Supplementary online material for "An ordinary differential equation based solution path algorithm" by Yichao Wu

Proof of Lemma 1. The new path updating direction is given by $-\boldsymbol{M}_{\mathcal{A}^*,\mathcal{A}^*}(\boldsymbol{\beta}(t^*))^{-1}\text{sign}(\boldsymbol{b}_{\mathcal{A}^*}(\boldsymbol{\beta}(t^*)))$. To facilitate our proof, we reshuffle the order and rewrite it as

$$-\begin{pmatrix} \boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\boldsymbol{\beta}(t^*)) & \boldsymbol{M}_{\mathcal{A},j^*}(\boldsymbol{\beta}(t^*)) \\ \boldsymbol{M}_{j^*,\mathcal{A}}(\boldsymbol{\beta}(t^*)) & \boldsymbol{M}_{j^*,j^*}(\boldsymbol{\beta}(t^*)) \end{pmatrix}^{-1} \begin{pmatrix} \text{sign}(\boldsymbol{b}_{\mathcal{A}}(\boldsymbol{\beta}(t^*))) \\ \text{sign}(b_{j^*}(\boldsymbol{\beta}(t^*))) \end{pmatrix}. \tag{14}$$

The last element of (14) is given by

$$\frac{1}{\gamma}\left[\boldsymbol{M}_{j^*,\mathcal{A}}(\boldsymbol{\beta}(t^*))\boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\boldsymbol{\beta}(t^*))^{-1}\text{sign}(\boldsymbol{b}_{\mathcal{A}}(\boldsymbol{\beta}(t^*))) - \text{sign}(b_{j^*}(\boldsymbol{\beta}(t^*)))\right], \tag{15}$$

where $\gamma = \boldsymbol{M}_{j^*,j^*}(\boldsymbol{\beta}(t^*)) - \boldsymbol{M}_{j^*,\mathcal{A}}(\boldsymbol{\beta}(t^*))\boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\boldsymbol{\beta}(t^*))^{-1}\boldsymbol{M}_{\mathcal{A},j^*}(\boldsymbol{\beta}(t^*)) < 0$ in that $\boldsymbol{M}(\boldsymbol{\beta})$ is negative definite when $n > p$ and $\boldsymbol{x}^{(j)}$, $j = 1, \cdots, p$ are linearly independent. The first term in (15) involves $\boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\boldsymbol{\beta}(t^*))^{-1}\text{sign}(\boldsymbol{b}_{\mathcal{A}}(\boldsymbol{\beta}(t^*)))$ which is exactly the opposite of the path updating direction calculated at t^* using the old active set $\mathcal{A}$ by ignoring the addition of predictor variable j^*.

Consider ignoring the addition of the new active variable j^* and updating path along the path updating direction evaluated by the old active predictor set $\mathcal{A}$. This leads to another solution path piece $\bar{\boldsymbol{\beta}}(t)$ defined by $\bar{\boldsymbol{\beta}}_{\mathcal{A}}(t) = \boldsymbol{\beta}_{\mathcal{A}}(t^*) - \int_{t^*}^{t} \boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\bar{\boldsymbol{\beta}}(\tau))^{-1}\text{sign}(\boldsymbol{b}_{\mathcal{A}}(\bar{\boldsymbol{\beta}}(\tau)))d\tau$ and $\bar{\boldsymbol{\beta}}_{\mathcal{A}^c}(t) = \boldsymbol{0}$ when t is inside a small neighborhood $[t^* - \triangle_t, t^* + \triangle_t]$ for some $\triangle_t > 0$. The neighborhood is chosen such that both solution component $\bar{\beta}_j(t)$ and the first-order partial derivative $b_j(\bar{\boldsymbol{\beta}}(t))$ do not change sign for $t \in [t^* - \triangle_t, t^* + \triangle_t]$ and $j \in \mathcal{A}$. Consequently when $t \in [t^* - \triangle_t, t^* + \triangle_t]$, $\frac{d}{dt}b_j(\bar{\boldsymbol{\beta}}(t)) = -\text{sign}(b_j(\bar{\boldsymbol{\beta}}(t))) = -\text{sign}(b_j(\bar{\boldsymbol{\beta}}(t^*)))$ for

