A Proof of Theorem 2

Proof. By standard conditions for optimality, $\hat{\beta}$ is a critical point if and only if there exits a subgradient $\hat{z} \in \partial\|\hat{\beta}\|_1 := \{\hat{z} \in \mathbb{R}^p | \hat{z}_j = \mathrm{sgn}(\hat{\beta}_j) \text{ for } \hat{\beta}_j \neq 0,\ |\hat{z}_j| \leq 1 \text{ otherwise}\}$ such that $\partial_{\hat{\beta}} L(\beta) = 0$. Because $\partial_\beta \frac{1}{2}|\beta|^\top R|\beta| = \mathrm{Diag}(R|\beta|)z$, the condition $\partial_{\hat{\beta}} L(\beta) = 0$ yields

$$-\frac{1}{n}X^\top(y - X\hat{\beta}) + \lambda\hat{z} + \lambda\alpha\mathrm{Diag}\left(R|\hat{\beta}|\right)\hat{z} = 0. \tag{A.1}$$

Substituting $y = X\beta^* + \epsilon$ in (A.1), we have

$$-\frac{1}{n}X^\top(X(\beta^* - \hat{\beta}) + \epsilon) + \lambda\hat{z} + \lambda\alpha\mathrm{Diag}\left(R|\hat{\beta}|\right)\hat{z} = 0. \tag{A.2}$$

Let the true active set $S = \{1, \cdots, s\}$ and inactive set $S^c = \{s+1, \cdots, p\}$ without loss of generality, then (A.2) is turned into

$$\frac{1}{n}X_S^\top X_S\left(\hat{\beta}_S - \beta_S^*\right) + \frac{1}{n}X_S^\top X_{S^c}\hat{\beta}_{S^c} - \frac{1}{n}X_S^\top\epsilon + \lambda\hat{z}_S + \lambda\alpha\mathrm{Diag}\left(R_{SS}|\hat{\beta}_S|\right)\hat{z}_S = 0, \tag{A.3}$$

$$\frac{1}{n}X_{S^c}^\top X_S\left(\hat{\beta}_S - \beta_S^*\right) + \frac{1}{n}X_{S^c}^\top X_{S^c}\hat{\beta}_{S^c} - \frac{1}{n}X_{S^c}^\top\epsilon + \lambda\hat{z}_{S^c} + \lambda\alpha\mathrm{Diag}\left(R_{S^cS}|\hat{\beta}_S|\right)\hat{z}_{S^c} = 0. \tag{A.4}$$

Hence, there exists a critical point with correct sign recovery if and only if there exists $\hat{\beta}$ and $\hat{z}$ such that (A.3), (A.4), $\hat{z} \in \partial\|\hat{\beta}\|_1$ and $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^*)$. The latter two conditions can be written as

$$\hat{z}_S = \mathrm{sgn}(\beta_S^*), \tag{A.5}$$

$$|\hat{z}_{S^c}| \leq 1, \tag{A.6}$$

$$\mathrm{sgn}(\hat{\beta}_S) = \mathrm{sgn}(\beta_S^*), \tag{A.7}$$

$$\hat{\beta}_{S^c} = 0. \tag{A.8}$$

The condition (A.5) and (A.8) yield

$$\frac{1}{n}X_S^\top X_S\left(\hat{\beta}_S - \beta_S^*\right) - \frac{1}{n}X_S^\top\epsilon + \lambda\,\mathrm{sgn}(\beta_S^*) + \lambda\alpha\mathrm{Diag}\left(R_{SS}|\hat{\beta}_S|\right)\mathrm{sgn}(\beta_S^*) = 0, \tag{A.9}$$

$$\frac{1}{n}X_{S^c}^\top X_S\left(\hat{\beta}_S - \beta_S^*\right) - \frac{1}{n}X_{S^c}^\top\epsilon + \lambda\hat{z}_{S^c} + \lambda\alpha\mathrm{Diag}\left(R_{S^cS}|\hat{\beta}_S|\right)\hat{z}_{S^c} = 0. \tag{A.10}$$

Since

$$\begin{aligned}\mathrm{Diag}(R_{SS}|\hat{\beta}_S|)\,\mathrm{sgn}(\beta_S^*) &= \mathrm{Diag}(\mathrm{sgn}(\beta_S^*))R_{SS}|\hat{\beta}_S| \\ &= \mathrm{Diag}(\mathrm{sgn}(\beta_S^*))R_{SS}\mathrm{Diag}(\mathrm{sgn}(\beta_S^*))\hat{\beta}_S,\end{aligned}$$

