition (MCC) (for the over-determined systems), positive cone condition (PCC) and the (strictly) diagonally dominant condition (SDDC/DDC).

entries is large. In the extreme case where $k = n-1$, MCC (2) rereads

$$\mu \leqslant \frac{1}{2n-3}, \tag{3}$$

which links the MCC and the DDC conditions by the following theorem and corollary.

Theorem 5: For full rank symmetric matrix $\boldsymbol{G} \in \mathbb{R}^{n\times n} (n > 2)$, if $g_{ii} > 0$ and $|g_{ij}|/g_{ii} \leqslant 1/(2n-3)$, then $\boldsymbol{G}$ is invertible and $\boldsymbol{G}^{-1}$ is SDD.

Proof: If $|g_{ij}|/g_{ii} \leqslant 1/(2n-3) (j \neq i)$, $\sum_{j\neq i} |g_{ij}|/g_{ii} \leqslant (n-1)/(2n-3) < 1$ for $n > 2 \Rightarrow$ $\boldsymbol{G}$ is SDD $\Rightarrow$ $\boldsymbol{G}$ is positive definite and nonsingular [11] $\Rightarrow$ its inverse $\boldsymbol{H} = \boldsymbol{G}^{-1}$ is also positive definite $\Rightarrow h_{ii} > 0$. From $\boldsymbol{HG} = \boldsymbol{I}$ we have:

$$\delta_{ij} = \sum_v h_{iv} g_{vj} = \sum_{v\neq j} h_{iv} g_{vj} + h_{ij} g_{jj},$$

here δ_{ij} is Kronecker symbol, and

$$\begin{aligned}
&\sum_{j\neq i} |h_{ij}| \\
&= \sum_{j\neq i} \left| \frac{\delta_{ij} - \sum_{v\neq j} h_{iv} g_{vj}}{g_{jj}} \right| \\
&= \sum_{j\neq i} \left| - \sum_{v\neq j} \frac{g_{vj}}{g_{jj}} h_{iv} \right| \\
&\leqslant \frac{1}{2n-3} \sum_{j\neq i} \sum_{v\neq j} |h_{iv}| \\
&= \frac{1}{2n-3} \left(\sum_{j\neq i} \sum_{v} |h_{iv}| - \sum_{j\neq i} |h_{ij}| \right) \\
&= \frac{1}{2n-3} \left((n-2) \sum_{j\neq i} |h_{ij}| + (n-1)|h_{ii}| \right).
\end{aligned}$$

By moving $\sum_{j\neq i} |h_{ij}|$ from the right to the left hand side, we have $\sum_{j\neq i} |h_{ij}| < |h_{ii}| = h_{ii}$, so $\boldsymbol{G}^{-1} = \boldsymbol{H}$ is SDD. ■

Corollary 2: Any symmetric matrix $\boldsymbol{G} \in \mathbb{R}^{n\times n}$ with $g_{ii} = 1$,$|g_{ij}| \leqslant 1/(2n-3) (i \neq j)$ is invertible and $\boldsymbol{G}^{-1}$ is SDD.

Corollary 2 coincides with the upper bound on μ in (3). As μ is equal to the maximum absolute value of the off-diagonal entries of the Gram matrix $\boldsymbol{G} = \boldsymbol{A}^T\boldsymbol{A}$, $\mu \leqslant 1/(2n-3)$ implies that $(\boldsymbol{A}^T\boldsymbol{A})^{-1}$ is DD. Therefore, the MCC and SDDC/DDC are connected *via* Corollary 2, as demonstrated in Fig. 1.

Three points need further discussion: first, the k-step solution property describes the performances of the given algorithms, while the monotonic increase of $\mathrm{Card}[\mathcal{I}(\boldsymbol{u}^*(\lambda))]$ with the decrease of λ describes the property of the solution path of Problem (1). As homotopy was proved to be able to find the solution path of Problem (1) [6], the k-step solution property is an indirect characterization of the monotonic increase of the entries of the solution vector. Therefore the MCC yields the k-step solution property directly, implying monotonic increase indirectly. On the contrary, the PCC and the SDDC/DDC yield monotonic increase directly. Second, as mentioned before, the MCC can be applied to both under-determined and over-determined systems, while the PCC and the SDDC/DDC are only applicable to over-determined systems. Third, the MCC was derived in the noise-free case, while PCC and the SDDC/DDC can be applied to the noisy case. However, MCC is more relaxable w.r.t. the sparsity of the signal when the sparsity is known *a priori*, as shown in (2), while the PCC and the SDDC/DDC are restrictive and independent of the sparsity of the signal. In fact, the noisy case can be considered as noise-free in the extreme case when $\tilde{\boldsymbol{u}}$ is not sparse at all, which was discussed in this section.