$j \in \mathcal{A}$ due to (7). Note that the definition of $\bar{\boldsymbol{\beta}}(t)$ implies

$$\frac{d}{dt} b_{j^*}(\bar{\boldsymbol{\beta}}(t)) = \sum_{j=1}^{p} m_{j^*j}(\bar{\boldsymbol{\beta}}(t)) \frac{d}{dt} \bar{\beta}_j(t) = -\boldsymbol{M}_{j^*,\mathcal{A}}(\bar{\boldsymbol{\beta}}(t)) \boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\bar{\boldsymbol{\beta}}(t))^{-1} \mathrm{sign}(\boldsymbol{b}_{\mathcal{A}}(\bar{\boldsymbol{\beta}}(t))). \quad (16)$$

Recall that for $t \in [t^* - \triangle_t, t^*]$, $\boldsymbol{\beta}(t) = \bar{\boldsymbol{\beta}}(t)$ and our QuasiLARS solution matches $\bar{\boldsymbol{\beta}}(t)$ exactly. Our QuasiLARS definition implies that $|b_{j^*}(\bar{\boldsymbol{\beta}}(t))| < |b_j(\bar{\boldsymbol{\beta}}(t))|$ for any $j \in \mathcal{A}$ and $t \in [t^* - \triangle_t, t^*)$. That means that predictor variable j^* has a smaller absolute value of the first-order partial derivative than active predictors in $\mathcal{A}$ for $t \in [t^* - \triangle_t, t^*)$ and catches up with active predictors in $\mathcal{A}$ at t^* by noting that predictor variable j^* joins the active predictor set $\mathcal{A}$ at t^*.

Next we prove our claim by contradiction. If our claim is wrong, then we have $[\boldsymbol{M}_{j^*,\mathcal{A}}(\boldsymbol{\beta}(t^*)) \boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\boldsymbol{\beta}(t^*))^{-1} \mathrm{sign}(\boldsymbol{b}_{\mathcal{A}}(\boldsymbol{\beta}(t^*))) - \mathrm{sign}(b_{j^*}(\boldsymbol{\beta}(t^*)))]\, \mathrm{sign}(b_{j^*}(\boldsymbol{\beta}(t^*))) > 0$ due to (15) and $\gamma < 0$. It implies $\boldsymbol{M}_{j^*,\mathcal{A}}(\boldsymbol{\beta}(t^*)) \boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\boldsymbol{\beta}(t^*))^{-1} \mathrm{sign}(\boldsymbol{b}_{\mathcal{A}}(\boldsymbol{\beta}(t^*))) \mathrm{sign}(b_{j^*}(\boldsymbol{\beta}(t^*))) > 1$. Note that $\bar{\boldsymbol{\beta}}(t) = \boldsymbol{\beta}(t)$ for $t \in [t^* - \triangle_t, t^*]$ implies the existence of $\epsilon \in (0, \triangle_t)$ such that

$$\boldsymbol{M}_{j^*,\mathcal{A}}(\bar{\boldsymbol{\beta}}(t)) \boldsymbol{M}_{\mathcal{A},\mathcal{A}}(\bar{\boldsymbol{\beta}}(t))^{-1} \mathrm{sign}(\boldsymbol{b}_{\mathcal{A}}(\bar{\boldsymbol{\beta}}(t))) \mathrm{sign}(b_{j^*}(\bar{\boldsymbol{\beta}}(t))) > 1 \text{ for } t \in (t^* - \epsilon, t^*) \quad (17)$$

due to continuity. By noting (16) and $\frac{d}{dt} b_j(\bar{\boldsymbol{\beta}}(t)) = -\mathrm{sign}(b_j(\bar{\boldsymbol{\beta}}(t)))$ for $j \in \mathcal{A}$ and $t \in (t^* - \epsilon, t^*)$, (17) contradicts the above conclusion that predictor j^* has a smaller absolute value of the first-order partial derivative than active predictors in $\mathcal{A}$ for $t \in [t^* - \triangle_t, t^*)$ and catches up with active predictors in $\mathcal{A}$ at t^*. This completes our proof. □

