Submitted to the Annals of Statistics

SUPPLEMENT A: PROOFS AND TECHNICAL DETAILS FOR "THE SOLUTION PATH OF THE GENERALIZED LASSO"

By Ryan J. Tibshirani and Jonathan Taylor

Stanford University

In this document we give supplementary details to the paper "The Solution Path of the Generalized Lasso". We use the prefix "GL" when referring to equations, sections, etc. in the original paper, as in equation (GL-1) or Section GL-1 (this stands for *G*eneralized *L*asso).

1. Proof of the boundary lemma. We prove the boundary lemma when $D = D_{1\mathrm{d}}$, but first we give a helpful lemma.

Lemma 1. *Let T_λ denote the function that truncates outside of the interval $[-\lambda, \lambda]$:*

$$T_\lambda(x) = \begin{cases} -\lambda & \text{if } x < -\lambda \\ x & \text{if } |x| \leq \lambda \\ \lambda & \text{if } x > \lambda. \end{cases}$$

Then for any λ_0, λ and x, y,

$$|T_{\lambda_0}(x) - T_\lambda(y)| \leq \max\{|x - y|, |\lambda_0 - \lambda|\}.$$

Proof. Suppose without a loss of generality that $\lambda_0 > \lambda$. We enumerate the possible cases:

- $x > \lambda,\ y > \lambda:\ \ |T_{\lambda_0}(x) - T_\lambda(y)| \leq \lambda_0 - \lambda$;
- $x \leq \lambda,\ y > \lambda:\ \ |T_{\lambda_0}(x) - T_\lambda(y)| \leq |x - y|$;
- $x > \lambda,\ y \leq \lambda:\ \ |T_{\lambda_0}(x) - T_\lambda(y)| \leq |x - y|$;
- $|x| \leq \lambda,\ |y| \leq \lambda:\ \ |T_{\lambda_0}(x) - T_\lambda(y)| = |x - y|$.

The remaining cases follow by symmetry. □

Proof of the boundary lemma. Our approach for the proof is a little unusual: we consider the use of coordinate descent to find the solution $\hat{u}_\lambda$, starting at the point $\hat{u}_{\lambda_0}$ as an initial guess. Because the coordinate updates are especially simple, we can track how the iterates change, and hence we can guarantee that $\hat{u}_\lambda$ and $\hat{u}_{\lambda_0}$ are close together. Namely, we show that

$$\|\hat{u}_{\lambda_0} - \hat{u}_\lambda\|_\infty \leq \lambda_0 - \lambda,$$

which implies the desired result.

First we describe the coordinate descent updates for finding the solution $\hat{u}_\lambda$ of the dual (GL-13), when $D = D_{1\mathrm{d}}$. We note that any limit point of the coordinate descent algorithm is indeed a solution by Theorem 4.1 of [1]. We take $u^{(0)} = \hat{u}_{\lambda_0}$ as an initial guess, and cycle through the coordinates in the order $i = 1, \ldots n - 1$. To derive the ith update, we fix u_j for all $j \neq i$ and minimize over u_i. Because of the simple structure of D, we only need to consider two terms:

$$\underset{u_i}{\text{minimize}}\ \frac{1}{2}\big(y_i - (u_i - u_{i-1})\big)^2 + \frac{1}{2}\big(y_i - (u_{i+1} - u_i)\big)^2 \ \text{ subject to } |u_i| \leq \lambda.$$

This is just a quadratic constrained to lie in an interval, and so the ith coordinate update is

$$u_i \leftarrow T_\lambda\left(\frac{y_{i+1} - y_i + u_{i+1} + u_{i-1}}{2}\right),$$

where we let $u_0 = u_n = 0$ for notational convenience.

Therefore in the first iteration of the coordinate descent algorithm, we get

$$u_i^{(1)} = T_\lambda\left(\frac{y_{i+1} - y_i + u_{i+1}^{(0)} + u_{i-1}^{(1)}}{2}\right).$$

Using the fact that $\hat{u}_{\lambda_0}$ is itself the solution corresponding to λ_0,

$$|\hat{u}_{\lambda_0,i} - u_i^{(1)}| = \left|T_{\lambda_0}\left(\frac{y_{i+1} - y_i + \hat{u}_{\lambda_0,i+1} + \hat{u}_{\lambda_0,i-1}}{2}\right) - T_\lambda\left(\frac{y_{i+1} - y_i + u_{i+1}^{(0)} + u_{i-1}^{(1)}}{2}\right)\right|.$$

But this is $\leq \max\{|\hat{u}_{\lambda_0,i-1} - u_{i-1}^{(1)}|/2,\ \lambda_0 - \lambda\}$ by the helpful lemma. Therefore, by induction, it follows that $\|\hat{u}_{\lambda_0} - u^{(1)}\|_\infty \leq \lambda_0 - \lambda$.

Continuing the same line of argument shows that $\|\hat{u}_{\lambda_0} - u^{(k)}\|_\infty \leq \lambda_0 - \lambda$ for all iterations k. Letting $k \to \infty$, we get $\|\hat{u}_{\lambda_0} - \hat{u}_\lambda\|_\infty \leq \lambda_0 - \lambda$, as desired. □

It is important to note that if DD^T is diagonally dominant, in other words

$$(DD^T)_{ii} \geq \sum_{j \neq i} |(DD^T)_{ij}|$$

for each $i = 1, \ldots m$, then the proof of the boundary lemma is similar to that given for the 1d fused lasso case. The coordinate updates are now

$$u_i \leftarrow T_\lambda\left(\frac{(Dy)_i - \sum_{j \neq i}(DD^T)_{ij} u_j}{(DD^T)_{ii}}\right),$$

but the rest of the proof remains the same, so the boundary lemma still holds.