IV. Conclusion

In this paper, we study the conditions concerning the solution path of the over-determined LASSO problem. We showed that two conditions, namely, the PCC and the SDDC, are equivalent. Under either of them, the number of nonzero entries in the optimizer of over-determined LASSO increases monotonically when the penalty parameter decreases. In practice, this means the 'forward' algorithm LARS yields the same solution path as the 'forward-backward' algorithm homotopy. Based on this fact, the computation in homotopy algorithm can thus be reduced. Through the equivalence, we also have a practical way of verifying the PCC, avoiding going through all possible configurations as suggested by the PCC definition. Furthermore, we showed that the SDDC/DDC and the PCC are related to the MCC, which yields the k-step property for the over-determined systems.

Appendix

First, we introduce Lemmas 1 and 2 to prove Lemma 3. Finally, based on Lemma 3, we give the proof of Theorem 2.

Lemma 1: If full rank symmetric matrix $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ is DD, then $\boldsymbol{R} = (\boldsymbol{J}_{(n-1)\times n} \boldsymbol{H}^{-1} \boldsymbol{J}^T_{(n-1)\times n})^{-1}$ is also DD.

Proof: Define

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{12}^T & \boldsymbol{H}_{22} \end{bmatrix}, \boldsymbol{G} = \boldsymbol{H}^{-1} = \begin{bmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{12}^T & \boldsymbol{G}_{22} \end{bmatrix},$$

where $\boldsymbol{H}_{11}, \boldsymbol{G}_{11}$ are of size $(n-1)\times(n-1)$. From block matrix inversion lemma [?], we know

$$\boldsymbol{R} = \boldsymbol{G}_{11}^{-1} = \boldsymbol{H}_{11} - \boldsymbol{H}_{12} \boldsymbol{H}_{22}^{-1} \boldsymbol{H}_{12}^T,$$

so the entry of $\boldsymbol{R}$ reads $r_{ij} = h_{ij} - (h_{in} h_{jn})/(h_{nn})$.

$$\begin{aligned}
\sum_{j\neq i,n} |r_{ij}| &= \sum_{j\neq i,n} \left| h_{ij} - \frac{h_{in} h_{jn}}{h_{nn}} \right| \\
&\leqslant \sum_{j\neq i,n} |h_{ij}| + \frac{|h_{in}|}{h_{nn}} \sum_{j\neq i,n} |h_{jn}| \\
&\overset{\mathrm{DD}}{\leqslant} \sum_{j\neq i,n} |h_{ij}| + \frac{|h_{in}|}{h_{nn}} (h_{nn} - |h_{in}|) \\
&= \sum_{j\neq i} |h_{ij}| - \frac{h_{in}^2}{h_{nn}} \overset{\mathrm{DD}}{\leqslant} h_{ii} - \frac{h_{in}^2}{h_{nn}} = r_{ii},
\end{aligned}$$

so $\boldsymbol{R}$ is DD. 'DD' above $\leqslant$ indicates that the DD condition $h_{ii} \geqslant \sum_{j\neq i} |h_{ij}| \geqslant 0$ is applied. ∎

Lemma 2: If full rank symmetric matrix $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ is DD, then $\boldsymbol{R}_k = (\boldsymbol{J}_{k\times n}\boldsymbol{H}^{-1}\boldsymbol{J}_{k\times n}^T)^{-1}$ is also DD for all $k = 1,\ldots,n-1$.

Proof: $\boldsymbol{H}$ is full rank symmetric, so $\boldsymbol{R}_i (i = 1,\ldots,n-1)$ is also full rank symmetric. By using Lemma 1 recursively: $\boldsymbol{H}$ is DD $\Rightarrow \boldsymbol{R}_{n-1}$ is DD $\Rightarrow \boldsymbol{R}_{n-2}$ is DD $\Rightarrow \cdots\cdots \Rightarrow \boldsymbol{R}_1$ is DD. ∎

Lemma 3 (DD Preservation Property): If a full rank symmetric matrix $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ is DD, then $(\boldsymbol{J}_{k\times n}\boldsymbol{P}\boldsymbol{H}^{-1}\boldsymbol{P}^T\boldsymbol{J}_{k\times n}^T)^{-1}$ is also DD for any $\boldsymbol{P}$ and for all $k = 1,\ldots,n$.