Proof of Lemma 2. For any $j \in \mathcal{N}(\hat{\boldsymbol{\beta}})$, differentiating the objective function in (8) with respect β_j, we get

$$-\frac{\partial}{\partial \beta_j} R(\boldsymbol{\beta}, \beta_0(\boldsymbol{\beta})) + \lambda \mathrm{sign}(\beta_j) \quad (18)$$

which has to be equal to zero at $\hat{\boldsymbol{\beta}}$ in that $\hat{\boldsymbol{\beta}}$ solves (8). This completes the proof by noting that $\lambda \geq 0$ and, when $\lambda = 0$, $\frac{\partial}{\partial \beta_j} R(\boldsymbol{\beta}, \beta_0(\boldsymbol{\beta})) = 0$ for all j. □

Proof of Lemma 3. Note that $\hat{\boldsymbol{\beta}}(s)$ solves (9) and has nonzero set $\mathcal{N}_s$, which is constant for $s \in \mathcal{S}$. Namely $\mathcal{N}_s = \mathcal{N}$ for some $\mathcal{N}$ and all $s \in \mathcal{S}$. Then $\hat{\boldsymbol{\beta}}_{\mathcal{N}}(s)$ minimizes $-R(\hat{\boldsymbol{\beta}}_{\mathcal{N}}, \beta_0(\hat{\boldsymbol{\beta}}_{\mathcal{N}})) \triangleq -\sum_{i=1}^n Q(g^{-1}(\beta_0 + \boldsymbol{x}_{i\mathcal{N}}^T \hat{\boldsymbol{\beta}}_{\mathcal{N}}), y_i)$, subject to

$$\boldsymbol{s}_{\mathcal{N}}^T \hat{\boldsymbol{\beta}}_{\mathcal{N}} = s \text{ and } \operatorname{sign}(\hat{\beta}_j) = s_j \text{ for } j \in \mathcal{N}, \tag{19}$$

where $s_j = \operatorname{sign}(b_j(\hat{\boldsymbol{\beta}}(s))$ for $j = 1, \cdots, p$ denote the sign of the current first-order partial derivatives and $\boldsymbol{s} = (s_1, \cdots, s_p)^T$ and the sign restriction is due to Lemma 2. Here $\boldsymbol{x}_{i\mathcal{N}}$ is the sub-vector of $\boldsymbol{x}_i$ with index in $\mathcal{N}$ and, for any $\hat{\boldsymbol{\beta}}_{\mathcal{N}}$, $\beta_0(\hat{\boldsymbol{\beta}}_{\mathcal{N}})$ is defined by $\beta_0(\hat{\boldsymbol{\beta}}_{\mathcal{N}}) = \text{argmax}_{\beta_0} \sum_{i=1}^n Q(g^{-1}(\beta_0 + \boldsymbol{x}_{i\mathcal{N}}^T \hat{\boldsymbol{\beta}}_{\mathcal{N}}), y_i)$. Note that the inequality constraint in (9) can be replaced by the equality constraint $\sum_{j=1}^p |\beta_j| = s$ as long as s is less than the one norm of the full quasi-likelihood solution to (1). This justifies (19). Note further that the optimal solution $\hat{\boldsymbol{\beta}}_{\mathcal{N}}(s)$ is strictly inside the simplex (19) since $\hat{\beta}_j(s) \neq 0$ for $j \in \mathcal{N}$ and $s \in \mathcal{S}$. This in combination with the strict convexity of the objective function $-R(\hat{\boldsymbol{\beta}}_{\mathcal{N}}, \beta_0(\hat{\boldsymbol{\beta}}_{\mathcal{N}}))$ implies that the second condition, $\operatorname{sign}(\hat{\beta}_j) = s_j$ for $j \in \mathcal{N}$, can be dropped. Consequently $\hat{\boldsymbol{\beta}}_{\mathcal{N}}(s)$ solves $\min -R(\hat{\boldsymbol{\beta}}_{\mathcal{N}}, \beta_0(\hat{\boldsymbol{\beta}}_{\mathcal{N}}))$ subject to $\boldsymbol{s}_{\mathcal{N}}^T \hat{\boldsymbol{\beta}}_{\mathcal{N}} = s$. By introducing a Lagrange multiplier λ, it becomes $\min -R(\hat{\boldsymbol{\beta}}_{\mathcal{N}}, \beta_0(\hat{\boldsymbol{\beta}}_{\mathcal{N}})) + \lambda \sum_{j \in \mathcal{N}} s_j \hat{\beta}_j$. Applying differential operator $\frac{\partial}{\partial \hat{\boldsymbol{\beta}}_{\mathcal{N}}}$, we get