2. Derivation details for Algorithm GL-2. This section is divided into two parts: 1) details for the algorithm's steps at each iteration, and 2) the insertion-deletion lemma. The first part relies on the insertion-deletion lemma when verifying the KKT conditions (hence establishing the algorithm's correctness), and we present and prove this lemma in the second part for the sake of clarity. The insertion-deletion lemma also proves that constructed solution path is continuous over λ.

2.1. *The algorithm at the kth iteration.* We propose a solution $\hat{u}_\lambda = f(\lambda)$, with $\gamma = g(\lambda)$ and $\alpha = h(\lambda)$, in order to satisfy the KKT conditions. First we define

$$\begin{aligned} f(\lambda)_\mathcal{B} &= \lambda s \\ f(\lambda)_{-\mathcal{B}} &= \big(D_{-\mathcal{B}}(D_{-\mathcal{B}})^T\big)^+ D_{-\mathcal{B}}\big(y - \lambda(D_\mathcal{B})^T s\big). \end{aligned}$$

Next we define $g(\lambda)$ and $h(\lambda)$ to satisfy the stationarity equation (GL-24). We examine this in two blocks: the interior coordinates, $-\mathcal{B}$, and the boundary coordinates, $\mathcal{B}$. For the first block, the equation is:

$$\begin{aligned}
&\big(DD^T f(\lambda)\big)_{-\mathcal{B}} - (Dy)_{-\mathcal{B}} + \alpha\gamma_{-\mathcal{B}} \\
&= \lambda D_{-\mathcal{B}}(D_{\mathcal{B}})^T s + D_{-\mathcal{B}}(D_{-\mathcal{B}})^T \big(D_{-\mathcal{B}}(D_{-\mathcal{B}})^T\big)^+ D_{-\mathcal{B}}\big(y - \lambda(D_{\mathcal{B}})^T s\big) - (Dy)_{-\mathcal{B}} + \alpha\gamma_{-\mathcal{B}} \\
&= \alpha\gamma_{-\mathcal{B}}. \qquad (1)
\end{aligned}$$

Now for the second block:

$$\begin{aligned}
&\big(DD^T f(\lambda)\big)_{\mathcal{B}} - (Dy)_{\mathcal{B}} + \alpha\gamma_{\mathcal{B}} \\
&= D_{\mathcal{B}}\Big[I - (D_{-\mathcal{B}})^T \big(D_{-\mathcal{B}}(D_{-\mathcal{B}})^T\big)^+ D_{-\mathcal{B}}\Big]\big(y - \lambda(D_{\mathcal{B}})^T s\big) + \alpha\gamma_{\mathcal{B}}. \qquad (2)
\end{aligned}$$

We want to choose $\gamma = g(\lambda)$ and $\alpha = h(\lambda)$ to make both (1) and (2) equal to zero. Consider defining

$$h(\lambda) = \Big\| D_{\mathcal{B}}\Big[I - (D_{-\mathcal{B}})^T \big(D_{-\mathcal{B}}(D_{-\mathcal{B}})^T\big)^+ D_{-\mathcal{B}}\Big]\big(y - \lambda(D_{\mathcal{B}})^T s\big)\Big\|_1.$$

If $h(\lambda) = 0$ then we let $g(\lambda)$ to be any subgradient of $\|f(\lambda)\|_\infty$. Otherwise we let $g(\lambda)_{-\mathcal{B}} = 0$ and

$$g(\lambda)_{\mathcal{B}} = \frac{1}{h(\lambda)} \cdot D_{\mathcal{B}}\Big[I - (D_{-\mathcal{B}})^T \big(D_{-\mathcal{B}}(D_{-\mathcal{B}})^T\big)^+ D_{-\mathcal{B}}\Big]\big(y - \lambda(D_{\mathcal{B}})^T s\big).$$

Now we must check that the constraints are met with $\hat{u}_\lambda = f(\lambda)$, $\gamma = g(\lambda)$, and $\alpha = h(\lambda)$ as defined above. First we consider the case $\lambda = \lambda_k$:

- (GL-25a): This holds because $\|f(\lambda_k)_{\mathcal{B}}\|_\infty = \lambda_k$, and $\|f(\lambda_k)_{-\mathcal{B}}\|_\infty \leq \lambda_k$ by Lemma 3 (we delay presenting this lemma until Section 2.2.
- (GL-25b): This is true by construction.
- (GL-25c): This is true because $\|f(\lambda_k)\|_\infty = \lambda_k$ when $\mathcal{B} \neq \emptyset$, and otherwise $h(\lambda_k) = 0$.
- (GL-25d) and (GL-25e): Here we need to show that $g(\lambda_k)$ is indeed a subgradient of $\|f(\lambda_k)\|_\infty$. This is true by definition when $h(\lambda_k) = 0$, so suppose $h(\lambda_k) \neq 0$. Note first that $\|g(\lambda_k)\|_1 = 1$ by construction. Further, $\mathrm{sign}\big(g(\lambda_k)_{\mathcal{B}}\big) = \mathrm{sign}\big(f(\lambda_k)_{\mathcal{B}}\big)$ by Lemma 4 (presented in Section 2.2), and hence $g(\lambda_k)^T f(\lambda_k) = \|f(\lambda_k)\|_\infty$. This verifies the subgradient constraint.