(A.9) can be rewritten as

$$U(\hat{\beta}_S - \beta_S^*) + V = 0,$$

where

$$\begin{aligned}U &:= \frac{1}{n}X_S^\top X_S + \lambda\alpha\mathrm{Diag}(\mathrm{sgn}(\beta_S^*))R_{SS}\mathrm{Diag}(\mathrm{sgn}(\beta_S^*)), \\ V &:= \lambda\,\mathrm{sgn}(\beta_S^*) + \lambda\alpha\mathrm{Diag}(\mathrm{sgn}(\beta_S^*))R_{SS}\mathrm{Diag}(\mathrm{sgn}(\beta_S^*))\beta_S^* - \frac{1}{n}X_S^\top\epsilon.\end{aligned}$$

If we assume U is invertible, we obtain

$$\hat{\beta}_S = \beta_S^* - U^{-1}V. \tag{A.11}$$

Substituting this in (A.10), we have

$$\frac{1}{n}X_{S^c}^\top X_S\left(-U^{-1}V\right) - \frac{1}{n}X_{S^c}^\top\epsilon + \lambda\hat{z}_{S^c} + \lambda\alpha\mathrm{Diag}\left(R_{S^cS}|\beta_S^* - U^{-1}V|\right)\hat{z}_{S^c} = 0,$$

that is,

$$\left(1+\alpha\mathrm{Diag}\left(R_{S^cS}|\beta_S^* - U^{-1}V|\right)\right)\lambda\hat{z}_{S^c} = \frac{1}{n}X_{S^c}^\top X_S U^{-1}V + \frac{1}{n}X_{S^c}^\top\epsilon. \tag{A.12}$$

Combining (A.6), (A.7), (A.11) and (A.12), we have the following conditions:

$$\mathrm{sgn}(\beta_S^* - U^{-1}V) = \mathrm{sgn}(\beta_S^*),$$
$$\left|\frac{1}{n}X_{S^c}^\top X_S U^{-1}V + \frac{1}{n}X_{S^c}^\top\right| \le \lambda\left(1+\alpha R_{S^cS}|\beta_S^* - U^{-1}V|\right).$$

□

B Proof of Theorem 3

First, we prepare the following lemma.

Lemma B.1. Suppose that Assumption 1 and

$$\frac{1}{n}\sum_{i=1}^n X_{ij}^2 \le 1 \quad (\forall j = 1,\dots,p),$$

are satisfied. For $\forall\delta > 0$, let $\gamma_n := \gamma_n(\delta)$ be

$$\gamma_n := \sigma\sqrt{\frac{2\log(2p/\delta)}{n}}.$$

Then, we have that

$$P\left(\left\|\frac{1}{n}X^\top\epsilon\right\|_\infty \ge \gamma_n\right) \le \delta.$$

Proof. The assertion can be shown in the standard way. First notice that

$$\begin{aligned} P\left(\left\|\frac{1}{n}X^\top\epsilon\right\|_\infty \ge \gamma\right) &= P\left(\max_{1\le j\le p}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{ij}\right| \ge \gamma\right) \\ &= P\left(\bigcup_{1\le j\le p}\left\{\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{ij}\right| \ge \gamma\right\}\right) \\ &\le \sum_{j=1}^p P\left(\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{ij}\right| \ge \gamma\right) \le p\max_{1\le j\le p} P\left(\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{ij}\right| \ge \gamma\right). \end{aligned}$$

Since $\frac{1}{n}\sum_{i=1}^n X_{ij}^2 \le 1$, $\xi_i = X_{ij}\epsilon_i$ satisfies $\mathrm{E}[e^{t\xi_i}] \le e^{\sigma^2t^2/2}$ $\forall t\in\mathbb{R}$. Hence, applying Hoeffding's inequality, we obtain the assertion. □

Then, we derive Theorem 3.