$$-\frac{\partial}{\partial \hat{\boldsymbol{\beta}}_{\mathcal{N}}} R(\hat{\boldsymbol{\beta}}_{\mathcal{N}}, \beta_0(\hat{\boldsymbol{\beta}}_{\mathcal{N}})) + \lambda \boldsymbol{s}_{\mathcal{N}}, \tag{20}$$

which is equal to $\mathbf{0}$ at $\hat{\boldsymbol{\beta}}_{\mathcal{N}} = \hat{\boldsymbol{\beta}}_{\mathcal{N}}(s)$ because $\hat{\boldsymbol{\beta}}_{\mathcal{N}}(s)$ is the corresponding optimal solution.

Now consider two different values $s^{(1)}$ and $s^{(2)}$ in $\mathcal{S}$ with $\underline{s} < s^{(1)} < s^{(2)}$. The corresponding Lagrange multiplies are denoted by $\lambda^{(1)}$ and $\lambda^{(2)}$ and they satisfy $\lambda^{(1)} > \lambda^{(2)}$. Putting them into (20) and differencing, we get

$$-\left(\frac{\partial}{\partial \hat{\boldsymbol{\beta}}_{\mathcal{N}}} R(\hat{\boldsymbol{\beta}}_{\mathcal{N}}, \beta_0(\hat{\boldsymbol{\beta}}_{\mathcal{N}}))|_{\hat{\boldsymbol{\beta}}_{\mathcal{N}}(s^{(2)})} - \frac{\partial}{\partial \hat{\boldsymbol{\beta}}_{\mathcal{N}}} R(\hat{\boldsymbol{\beta}}_{\mathcal{N}}, \beta_0(\hat{\boldsymbol{\beta}}_{\mathcal{N}}))|_{\hat{\boldsymbol{\beta}}_{\mathcal{N}}(s^{(1)})} \right) = (\lambda^{(1)} - \lambda^{(2)}) \boldsymbol{s}_{\mathcal{N}}. \tag{21}$$

Note that $\hat{\boldsymbol{\beta}}_{\mathcal{N}^c}(s) = \mathbf{0}$ for any $s \in \mathcal{S}$. Thus (21) is the same as

$$-\Big(\boldsymbol{b}_{\mathcal{N}}(\hat{\boldsymbol{\beta}}(s^{(2)})) - \boldsymbol{b}_{\mathcal{N}}(\hat{\boldsymbol{\beta}}(s^{(1)}))\Big) = (\lambda^{(1)} - \lambda^{(2)})\boldsymbol{s}_{\mathcal{N}}. \tag{22}$$