As we decrease λ, note that only two of the above conditions can break: $\|f(\lambda)_{-\mathcal{B}}\|_\infty \leq \lambda$, or $\mathrm{sign}\big(g(\lambda)_{\mathcal{B}}\big) = \mathrm{sign}\big(f(\lambda)_{\mathcal{B}}\big)$. The first one will break when one of the interior coordinate paths crosses the boundary. This occurs at the next hitting time. Writing $f(\lambda)_{-\mathcal{B}} = a - \lambda b$ and solving $a_i - \lambda b_i = \pm\lambda$ for $i \notin \mathcal{B}$, we find that the hitting times are

$$t_i^{(\mathrm{hit})} = \frac{a_i}{b_i \pm 1} = \frac{\Big[\big(D_{-\mathcal{B}}(D_{-\mathcal{B}})^T\big)^+ D_{-\mathcal{B}} y\Big]_i}{\Big[\big(D_{-\mathcal{B}}(D_{-\mathcal{B}})^T\big)^+ D_{-\mathcal{B}}(D_{\mathcal{B}})^T s\Big]_i \pm 1},$$

where only one of $+1$ or -1 above will yield a value in $[0, \lambda_k]$. Thus the next hitting time is

$$h_{k+1} = \max_i t_i^{(\mathrm{hit})},$$

and the hitting coordinate and its sign are

$$i_{k+1}^{(\mathrm{hit})} = \operatorname*{argmax}_i t_i^{(\mathrm{hit})} \quad \text{and} \quad s_{k+1}^{(\mathrm{hit})} = \mathrm{sign}\Big(f(h_{k+1})_{i_{k+1}^{(\mathrm{hit})}}\Big).$$

The second condition, $\text{sign}\big(g(\lambda)_{\mathcal{B}}\big) = \text{sign}\big(f(\lambda)_{\mathcal{B}}\big)$, can be expressed as

$$s_i \cdot \Big[D_{\mathcal{B}} \Big[I - (D_{-\mathcal{B}})^T \big(D_{-\mathcal{B}} (D_{-\mathcal{B}})^T \big)^+ D_{-\mathcal{B}} \Big] \big(y - \lambda (D_{\mathcal{B}})^T s \big) \Big]_i \geq 0$$

for all $i \in \mathcal{B}$. Letting

$$c_i = s_i \cdot \Big[D_{\mathcal{B}} \Big[I - (D_{-\mathcal{B}})^T \big(D_{-\mathcal{B}} (D_{-\mathcal{B}})^T \big)^+ D_{-\mathcal{B}} \Big] y \Big]_i$$
$$d_i = s_i \cdot \Big[D_{\mathcal{B}} \Big[I - (D_{-\mathcal{B}})^T \big(D_{-\mathcal{B}} (D_{-\mathcal{B}})^T \big)^+ D_{-\mathcal{B}} \Big] (D_{\mathcal{B}})^T s \Big]_i ,$$

we can rewrite this as

$$c_i - \lambda d_i \geq 0 \tag{3}$$

for all $i \in \mathcal{B}$. Because we know that $c_i - \lambda_k d_i \geq 0$, the inequality (3) can only fail at some $\lambda \leq \lambda_k$ if c_i and d_i are both negative. Accordingly, the leaving times are

$$t_i^{(\text{leave})} = \begin{cases} c_i/d_i & \text{if } \; c_i < 0 \; \text{ and } \; d_i < 0 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore the next leaving time is

$$l_{k+1} = \max_i \, t_i^{(\text{leave})},$$

and the leaving coordinate and its sign are

$$i_{k+1}^{(\text{leave})} = \operatorname*{argmax}_i \, t_i^{(\text{leave})} \;\; \text{and} \;\; s_{k+1}^{(\text{leave})} = \text{sign}\Big(f(l_{k+1})_{i_{k+1}^{(\text{leave})}} \Big).$$

For the final step of the iteration, we take

$$\lambda_{k+1} = \max\{h_{k+1}, l_{k+1}\}.$$

This ensures the algorithm's correctness through the kth iteration, because we have satisfied the KKT conditions for all $\lambda \geq \lambda_{k+1}$. In preparation for the next iteration: if $h_{k+1} > l_{k+1}$ we add the hitting coordinate $i_{k+1}^{(\text{hit})}$ to $\mathcal{B}$ and append its sign $s_{k+1}^{(\text{hit})}$ to s; otherwise we delete the leaving coordinate $i_{k+1}^{(\text{leave})}$ from $\mathcal{B}$ and its sign $s_{k+1}^{(\text{leave})}$ from s.

2.2. *The insertion-deletion lemma.* The insertion-deletion lemma is important because it leads to Lemmas 3 and 4, which we used in the previous section to argue the correctness of our constructed path $\hat{u}_\lambda$. Moreover, it directly gives the continuity of $\hat{u}_\lambda$ with respect to λ.