Proof. By $L_{\lambda_n}(\hat\beta) \le L_{\lambda_n}(\beta^*)$ and $y = X\beta^* + \epsilon$, it holds that

$$\begin{aligned} &\frac{1}{2n}\|X(\hat\beta - \beta^*) - \epsilon\|_2^2 + \lambda_n\psi(\hat\beta) \le \frac{1}{2n}\|\epsilon\|_2^2 + \lambda_n\psi(\beta^*) \\ \Rightarrow\ &\frac{1}{2n}\|X(\hat\beta - \beta^*)\|_2^2 + \lambda_n\psi(\hat\beta) \le \frac{1}{n}\epsilon^\top X(\hat\beta - \beta^*) + \lambda_n\psi(\beta^*), \end{aligned} \tag{B.1}$$

where $\psi(\beta) = \lambda_n\left(\|\beta\|_1 + \frac{\alpha}{2}|\beta|^\top R|\beta|\right)$. By Lemma B.1, it holds that

$$P\left(\left\|\frac{1}{n}X^\top\epsilon\right\|_\infty > \gamma_n\right) \le \delta.$$

Hereafter, we assume that the event $\{\left\|\frac{1}{n}X^\top\epsilon\right\|_\infty \le \gamma_n\}$ is happening.

Then, if $\gamma_n \leq \lambda_n/3$, by (B.1),

$$\frac{1}{2n}\|X(\hat{\beta}-\beta^*)\|_2^2 + \lambda_n\psi(\hat{\beta}) \leq \frac{1}{n}\|\epsilon^\top X\|_\infty\|\beta^* - \hat{\beta}\|_1 + \lambda_n\psi(\beta^*)$$
$$\leq \gamma_n\|\beta^* - \hat{\beta}\|_1 + \lambda_n\psi(\beta^*) \leq \frac{1}{3}\lambda_n\|\beta^* - \hat{\beta}\|_1 + \lambda_n\psi(\beta^*). \quad \text{(B.2)}$$

Since

$$\|\hat{\beta}-\beta^*\|_1 = \|\hat{\beta}_S - \beta^*_S\|_1 + \|\hat{\beta}_{S^c} - \beta^*_{S^c}\|_1 = \|\hat{\beta}_S - \beta^*_S\|_1 + \|\hat{\beta}_{S^c}\|_1,$$

and

$$\begin{aligned}|\beta^*_S|^\top R_{SS}|\beta^*_S| - |\hat{\beta}_S|^\top R_{SS}|\hat{\beta}_S| &\leq \sum_{(j,k)\in S\times S} R_{jk}|\beta^*_j\beta^*_k - \hat{\beta}_j\hat{\beta}_k| \\ &\leq 2\sum_{(j,k)\in S\times S} R_{jk}|\beta^*_j(\beta^*_k - \hat{\beta}_k)| + \sum_{(j,k)\in S\times S} R_{jk}|(\beta^*_j - \hat{\beta}_j)(\beta^*_k - \hat{\beta}_k)| \\ &= 2|\beta^*_S|^\top R_{SS}|\beta^*_S - \hat{\beta}_S| + |\beta^*_S - \hat{\beta}_S|^\top R_{SS}|\beta^*_S - \hat{\beta}_S| \\ &\leq 2\|R_{SS}|\beta^*_S|\|_\infty\|\beta^*_S - \hat{\beta}_S\|_1 + D\|\beta^*_S - \hat{\beta}_S\|_1^2,\end{aligned}$$

we obtain that

$$\begin{aligned}&\frac{1}{2n}\|X(\hat{\beta}-\beta^*)\|_2^2 + \lambda_n\left(\|\hat{\beta}_S\|_1 + \|\hat{\beta}_{S^c}\|_1 + \frac{\alpha}{2}|\hat{\beta}_S|^\top R_{SS}|\hat{\beta}_S| + \frac{\alpha}{2}\sum_{(j,k)\notin S\times S} R_{jk}|\hat{\beta}_j\hat{\beta}_k|\right) \\ &\leq \frac{1}{3}\lambda_n(\|\hat{\beta}_S - \beta^*_S\|_1 + \|\hat{\beta}_{S^c}\|_1) + \lambda_n\left(\|\beta^*_S\|_1 + \frac{\alpha}{2}|\beta^*_S|^\top R_{SS}|\beta^*_S|\right) \\ \Rightarrow &\frac{1}{2n}\|X(\hat{\beta}-\beta^*)\|_2^2 + \lambda_n\left(\frac{2}{3}\|\hat{\beta}_{S^c}\|_1 + \frac{\alpha}{2}\sum_{(j,k)\notin S\times S} R_{jk}|\hat{\beta}_j\hat{\beta}_k|\right) \\ &\leq \frac{1}{3}\lambda_n\|\hat{\beta}_S - \beta^*_S\|_1 + \lambda_n\left(\|\beta^*_S\|_1 - \|\hat{\beta}_S\|_1 + \alpha\|R_{SS}|\beta^*_S|\|_\infty\|\beta^*_S - \hat{\beta}_S\|_1 + \frac{\alpha D}{2}\|\beta^*_S - \hat{\beta}_S\|_1^2\right) \\ \Rightarrow &\frac{1}{2n}\|X(\hat{\beta}-\beta^*)\|_2^2 + \lambda_n\left(\frac{2}{3}\|\hat{\beta}_{S^c}\|_1 + \frac{\alpha}{2}\sum_{(j,k)\notin S\times S} R_{jk}|\hat{\beta}_j\hat{\beta}_k|\right) \\ &\leq \lambda_n\left(\frac{4}{3}\|\hat{\beta}_S - \beta^*_S\|_1 + \alpha\|R_{SS}|\beta^*_S|\|_\infty\|\beta^*_S - \hat{\beta}_S\|_1 + \frac{\alpha D}{2}\|\beta^*_S - \hat{\beta}_S\|_1^2\right). \quad \text{(B.3)}\end{aligned}$$