Dividing both sides of (22) by $s^{(2)} - s^{(1)}$ and letting $s^{(2)} \to s^{(1)}$, we get

$$-\frac{d}{ds}\boldsymbol{b}_{\mathcal{N}}(\hat{\boldsymbol{\beta}}(s))|_{s^{(1)}} = -\lambda'(s^{(1)})\boldsymbol{s}_{\mathcal{N}}, \tag{23}$$

where $\lambda'(s) = \frac{d}{ds}\lambda(s)$ is negative. Noting that $\frac{d}{ds}\boldsymbol{b}(\hat{\boldsymbol{\beta}}(s)) = \boldsymbol{M}(\hat{\boldsymbol{\beta}}(s))\frac{d}{ds}\hat{\boldsymbol{\beta}}(s)$, $\hat{\boldsymbol{\beta}}_{\mathcal{N}^c}(s) = \mathbf{0}$ for $s \in \mathcal{S}$, (23) becomes $-\boldsymbol{M}_{\mathcal{N},\mathcal{N}}(\hat{\boldsymbol{\beta}}(s^{(1)}))\frac{d}{ds}\hat{\boldsymbol{\beta}}_{\mathcal{N}}(s)|_{s^{(1)}} = -\lambda'(s^{(1)})\boldsymbol{s}_{\mathcal{N}}$, which leads to $\frac{d}{ds}\hat{\boldsymbol{\beta}}_{\mathcal{N}}(s)|_{s^{(1)}} = \lambda'(s^{(1)})(\boldsymbol{M}_{\mathcal{N},\mathcal{N}}(\hat{\boldsymbol{\beta}}(s^{(1)})))^{-1}\boldsymbol{s}_{\mathcal{N}}$. By noting that $\lambda'(s) < 0$, this shows that for any $s \in \mathcal{S}$, the LASSO regularized quasi-likelihood solution updating direction matches our QuasiLARS path updating direction. It also holds for $\underline{s}$ due to continuity. □

Proof of Lemma 4. Due to (18), $|b_j(\hat{\boldsymbol{\beta}}(s))| = |b_{j'}(\hat{\boldsymbol{\beta}}(s))|$ for any $j,\ j' \in \mathcal{N}$. Thus it is enough to prove that $|b_l(\hat{\boldsymbol{\beta}}(s))| \leq |b_j(\hat{\boldsymbol{\beta}}(s))|$ for any $l \notin \mathcal{N}$, $j \in \mathcal{N}$, $s \in \mathcal{S} \cup \{\underline{s}\}$.

We first prove the statement for $s \in \mathcal{S}$ by contradiction. Suppose there is some $j^* \notin \mathcal{N}$ and some $s^* \in \mathcal{S}$ such that

$$|b_{j^*}(\hat{\boldsymbol{\beta}}(s^*))| > |b_j(\hat{\boldsymbol{\beta}}(s^*))|. \tag{24}$$

Let $\boldsymbol{d} = (d_1, \cdots, d_p)^T$ with $d_j = -\text{sign}(\hat{\beta}_j(s^*))(= -\text{sign}(b_j(\hat{\boldsymbol{\beta}}(s^*)))$, due to Lemma 2) for $j \in \mathcal{N}$, $d_{j^*} = n_{\mathcal{N}}\text{sign}(b_{j^*}(\hat{\boldsymbol{\beta}}(s^*)))$, and $d_{j'} = 0$ for $j \in (\mathcal{N} \cup \{j^*\})^c$, where $n_{\mathcal{N}}$ denote the size of $\mathcal{N}$.

Consider $R(\hat{\boldsymbol{\beta}}(s^*) + u\boldsymbol{d}, \boldsymbol{\beta}_0(\hat{\boldsymbol{\beta}}(s^*) + u\boldsymbol{d}))$ as a function of u. Its derivative is given by

$$\frac{d}{du}R(\hat{\boldsymbol{\beta}}(s^*) + u\boldsymbol{d}, \boldsymbol{\beta}_0(\hat{\boldsymbol{\beta}}(s^*) + u\boldsymbol{d})) \ = \ \sum_{j=1}^{p} b_j(\hat{\boldsymbol{\beta}}(s^*) + u\boldsymbol{d})d_j + O(u). \tag{25}$$