While it may appear complicated, its concept is pretty simple: the insertion-deletion lemma states that the point $f(\lambda_{k+1})$ is the same with f as defined in iteration k or iteration $k+1$ (in other words, it is the same if we define f using the boundary set and signs from iteration k or iteration $k+1$). Note that iteration k could have ended in one of two ways: a coordinate was added to $\mathcal{B}$ (insertion), or a coordinate was removed from $\mathcal{B}$ (deletion). Therefore the lemma has two statements, corresponding to these two cases.

LEMMA 2 (**The insertion-deletion lemma**). *At the kth iteration of the algorithm, let $\mathcal{B}$ and s denote the boundary coordinates and their signs, and let $\mathcal{B}^*$ and s^* denote the same quantities at the beginning of the next iteration. The two possibilities are:*

1. *(Insertion) If a coordinate hit the boundary at λ_{k+1}, that is, $\mathcal{B}^*$ and s^* are given by adding elements to $\mathcal{B}$ and s, then:*

$$\begin{bmatrix} \big(f(\lambda_{k+1})_{-\mathcal{B}}\big)_{-i_{k+1}^{(\mathrm{hit})}} \\ f(\lambda_{k+1})_{i_{k+1}^{(\mathrm{hit})}} \end{bmatrix} = \begin{bmatrix} \big(D_{-\mathcal{B}^*}(D_{-\mathcal{B}^*})^T\big)^+ D_{-\mathcal{B}^*}\big(y - \lambda_{k+1}(D_{\mathcal{B}^*})^T s^*\big) \\ \lambda_{k+1} \cdot s_{k+1}^{(\mathrm{hit})} \end{bmatrix}. \tag{4}$$

2. *(Deletion) If a coordinate left the boundary at λ_{k+1}, that is, $\mathcal{B}^*$ and s^* are given by deleting elements from $\mathcal{B}$ and s, then:*

$$\begin{bmatrix} f(\lambda_{k+1})_{-\mathcal{B}} \\ \lambda_{k+1} \cdot s_{k+1}^{(\mathrm{leave})} \end{bmatrix} = \begin{bmatrix} \Big[\big(D_{-\mathcal{B}^*}(D_{-\mathcal{B}^*})^T\big)^+ D_{\mathcal{B}^*}\big(y - \lambda_{k+1}(D_{\mathcal{B}^*})^T s^*\big)\Big]_{-i_{k+1}^{(\mathrm{leave})}} \\ \Big[\big(D_{-\mathcal{B}^*}(D_{-\mathcal{B}^*})^T\big)^+ D_{\mathcal{B}^*}\big(y - \lambda_{k+1}(D_{\mathcal{B}^*})^T s^*\big)\Big]_{i_{k+1}^{(\mathrm{leave})}} \end{bmatrix}. \tag{5}$$

PROOF. The proof of each part relies on a block matrix decomposition. The arguments are not conceptually difficult but detailed. We treat the two cases separately.

Case 1: Insertion. Let

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \big(f(\lambda_{k+1})_{-\mathcal{B}}\big)_{-i_{k+1}^{(\mathrm{hit})}} \\ f(\lambda_{k+1})_{i_{k+1}^{(\mathrm{hit})}} \end{bmatrix},$$

the left-hand side of (4). By definition $i_{k+1}^{(\mathrm{hit})}$ hits the boundary at λ_{k+1}, so that exactly

$$x_2 = f(\lambda_{k+1})_{i_{k+1}^{(\mathrm{hit})}} = \lambda_{k+1} \cdot s_{k+1}^{(\mathrm{hit})}.$$

Now we consider x_1. Assume without a loss of generality that $i_{k+1}^{(\mathrm{hit})}$ is the last of the interior coordinates. Then we can write

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = f(\lambda_{k+1})_{-\mathcal{B}} = \big(D_{-\mathcal{B}}(D_{-\mathcal{B}})^T\big)^+ D_{-\mathcal{B}}\big(y - \lambda_{k+1}(D_{\mathcal{B}})^T s\big).$$

The point $(x_1, x_2)^T$ is the minimum ℓ_2 norm solution to the linear equation:

$$D_{-\mathcal{B}}(D_{-\mathcal{B}})^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = D_{-\mathcal{B}}\big(y - \lambda_{k+1}(D_{\mathcal{B}})^T s\big).$$

Decomposing this into blocks, we get

$$\begin{bmatrix} D_{-\mathcal{B}^*}(D_{-\mathcal{B}^*})^T & D_{-\mathcal{B}^*}\big(D_{i_{k+1}^{(\mathrm{hit})}}\big)^T \\ D_{i_{k+1}^{(\mathrm{hit})}}(D_{-\mathcal{B}^*})^T & D_{i_{k+1}^{(\mathrm{hit})}}\big(D_{i_{k+1}^{(\mathrm{hit})}}\big)^T \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} D_{-\mathcal{B}^*} \\ D_{i_{k+1}^{(\mathrm{hit})}} \end{bmatrix} \big(y - \lambda_{k+1}(D_{\mathcal{B}})^T s\big).$$