On the other hand, (B.2) also gives

$$\begin{aligned}&\|\hat{\beta}_S\|_1 + \|\hat{\beta}_{S^c}\|_1 \leq \frac{1}{3}(\|\hat{\beta}_S - \beta^*_S\|_1 + \|\hat{\beta}_{S^c}\|_1) + \|\beta^*_S\|_1 + \frac{\alpha}{2}|\beta^*_S|^\top R_{SS}|\beta^*_S| \\ \Rightarrow\quad &\frac{2}{3}\|\hat{\beta}_S - \beta^*_S\|_1 + \frac{2}{3}\|\hat{\beta}_{S^c}\|_1 \leq 2\|\beta^*_S\|_1 + \frac{\alpha}{2}|\beta^*_S|^\top R_{SS}|\beta^*_S| \\ \Rightarrow\quad &\|\hat{\beta}_S - \beta^*_S\|_1 \leq 3\|\beta^*_S\|_1 + \frac{3}{4}\alpha|\beta^*_S|^\top R_{SS}|\beta^*_S| \\ \Rightarrow\quad &\|\hat{\beta}_S - \beta^*_S\|_1 \leq \left(3 + \frac{3}{4}\alpha\|R_{SS}|\beta^*_S|\|_\infty\right)\|\beta^*_S\|_1.\end{aligned}$$

Therefore, (B.3) gives

$$\begin{aligned}&\frac{2}{3}\|\hat{\beta}_{S^c}\|_1 + \frac{\alpha}{2}\sum_{(j,k)\notin S\times S} R_{jk}|\hat{\beta}_j\hat{\beta}_k| \\ &\leq \left(\frac{4}{3} + \alpha\|R_{SS}|\beta^*_S|\|_\infty + \frac{3}{2}\alpha D\|\beta^*_S\|_1\left(1 + \frac{\alpha}{4}\|R_{SS}|\beta^*_S|\|_\infty\right)\right)\|\hat{\beta}_S - \beta^*_S\|_1. \quad \text{(B.4)}\end{aligned}$$

The second term of the left side is evaluated as

$$\sum_{(j,k)\notin S\times S} R_{jk}|\hat{\beta}_j\hat{\beta}_k| = \sum_{j\in S^c,k\in S^c} R_{jk}|\hat{\beta}_j\hat{\beta}_k| + 2\sum_{j\in S,k\in S^c} R_{jk}|(\hat{\beta}_j - \beta_S^* + \beta_S^*)\hat{\beta}_k|$$
$$=|\hat{\beta}_{S^c}|^\top R_{S^cS^c}|\hat{\beta}_{S^c}| + 2|\hat{\beta}_{S^c}|^\top R_{S^cS}|\hat{\beta}_S - \beta_S^* + \beta_S^*|.$$