When $u = 0$, the right hand side of (25) becomes

$$-n_{\mathcal{N}}|b_j(\hat{\boldsymbol{\beta}}(s^*))| + n_{\mathcal{N}}|b_{j^*}(\hat{\boldsymbol{\beta}}(s^*))| > 0, \tag{26}$$

where $j \in \mathcal{N}$ and positivity is due to (24). Note that $\min_{j\in\mathcal{N}} |\hat{\beta}_j(s^*)| > 0$ since $s^* \in \mathcal{S}$. When $0 < u < \min_{j\in\mathcal{N}} |\hat{\beta}_j(s^*)|$, $\sum_{j=1}^{p} |\hat{\beta}_j(s^*)| = \sum_{j=1}^{p} |\hat{\beta}_j(s^*) + ud_j|$ by noting the above definition of $\boldsymbol{d}$. However (26) contradicts the fact that $\hat{\boldsymbol{\beta}}(s^*)$ is a solution of the LASSO regularized quasi-likelihood (18). This proves our claim for $s \in \mathcal{S}$. Our claim holds at $\underline{s}$ simply due to continuity. □

Proof of Lemma 5. Note that $\dot{S}(0) = \sum_{j\in\mathcal{A}_1} \mathrm{sign}(b_j(\hat{\boldsymbol{\beta}}))d_j + \sum_{j\in\mathcal{A}_0} |d_j| + \sum_{j\in\mathcal{A}_{10}^c} |d_j|$ due to Lemma 2 and $\dot{T}(0) = \sum_{j\in\mathcal{A}_1} b_j(\hat{\boldsymbol{\beta}})d_j + \sum_{j\in\mathcal{A}_0} b_j(\hat{\boldsymbol{\beta}})d_j + \sum_{j\in\mathcal{A}_{10}^c} b_j(\hat{\boldsymbol{\beta}})d_j$. Thus due to Lemma 4 and the above definition of $\mathcal{A}_0$, we have

$$\dot{T}(0)/\dot{S}(0) = \hat{D}(\hat{\boldsymbol{\beta}}) \frac{\sum_{j\in\mathcal{A}_1} \mathrm{sign}(b_j(\hat{\boldsymbol{\beta}}))d_j + \sum_{j\in\mathcal{A}_0} \mathrm{sign}(b_j(\hat{\boldsymbol{\beta}}))d_j + \sum_{j\in\mathcal{A}_{10}^c} d_j b_j(\hat{\boldsymbol{\beta}})/\hat{D}(\hat{\boldsymbol{\beta}})}{\sum_{j\in\mathcal{A}_1} \mathrm{sign}(b_j(\hat{\boldsymbol{\beta}}))d_j + \sum_{j\in\mathcal{A}_0} |d_j| + \sum_{j\in\mathcal{A}_{10}^c} |d_j|},$$

which is analogous to Equation (5.40) of Efron et al. (2004). It is enough to consider $\boldsymbol{d}$ satisfying $\sum_{j\in\mathcal{A}_1} \mathrm{sign}(b_j(\hat{\boldsymbol{\beta}}))d_j + \sum_{j\in\mathcal{A}_0} |d_j| + \sum_{j\in\mathcal{A}_{10}^c} |d_j| > 0$ which corresponds to $\dot{S}(0) > 0$, for which case $S(\gamma)$ is increasing in γ at the origin. Thus we need $d_j \mathrm{sign}(b_j(\hat{\boldsymbol{\beta}})) \geq 0$ for $j \in \mathcal{A}_0 \cup (\mathcal{A}_{10}^c)$ in order to maximize $Z(\boldsymbol{d})$. In this case we have

$$Z(\boldsymbol{d}) = \hat{D}(\hat{\boldsymbol{\beta}}) \frac{\sum_{j\in\mathcal{A}_1} \mathrm{sign}(b_j(\hat{\boldsymbol{\beta}}))d_j + \sum_{j\in\mathcal{A}_0} |d_j| + \sum_{j\in\mathcal{A}_{10}^c} |d_j| \frac{|b_j(\hat{\boldsymbol{\beta}})|}{\hat{D}(\hat{\boldsymbol{\beta}})}}{\sum_{j\in\mathcal{A}_1} \mathrm{sign}(b_j(\hat{\boldsymbol{\beta}}))d_j + \sum_{j\in\mathcal{A}_0} |d_j| + \sum_{j\in\mathcal{A}_{10}^c} |d_j|},$$

which is $< \hat{D}(\hat{\boldsymbol{\beta}})$ unless $d_j = 0$ for $j \in \mathcal{A}_{10}^c$ since $|b_j(\hat{\boldsymbol{\beta}})| < \hat{D}(\hat{\boldsymbol{\beta}})$ for $j \in \mathcal{A}_{10}^c$. This proves (11). In this case a second order Taylor expansion leads to (12). □