Solving for x_1 gives

$$\begin{aligned} x_1 &= \big(D_{-\mathcal{B}^*}(D_{-\mathcal{B}^*})^T\big)^+ D_{-\mathcal{B}^*}\Big[y - \lambda_{k+1}(D_{\mathcal{B}})^T s - \big(D_{i_{k+1}^{(\mathrm{hit})}}\big)^T x_2\Big] + b \\ &= \big(D_{-\mathcal{B}^*}(D_{-\mathcal{B}^*})^T\big)^+ D_{-\mathcal{B}^*}\big(y - \lambda_{k+1}(D_{\mathcal{B}^*})^T s^*\big) + b, \end{aligned}$$

where $b \in \mathrm{null}\big((D_{-\mathcal{B}^*})^T\big)$. The value of b can be determined by considering the squared norm of x_1,

$$\|x_1\|_2^2 = \Big\|\big(D_{-\mathcal{B}^*}(D_{-\mathcal{B}^*})^T\big)^+ D_{-\mathcal{B}^*}\big(y - \lambda_{k+1}(D_{\mathcal{B}^*})^T s^*\big)\Big\|_2^2 + \|b\|_2^2,$$

which is minimal when $b = 0$. This completes the proof.

Case 2: Deletion. This case is similar but a little more complicated. Let

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \Big[\big(D_{-\mathcal{B}^*}(D_{-\mathcal{B}^*})^T\big)^+ D_{\mathcal{B}^*}\big(y - \lambda_{k+1}(D_{\mathcal{B}^*})^T s^*\big)\Big]_{-i_{k+1}^{(\text{leave})}} \\ \Big[\big(D_{-\mathcal{B}^*}(D_{-\mathcal{B}^*})^T\big)^+ D_{\mathcal{B}^*}\big(y - \lambda_{k+1}(D_{\mathcal{B}^*})^T s^*\big)\Big]_{i_{k+1}^{(\text{leave})}} \end{bmatrix}.$$

If we assume without a loss of generality that $i_{k+1}^{(\text{leave})}$ is the largest of all the boundary coordinates, then $(x_1, x_2)^T$ is the minimum ℓ_2 norm solution of the equation:

$$\begin{bmatrix} D_{-\mathcal{B}}(D_{-\mathcal{B}})^T & D_{-\mathcal{B}}\big(D_{i_{k+1}^{(\text{leave})}}\big)^T \\ D_{i_{k+1}^{(\text{leave})}}(D_{-\mathcal{B}})^T & D_{i_{k+1}^{(\text{leave})}}\big(D_{i_{k+1}^{(\text{leave})}}\big)^T \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} D_{-\mathcal{B}} \\ D_{i_{k+1}^{(\text{leave})}} \end{bmatrix} \big(y - \lambda_{k+1}(D_{\mathcal{B}^*})^T s^*\big).$$

Solving this system for x_1 in terms of x_2 yields

$$x_1 = \big(D_{-\mathcal{B}}(D_{-\mathcal{B}})^T\big)^+ D_{-\mathcal{B}}\Big[y - \lambda_{k+1}(D_{\mathcal{B}^*})^T s^* - \big(D_{i_{k+1}^{(\text{leave})}}\big)^T x_2\Big] + b,$$

where $b \in \text{null}\big((D_{-\mathcal{B}})^T\big)$, and as we argued before, we must have $b = 0$ in order for x_1 to have minimal ℓ_2 norm. Therefore it suffices to show that $x_2 = \lambda_{k+1} \cdot s_{k+1}^{(\text{leave})}$. To this end, we continue the block elimination and solve for x_2. After a bit of algebra, this is

$$x_2 = \Big[D_{i_{k+1}^{(\text{leave})}} P\big(D_{i_{k+1}^{(\text{leave})}}\big)^T\Big]^{-1} D_{i_{k+1}^{(\text{leave})}} P\Big[y - \lambda_{k+1}(D_{\mathcal{B}})^T s + \lambda_{k+1}\big(D_{i_{k+1}^{(\text{leave})}}\big)^T s_{k+1}^{(\text{leave})}\Big], \tag{6}$$

where $P = P_{\text{null}(D_{-\mathcal{B}})}$. But by definition of $i_{k+1}^{(\text{leave})}$,

$$D_{i_{k+1}^{(\text{leave})}} P\big(y - \lambda_{k+1}(D_{\mathcal{B}})^T s\big) = 0.$$

Furthermore $D_{i_{k+1}^{(\text{leave})}} P\big(D_{i_{k+1}^{(\text{leave})}}\big)^T$ is just a scalar, and it is nonzero (otherwise this implies that $P\big(D_{i_{k+1}^{(\text{leave})}}\big)^T = 0$ as P is a projection matrix, and so $\lambda_{k+1} = 0$ by definition of the leaving time, which makes the result trivial). Therefore (6) becomes $x_2 = \lambda_{k+1} \cdot s_{k+1}^{(\text{leave})}$, which completes the proof. □

Now we give Lemmas 3 and 4, which we used in Section 2.1 to verify the KKT conditions.

LEMMA 3. *At the kth iteration of the algorithm,* $\|f(\lambda_k)_{-\mathcal{B}}\|_\infty \leq \lambda_k$.

PROOF. The proof is a straightforward application of induction and Lemma 2. At $k = 0$ this is trivially true since $\lambda_0 = \infty$. Now assume the statement holds for iteration k. Depending on whether a coordinate hit or left the boundary in iteration k, the statement for $k + 1$ is verified by taking the ℓ_∞ norm of the right-hand side of (4) or (5), respectively. □

LEMMA 4. *At the kth iteration of the algorithm,* $\text{sign}\big(g(\lambda_k)_{\mathcal{B}}\big) = \text{sign}\big(f(\lambda_k)_{\mathcal{B}}\big)$.