From Lemma 2, Lemma 3 is straightforward. The same result can also be found in [12].

Based on Lemma 3, the proof of Theorem 2 is as follows.

Proof of Theorem 2: The subdifferential of $E(\boldsymbol{u},\lambda)$ is:

$$\partial E(\boldsymbol{u},\lambda) = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{u}(\lambda) - \boldsymbol{y}) + \lambda_(\lambda).$$

Here $\boldsymbol{s}$ is the subdifferential of $\|\boldsymbol{u}\|_1$, which is defined as:

$$\boldsymbol{s} = \partial\|\boldsymbol{u}\|_1 = \begin{cases} s_i = 1, & \text{if } u_i > 0; \\ s_i = -1, & \text{if } u_i < 0; \\ s_i \in [-1,1], & \text{otherwise.} \end{cases}$$

A necessary condition for $\boldsymbol{u}^*$ to be a minimizer of the optimization problem (1) is to have $\boldsymbol{0} \in \partial E(\boldsymbol{u}^*,\lambda)$. Therefore, we have the following system:

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{u}^*(\lambda) + \lambda\boldsymbol{s}^*(\lambda) = \boldsymbol{A}^T\boldsymbol{y}. \tag{4}$$

Because $\boldsymbol{u}^*(\lambda)$ is piecewise linear [2], for each piece $[\lambda_p,\lambda_{p-1})$, the entries of $\boldsymbol{s}^*(\lambda)$ corresponding to the nonzero entries in $\boldsymbol{u}^*(\lambda)$ are constant. Thus we can find a permutation $\boldsymbol{P}$ such that the nonzero entries and zero entries in $\boldsymbol{u}^*$ are rearranged to be the first $\boldsymbol{u}^*_{\text{on}}(\neq \boldsymbol{0})$ and last entries $\boldsymbol{u}^*_{\text{off}}(= \boldsymbol{0})$ respectively. In the following, we omit the dependency with respect to λ for the sake of brevity.

$$\begin{aligned} \boldsymbol{u}^* &= \boldsymbol{P}^T\begin{bmatrix}\boldsymbol{u}^*_{\text{on}}\\ \boldsymbol{u}^*_{\text{off}}\end{bmatrix}, \\ \boldsymbol{s}^* &= \boldsymbol{P}^T\begin{bmatrix}\boldsymbol{s}^*_{\text{on}}\\ \boldsymbol{s}^*_{\text{off}}\end{bmatrix}, \boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{P}^T\begin{bmatrix}\boldsymbol{x}_{\text{on}}\\ \boldsymbol{x}_{\text{off}}\end{bmatrix} \end{aligned} \tag{5}$$

By substituting (5) into (4), and left multiplying $\boldsymbol{P}$, since $\boldsymbol{P}^T = \boldsymbol{P}^{-1}$, we have

$$\boldsymbol{P}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{P}^T\begin{bmatrix}\boldsymbol{u}^*_{\text{on}}\\ \boldsymbol{0}\end{bmatrix} + \lambda\begin{bmatrix}\boldsymbol{s}^*_{\text{on}}\\ \boldsymbol{s}^*_{\text{off}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{x}_{\text{on}}\\ \boldsymbol{x}_{\text{off}}\end{bmatrix},$$

which can be rewritten as

$$\begin{bmatrix}\boldsymbol{\Psi} & \boldsymbol{\Upsilon}\\ \boldsymbol{\Upsilon}^T & \boldsymbol{\Phi}\end{bmatrix}\begin{bmatrix}\boldsymbol{u}^*_{\text{on}}\\ \boldsymbol{0}\end{bmatrix} + \lambda\begin{bmatrix}\boldsymbol{s}^*_{\text{on}}\\ \boldsymbol{s}^*_{\text{off}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{x}_{\text{on}}\\ \boldsymbol{x}_{\text{off}}\end{bmatrix}, \tag{6}$$