Hence, (B.4) gives

$$\frac{2}{3}\|\hat{\beta}_{S^c}\|_1 + \frac{\alpha}{2}|\hat{\beta}_{S^c}|^\top R_{S^cS^c}|\hat{\beta}_{S^c}| + \alpha|\hat{\beta}_{S^c}|^\top R_{S^cS}|\hat{\beta}_S - \beta_S^* + \beta_S^*|$$
$$\leq \left(\frac{4}{3} + \alpha\|R_{SS}|\beta_S^*|\|_\infty + \frac{3}{2}\alpha D\|\beta_S^*\|_1\left(1 + \frac{\alpha}{4}\|R_{SS}|\beta_S^*|\|_\infty\right)\right)\|\hat{\beta}_S - \beta_S^*\|_1$$
$$\Rightarrow \quad \|\hat{\beta}_{S^c}\|_1 + \frac{3}{4}\alpha|\hat{\beta}_{S^c}|^\top R_{S^cS^c}|\hat{\beta}_{S^c}| + \frac{3}{2}\alpha|\hat{\beta}_{S^c}|^\top R_{S^cS}|\hat{\beta}_S - \beta_S^* + \beta_S^*|$$
$$\leq \left(2 + \frac{15}{4}\alpha D\|\beta_S^*\|_1 + \frac{9}{16}(\alpha D\|\beta_S^*\|_1)^2\right)\|\hat{\beta}_S - \beta_S^*\|_1. \tag{B.5}$$

If $\alpha \leq \frac{1}{4D\|\beta_S^*\|_1}$, we have

$$\|\hat{\beta}_{S^c}\|_1 + \frac{3}{4}\alpha|\hat{\beta}_{S^c}|^\top R_{S^cS^c}|\hat{\beta}_{S^c}| + \frac{3}{2}\alpha|\hat{\beta}_{S^c}|^\top R_{S^cS}|\hat{\beta}_S - \beta_S^* + \beta_S^*| \leq 3\|\hat{\beta}_S - \beta_S^*\|_1.$$

Therefore, we can see that

$$\Delta\beta \in \mathcal{B}(S, C, C'),$$

where $\Delta\beta = \hat{\beta} - \beta^*$, $C = 3$ and $C' = \frac{3}{2}$. By applying the definition of ϕ_{GRE} to (B.3), it holds that

$$\frac{\phi_{\mathrm{GRE}}}{2}\|\hat{\beta} - \beta^*\|_2^2 \leq \lambda_n\left(\frac{4}{3} + \frac{5}{2}\alpha D\|\beta_S^*\|_1 + \frac{3}{8}(\alpha D\|\beta_S^*\|_1)^2\right)\|\hat{\beta}_S - \beta_S^*\|_1$$

Because $\|\hat{\beta}_S - \beta_S^*\|_1^2 \leq s\|\hat{\beta}_S - \beta_S^*\|_2^2$, we have

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{\left(\frac{8}{3} + 5\alpha D\|\beta_S^*\|_1 + \frac{3}{4}(\alpha D\|\beta_S^*\|_1)^2\right)\sqrt{s}\lambda_n}{\phi_{\mathrm{GRE}}}$$
$$\Rightarrow \quad \|\hat{\beta} - \beta^*\|_2^2 \leq \frac{\left(\frac{8}{3} + 5\alpha D\|\beta_S^*\|_1 + \frac{3}{4}(\alpha D\|\beta_S^*\|_1)^2\right)^2 s\lambda_n^2}{\phi_{\mathrm{GRE}}^2} \leq \frac{16s\lambda_n^2}{\phi_{\mathrm{GRE}}^2} \tag{B.6}$$

This concludes the assertion. □

C Corollary of Theorem 3

For comparison with IILasso and Lasso, we use the following a little bit stricter bound.

Corollary C.1. *Suppose the same assumption of Theorem 3 except for* $\alpha \leq \frac{1}{4D\|\beta_S^*\|_1}$ *and Assumption GRE*$(S, 3, \frac{3}{2})$. *Instead, suppose that Assumption GRE*$(S, C, \frac{3}{2})$ *(Definition 1) where* $C = 2 + \frac{15}{4}\alpha D\|\beta_S^*\|_1 + \frac{9}{16}(\alpha D\|\beta_S^*\|_1)^2$ *is satisfied. Then, it holds that*

$$\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{\left(\frac{8}{3} + 5\alpha D\|\beta_S^*\|_1 + \frac{3}{4}(\alpha D\|\beta_S^*\|_1)^2\right)^2 s\lambda_n^2}{\phi_{\mathrm{GRE}}^2},$$

with probability $1 - \delta$.

Proof. This is derived basically in the same way as Theorem 3. From (B.5), we can see directly that

$$\Delta\beta \in \mathcal{B}(S, C, C'),$$

where $\Delta\beta = \hat{\beta} - \beta^*$, $C = 2 + \frac{15}{4}\alpha D\|\beta_S^*\|_1 + \frac{9}{16}(\alpha D\|\beta_S^*\|_1)^2$ and $C' = \frac{3}{2}$. This and (B.6) concludes the assertion. □

From this corollary, we can compare Lasso and IILasso with $R_{SS} = O$.