PROOF. Again we use induction. At $k = 0$ this is trivially true because $\mathcal{B} = \emptyset$. Suppose that the statement holds for all iterations $\leq k$. Given that we have already proved Lemma 3, the inductive hypothesis is really that the constructed path $\hat{u}_\lambda$ is the solution path for all $\lambda \geq \lambda_{k+1}$. Let $\mathcal{B}, s, f$, and g refer to the versions defined at the beginning of iteration $k+1$. By Lemma 2 we know that $\hat{u}_{\lambda_{k+1}} = f(\lambda_{k+1})$ is indeed the solution at λ_{k+1}. Hence $\hat{\beta}_{\lambda_{k+1}} = y - D^T f(\lambda_{k+1})$ is indeed the primal solution at λ_{k+1}. Noting that $g(\lambda_{k+1})_\mathcal{B} = D_\mathcal{B}\hat{\beta}_{\lambda_{k+1}}$, and recalling the relationship (GL-15), we have $\mathrm{sign}\big(g(\lambda_{k+1})_\mathcal{B}\big) = \mathrm{sign}\big(f(\lambda_{k+1})_\mathcal{B}\big)$. □

3. Proof of the primal-dual correspondence for a general D. Here we prove that the primal solution changes slope at λ_{k+1} if and only if the null space of $D_{-\mathcal{B}}$ changes from iterations k to $k+1$. Again we use the notation $\mathcal{B}, s$ and $\mathcal{B}^*, s^*$ to denote the boundary set and signs at iteration k, respectively $k+1$. This was claimed in Section GL-6.2 for the case $X = I$, and later in Section GL-7.1 for a general X with $\mathrm{rank}(X) = p$. We divide our proof into two parts, accordingly.

3.1. *The case $X = I$.* Consider the vector of coordinate-wise slopes of the solution path $\hat{\beta}_\lambda$, as a function of λ. Using (GL-33), the limits of this as $\lambda \to \lambda_{k+1}$ from above and below are

$$a_+ = P_{\mathrm{null}(D_{-\mathcal{B}})}(D_\mathcal{B})^T s \quad \text{and} \quad a_- = P_{\mathrm{null}(D_{-\mathcal{B}^*})}(D_{\mathcal{B}^*})^T s^*, \tag{7}$$

respectively. Suppose that a coordinate hit the boundary at λ_{k+1}. Then we have $(D_{\mathcal{B}^*})^T s^* = (D_\mathcal{B})^T s + \big(D_{i_{k+1}^{(\mathrm{hit})}}\big)^T s_{k+1}^{(\mathrm{hit})}$, and if $\mathrm{null}(D_{-\mathcal{B}}) = \mathrm{null}(D_{-\mathcal{B}^*})$ then

$$a_- = P_{\mathrm{null}(D_{-\mathcal{B}})}(D_\mathcal{B})^T s + P_{\mathrm{null}(D_{-\mathcal{B}})}\big(D_{i_{k+1}^{(\mathrm{hit})}}\big)^T s_{k+1}^{(\mathrm{hit})} = a_+,$$

where the identity $P_{\mathrm{null}(D_{-\mathcal{B}})}\big(D_{i_{k+1}^{(\mathrm{hit})}}\big)^T s_{k+1}^{(\mathrm{hit})} = 0$ follows from

$$\big(D_{i_{k+1}^{(\mathrm{hit})}}\big)^T s_{k+1}^{(\mathrm{hit})} \in \mathrm{row}(D_{-\mathcal{B}}) \perp \mathrm{null}(D_{-\mathcal{B}}).$$

A similar argument holds in the case that a coordinate left the boundary at λ_{k+1}. Therefore the slope of $\hat{\beta}_\lambda$ changes at λ_{k+1} only if $\mathrm{null}(D_{-\mathcal{B}}) \neq \mathrm{null}(D_{-\mathcal{B}^*})$.

The converse statement, that the slope of $\hat{\beta}_\lambda$ changes at λ_{k+1} if $\mathrm{null}(D_{-\mathcal{B}}) \neq \mathrm{null}(D_{-\mathcal{B}^*})$, is only true for (Lebesgue) almost every $y \in \mathbb{R}^n$. Hence, for any reasonable model of the data y, it holds with probability one. To show this, we first note that the limits of $\hat{\beta}_\lambda$ as $\lambda \to \lambda_{k+1}$ from above and below can be expressed as

$$\hat{\beta}_+ = P_{\mathrm{null}(D_{-\mathcal{B}})}(y) - \lambda_{k+1} a_+ \quad \text{and} \quad \hat{\beta}_- = P_{\mathrm{null}(D_{-\mathcal{B}^*})}(y) - \lambda_{k+1} a_-,$$

respectively, where a_-, a_+ are defined in (7). By the continuity of $\hat{\beta}_\lambda$, we know that $\hat{\beta}_+ = \hat{\beta}_-$. Suppose that $\mathrm{null}(D_{-\mathcal{B}}) \neq \mathrm{null}(D_{-\mathcal{B}^*})$. Then these two linear spaces differ in dimension by one (depending on whether or not a coordinate hit or left the boundary at λ_{k+1}). Hence $P_{\mathrm{null}(D_{-\mathcal{B}})}(y) \neq P_{\mathrm{null}(D_{-\mathcal{B}^*})}(y)$ for almost every $y \in \mathbb{R}^n$. Therefore, for any such y, we must have $a_+ \neq a_-$ in order to satisfy $\hat{\beta}_+ = \hat{\beta}_-$.