where $\boldsymbol{\Psi} = \boldsymbol{J}_{k\times n}\boldsymbol{P}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{P}^T\boldsymbol{J}_{k\times n}^T$ and k is the length of $\boldsymbol{u}^*_{\text{on}}$. Under the condition that $(\boldsymbol{A}^T\boldsymbol{A})^{-1}$ is DD, from Lemma 3, $\boldsymbol{R} \triangleq \boldsymbol{\Psi}^{-1}$ is DD. From (6) we have

$$\frac{d\boldsymbol{u}^*_{\text{on}}}{d\lambda} = -\boldsymbol{R}\boldsymbol{s}^*_{\text{on}}.$$

For the i-th entry of $\boldsymbol{u}^*_{\text{on}}$, i.e., $u^*_{\text{on},i} (i = 1,\ldots,k)$

$$\frac{du^*_{\text{on},i}}{d\lambda} = -\sum_{j=1}^{k} r_{ij}s^*_{\text{on},j} = -r_{ii}s^*_{\text{on},i} - \sum_{j\neq i} r_{ij}s^*_{\text{on},j}$$

For positive $u^*_{\text{on},i}$, since $s^*_{\text{on},i} = 1$ and $s^*_{\text{on},j} \in \{-1,1\}$,

$$\begin{aligned} \frac{du^*_{\text{on},i}}{d\lambda} &= -r_{ii} - \sum_{j\neq i} r_{ij}s^*_{\text{on},j} \\ &\overset{\text{DD}}{\leqslant} -\sum_{j\neq i}(|r_{ij}| + r_{ij}s^*_{\text{on},j}) \leqslant 0 \end{aligned}$$

For the negative case, the derivative is non-negative. We can see that $|u^*_{\text{on},i}(\lambda)|$ non-increases monotonically as λ increases, while $|u^*_{\text{off},i}(\lambda)|$ is equal to zero in piece $[\lambda_p,\lambda_{p-1})$.

Because $u^*_i(\lambda)$ is continuous for $\lambda > 0$ [2], it is straightforward to extend the result to all λ: when λ increases, the absolute values of the nonzero entries in $\boldsymbol{u}^*(\lambda)$ decrease and tend towards zero, while the zero entries remain zeros. Therefore, $\mathrm{Card}[\mathcal{I}(\boldsymbol{u}^*(\lambda))]$ non-increases monotonically when λ increases. In other words, $\mathrm{Card}[\mathcal{I}(\boldsymbol{u}^*(\lambda))]$ increases monotonically when λ decreases. ∎

REFERENCES

[1] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[2] M. R. Osborne, B. Presnell, and B. A. Turlach, "A new approach to variable selection in least squares problems," *IMA J. Numer. Anal.*, vol. 20, no. 3, pp. 389–403, 2000.

[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, Nov. 2008.

[4] C. Soussen, J. Idier, D. Brie, and J. Duan, "From Bernoulli-Gaussian deconvolution to sparse signal restoration," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4572–4584, 2011.

[5] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. R. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[6] D. M. Malioutov, M. Cetin, and A. S. Willsky, "Homotopy continuation for sparse signal representation," in *Proc. IEEE ICASSP*, Philadephia, PA, Mar. 2005, vol. V, pp. 733–736.

[7] N. Meinshausen, "Relaxed Lasso," *Comput. Statist. Data Anal.*, vol. 52, no. 1, pp. 374–393, Sep. 2007.

[8] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani, "Strong rules for discarding predictors in lasso-type problems," *J. R. Statist. Soc. B*, vol. 74, pp. 245–266, 2012.

[9] R. Tibshirani and J. Taylor, "The solution path of the generalized lasso," *Ann. Statist.*, vol. 39, no. 3, pp. 1335–1371, 2011.

[10] N. Meinshausen, G. Rocha, and B. Yu, "Discussion: A tale of three cousins: Lasso, L2Boosting and Dantzig," *Ann. Statist.*, vol. 35, no. 6, pp. 2373–2384, 2007.

[11] D. Bernstein, *Matrix Mathematics*. Princeton, NJ: Princeton Univ. Press, 2009.

[12] D. Carlson and T. Markham, "Schur complements of diagonally dominant matrices," *Czech. Math. J.*, vol. 29, no. 2, pp. 246–251, 1979.