Proof of Lemma 6. By noting that $\boldsymbol{M}_{\mathcal{A}_{10},\mathcal{A}_{10}}(\hat{\boldsymbol{\beta}})$ is negative definite, (13) is equivalent to

$$\begin{aligned} \max \quad & \boldsymbol{d}_{\mathcal{A}_{10}^T}^T \boldsymbol{M}_{\mathcal{A}_{10},\mathcal{A}_{10}}(\hat{\boldsymbol{\beta}}) \boldsymbol{d}_{\mathcal{A}_{10}} \\ \text{subject to} \quad & \sum_{j\in\mathcal{A}_1} \mathrm{sign}(b_j(\hat{\boldsymbol{\beta}}))d_j + \sum_{j\in\mathcal{A}_0} |d_j| \geq \triangle \text{ and } \mathrm{sign}(d_j) = \mathrm{sign}(b_j(\hat{\boldsymbol{\beta}})) \text{ for } j \in \mathcal{A}_0. \end{aligned} \tag{27}$$

Consider combining the two constraints in (27) into one and solve a simpler version

$$\max \boldsymbol{d}_{\mathcal{A}_{10}^T}^T \boldsymbol{M}_{\mathcal{A}_{10},\mathcal{A}_{10}}(\hat{\boldsymbol{\beta}})\boldsymbol{d}_{\mathcal{A}_{10}} \text{ subject to } \sum_{j\in\mathcal{A}_1} \text{sign}(b_j(\hat{\boldsymbol{\beta}}))d_j + \sum_{j\in\mathcal{A}_0} \text{sign}(b_j(\hat{\boldsymbol{\beta}}))d_j \geq \triangle. \quad (28)$$

Later we will show that the solution to (28) satisfies the sign constraint in (27).

We solve $\max \boldsymbol{d}_{\mathcal{A}_{10}^T}^T \boldsymbol{M}_{\mathcal{A}_{10},\mathcal{A}_{10}}(\hat{\boldsymbol{\beta}})\boldsymbol{d}_{\mathcal{A}_{10}} + \lambda \left(\sum_{j\in\mathcal{A}_1} \text{sign}(b_j(\hat{\boldsymbol{\beta}}))d_j + \sum_{j\in\mathcal{A}_0} \text{sign}(b_j(\hat{\boldsymbol{\beta}}))d_j - \triangle\right)$ by introducing a Lagrange multiplier λ. Differentiating with respect to $\boldsymbol{d}_{\mathcal{A}_{10}}$ and solving for $\boldsymbol{d}_{\mathcal{A}_{10}}$, we get the optimal solution $\boldsymbol{d}_{\mathcal{A}_{10}}^{opt} = -\lambda(\boldsymbol{M}_{\mathcal{A}_{10},\mathcal{A}_{10}})^{-1}\text{sign}(\boldsymbol{b}_{\mathcal{A}_{10}}(\hat{\boldsymbol{\beta}}))$, which corresponds exactly to our QuasiLARS updating direction by noting $\lambda > 0$.

Note that we assume "one at a time" condition. Thus $\mathcal{A}_0$ is a singleton. Consequently the second constraint in (27) is satisfied due to Lemma 1. This completes the proof. □

Proof of Theorem 1. Lemmas 2-5 are extensions of Lemmas 7-10 of Efron et al. (2004), which are key results for establishing that the LASSO modification leads to the LASSO solutions. Their proof by induction can be extended to prove our Theorem based on Lemmas 1-5 and parallel extension of Constraints 1-4 on page 437 of Efron et al. (2004). We skip the details here to save space. □