3.2. *The case of a general X, $\mathrm{rank}(X) = p$.* The proof is quite similar. Now the limiting slopes are, from equation (GL-38),

$$a_+ = X^+ P_{\mathrm{null}(\widetilde{D}_{-\mathcal{B}})}(\widetilde{D}_\mathcal{B})^T s \quad \text{and} \quad a_- = X^+ P_{\mathrm{null}(\widetilde{D}_{-\mathcal{B}^*})}(\widetilde{D}_{\mathcal{B}^*})^T s^*.$$

Recalling that $\widetilde{D}_{-\mathcal{B}} = D_{-\mathcal{B}}X^+$, we have

$$\text{null}(D_{-\mathcal{B}}) = \text{null}(D_{-\mathcal{B}^*}) \;\;\Rightarrow\;\; \text{null}(\widetilde{D}_{-\mathcal{B}}) = \text{null}(\widetilde{D}_{-\mathcal{B}^*}),$$

which implies that $a_- = a_+$, using the same arguments as we gave for the case $X = I$.

The converse is again true for almost every $y \in \mathbb{R}^n$. This is because

$$\text{null}(D_{-\mathcal{B}}) \neq \text{null}(D_{-\mathcal{B}^*}) \;\;\Rightarrow\;\; \text{null}(\widetilde{D}_{-\mathcal{B}}) \neq \text{null}(\widetilde{D}_{-\mathcal{B}^*}),$$

as X^+ has rank p, which implies that $a_+ \neq a_-$ for almost every y using similar arguments to those given above.

4. Proof of Lemma GL-3. Note that for a set $\mathcal{B} \subseteq \{1, \ldots m\}$, the matrix $D_{\mathcal{B}}P_{\text{null}(D_{-\mathcal{B}})}$ may have some rows that are entirely zero. We let $Z(\mathcal{B})$ denote the set of such rows. Now define

$$\mathcal{N}_\lambda \;=\; \bigcup_{\mathcal{B},s} \bigcup_{i \in \mathcal{B}\setminus Z(\mathcal{B})} \big\{x : D_i P_{\text{null}(D_{-\mathcal{B}})} x = \lambda D_i P_{\text{null}(D_{-\mathcal{B}})} (D_{\mathcal{B}})^T s\big\} \;\subseteq\; \mathbb{R}^n,$$

where the first union is taken over all subsets $\mathcal{B} \subseteq \{1, \ldots m\}$ and all sign vectors $s \in \{-1, 1\}^{|\mathcal{B}|}$. Note that $\mathcal{N}_\lambda$ is a finite union of affine subspaces of dimension $n-1$, and hence has measure zero. This establishes part (a) of the lemma.

Now, for $y \notin \mathcal{N}_\lambda$, let $\hat{u}_\lambda(y)$ be a dual solution with boundary set $\mathcal{B}$ and signs s. We show that:

1. there is a neighborhood U of y such that for any $y' \in U$, there exists a dual solution $\hat{u}_\lambda(y')$ with the same boundary set $\mathcal{B}$ and signs s;
2. if $u^*_\lambda(y)$ is another dual solution at y, with a different boundary set $\mathcal{B}^*$ and signs s^*, then

$$\lambda (D_{\mathcal{B}})^T s + \text{row}(D_{-\mathcal{B}}) = \lambda (D_{\mathcal{B}^*})^T s^* + \text{row}(D_{-\mathcal{B}^*}).$$

If we show these two statements then this would imply part (b) of the lemma.

4.1. *Proof of statement 1.* First note that we can rewrite the optimality conditions (GL-24) and (GL-25a)–(GL-25e) for our dual problem as

$$\|\hat{u}_\lambda\|_\infty \leq \lambda \tag{8}$$

$$D(y - D^T \hat{u}_\lambda) \in K, \tag{9}$$

where $K \subseteq \mathbb{R}^m$ is the cone generated by $\big\{\text{sign}(\hat{u}_{\lambda,i}) \cdot e_i : i \in \{j : |u_j| = \lambda\}\big\}$ (and e_i denotes the ith standard basis vector). Focusing first on the point y, let $a = D\big(y - D^T \hat{u}_\lambda(y)\big) \in K$, and note that K is generated by $\{s_i \cdot e_i : i \in \mathcal{B}\}$. Note also that $a_{-\mathcal{B}} = 0$, and using the fact that $\hat{u}_{\lambda,\mathcal{B}}(y) = \lambda \cdot s$, this means

$$D_{-\mathcal{B}}\big(y - \lambda (D_{\mathcal{B}})^T s\big) - D_{-\mathcal{B}}(D_{-\mathcal{B}})^T \hat{u}_{\lambda,-\mathcal{B}}(y) = 0.$$

Hence we can write the dual solution $\hat{u}_\lambda(y)$ as

$$\begin{aligned}
\hat{u}_{\lambda,\mathcal{B}}(y) &= \lambda \cdot s \\
\hat{u}_{\lambda,-\mathcal{B}}(y) &= \big(D_{-\mathcal{B}}(D_{-\mathcal{B}})^T\big)^+ D_{-\mathcal{B}}\big(y - \lambda (D_{\mathcal{B}})^T s\big) + b,
\end{aligned}$$

where $b \in \text{null}\big((D_{-\mathcal{B}})^T\big)$. By definition of the boundary set, $\|\hat{u}_{\lambda,-\mathcal{B}}(y)\|_\infty < \lambda$. Now we can also write

$$a_{\mathcal{B}} = D_{\mathcal{B}} P_{\text{null}(D_{-\mathcal{B}})}\big(y - \lambda (D_{\mathcal{B}})^T s\big).$$

Some rows of the matrix $D_{\mathcal{B}}P_{\mathrm{null}(D_{-\mathcal{B}})}$ may be entirely zero; recall that these are denoted by $Z(\mathcal{B})$. Since $y \notin \mathcal{N}_\lambda$, we know that $a_i \neq 0$ for all $i \in \mathcal{B} \setminus Z(\mathcal{B})$.

For a new point y', consider defining $\hat{u}_\lambda(y')$ as

$$\begin{aligned}\hat{u}_{\lambda,\mathcal{B}}(y') &= \lambda \cdot s \\ \hat{u}_{\lambda,-\mathcal{B}}(y') &= \big(D_{-\mathcal{B}}(D_{-\mathcal{B}})^T\big)^+ D_{-\mathcal{B}}\big(y' - \lambda(D_{\mathcal{B}})^T s\big) + b.\end{aligned}$$

By continuity of the affine mapping (note that $\mathcal{B}, s, b$ are fixed)

$$x \mapsto \big(D_{-\mathcal{B}}(D_{-\mathcal{B}})^T\big)^+ D_{-\mathcal{B}}\big(x - \lambda(D_{\mathcal{B}})^T s\big) + b,$$

there exists a neighborhood U_1 of y such that $\|\hat{u}_{\lambda,-\mathcal{B}}(y')\|_\infty < \lambda$ for all y' in U_1. This establishes the first optimality condition (8) and shows that $\hat{u}_\lambda(y')$ has boundary set $\mathcal{B}$ and signs s, for all $y' \in U_1$. To establish the second optimality condition (9), we must check that

$$a' = DP_{\mathrm{null}(D_{-\mathcal{B}})}\big(y' - \lambda(D_{\mathcal{B}})^T s\big) \in K.$$

Well $a'_{-\mathcal{B}} = 0$ and also $a'_{Z(\mathcal{B})} = 0$. By continuity of the affine mapping (again $\mathcal{B}, s$ are fixed)

$$x \mapsto D_{\mathcal{B}\setminus Z(\mathcal{B})}P_{\mathrm{null}(D_{-\mathcal{B}})}\big(x - \lambda(D_{\mathcal{B}})^T s\big),$$

there exists another neighborhood U_2 of y such that $a'_i \neq 0$ and further $\mathrm{sign}(a'_i) = \mathrm{sign}(a_i)$ for all $i \in \mathcal{B} \setminus Z(\mathcal{B})$ and $y' \in U_2$. As $a \in K$, this means that $a' \in K$ for all $y' \in U_2$. Letting $U = U_1 \cap U_2$, we have verified that $\hat{u}_\lambda(y')$ has boundary set $\mathcal{B}$ and signs s, and is indeed a dual solution, for all $y' \in U$.

4.2. *Proof of statement 2.* Given another dual solution $u^*_\lambda(y)$ at y with a different boundary set $\mathcal{B}^*$ and signs s^*, we know from statement 1 that there is a neighborhood U^* of y such that

$$\begin{aligned}u^*_{\lambda,\mathcal{B}^*}(y') &= \lambda \cdot s^* \\ u^*_{\lambda,-\mathcal{B}^*}(y') &= \big(D_{-\mathcal{B}^*}(D_{-\mathcal{B}^*})^T\big)^+ D_{-\mathcal{B}^*}\big(y' - \lambda(D_{\mathcal{B}^*})^T s^*\big) + b^*\end{aligned}$$

is a dual solution for all $y' \in U^*$, where $b^* \in \mathrm{null}\big((D_{-\mathcal{B}^*})^T\big)$. By uniqueness of the dual fit, we have

$$\lambda(D_{\mathcal{B}})^T s + P_{\mathrm{row}(D_{-\mathcal{B}})}\big(y' - \lambda(D_{\mathcal{B}})^T s\big) \;=\; \lambda(D_{\mathcal{B}^*})^T s^* + P_{\mathrm{row}(D_{-\mathcal{B}^*})}\big(y' - \lambda(D_{\mathcal{B}^*})^T s^*\big)$$

for all $y' \in U \cap U^* \neq \emptyset$, and therefore

$$\lambda(D_{\mathcal{B}})^T s + \mathrm{row}(D_{-\mathcal{B}}) = \lambda(D_{\mathcal{B}^*})^T s^* + \mathrm{row}(D_{-\mathcal{B}^*}).$$

References.

[1] Tseng, P. [2001], 'Convergence of a block coordinate descent method for nondifferentiable minimization', *Journal of Optimization Theory and Applications* **109**(3), 475–494.