- If $\alpha = 0$, we have
$$\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{64 s \lambda_n^2}{9\phi_{\mathrm{GRE}}^2},$$
with $\mathcal{B}(S, C, C')$ where $C = 2$ and $C' = 0$. This is a standard Lasso result.
- If $D = 0$, we have
$$\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{64 s \lambda_n^2}{9\phi_{\mathrm{GRE}}^2},$$
with $\mathcal{B}(S, C, C')$ where $C = 2$ and $C' = \frac{3}{2}$. Since ϕ_{GRE} is the minimum eigenvalue restricted by $\mathcal{B}(S, C, C')$, ϕ_{GRE} of IILasso is larger than that of Lasso.

D Proof of Theorem 4

Proof. Let
$$\check{\beta} := \mathop{\arg\min}_{\beta \in \mathbb{R}^p : \beta_{S^c} = 0} \|y - X\beta\|_2^2.$$
That is, $\check{\beta}$ is the least squares estimator with the true non-zero coefficients. Let $\tilde{\beta}$ be a local optimal solution. For $0 < h < 1$, letting $\beta(h) := \tilde{\beta} + h(\check{\beta} - \tilde{\beta})$, then it holds that
$$\begin{aligned} L_{\lambda_n}(\beta(h)) - L_{\lambda_n}(\tilde{\beta}) = & \frac{h^2 - 2h}{2n} \|X(\tilde{\beta} - \check{\beta})\|_2^2 - \frac{h}{n}(X\check{\beta} - y)^\top X(\tilde{\beta} - \check{\beta}) \\ & + \lambda_n(\|\beta(h)\|_1 - \|\tilde{\beta}\|_1) + \frac{\lambda_n \alpha}{2}(|\beta(h)|^\top R |\beta(h)| - |\tilde{\beta}|^\top R |\tilde{\beta}|). \end{aligned} \quad \text{(D.1)}$$

First we evaluate the term $\frac{1}{n}(X\check{\beta} - y)^\top X(\tilde{\beta} - \check{\beta}) = \frac{1}{n}(X\check{\beta} - y)^\top X_S(\tilde{\beta}_S - \check{\beta}_S) + \frac{1}{n}(X\check{\beta} - y)^\top X_{S^c}(\tilde{\beta}_{S^c} - \check{\beta}_{S^c})$ as follows:
(1) Since $\check{\beta}$ is the least squares estimator and $\frac{1}{n}X_S^\top X_S$ is invertible by the assumption, we have
$$\check{\beta}_S = (X_S^\top X_S)^{-1} X_S^\top y, \quad \check{\beta}_{S^c} = 0.$$
Therefore,
$$\frac{1}{n} X_S^\top (X\check{\beta} - y) = \frac{1}{n} X_S^\top (X_S (X_S^\top X_S)^{-1} X_S^\top - I) y.$$
Here, $I - X_S(X_S^\top X_S)^\top X_S^\top$ is the projection matrix to the orthogonal complement of the image of $(X_S^\top X_S)^\top$. Hence, $\frac{1}{n}(X\check{\beta} - y)^\top X_S(\tilde{\beta}_S - \check{\beta}_S) = 0$.
(2) Noticing that
$$\begin{aligned} \frac{1}{n} X_{S^c}^\top (X\check{\beta} - y) &= -\frac{1}{n} X_{S^c}^\top (I - X_S(X_S^\top X_S)^{-1} X_S^\top) y \\ &= -\frac{1}{n} X_{S^c}^\top (I - X_S(X_S^\top X_S)^{-1} X_S^\top)(X_S \beta_S^* + \epsilon) \\ &= -\frac{1}{n} X_{S^c}^\top (I - X_S(X_S^\top X_S)^{-1} X_S^\top)\epsilon, \end{aligned}$$
where we used $(I - X_S(X_S^\top X_S)^{-1} X_S^\top) X_{S^c} = 0$ in the last line. Because $(I - X_S(X_S^\top X_S)^\top X_S^\top)$ is a projection matrix, we have $\|(I - X_S(X_S^\top X_S)^{-1} X_S^\top) X_j\|_2^2 \leq \|X_j\|_2^2$. This and Lemma B.1 gives
$$\left\| \frac{1}{n} X_{S^c}^\top (X\check{\beta} - y) \right\|_\infty \leq \gamma_n,$$
with probability $1 - \delta$. Hence, let $V := \mathrm{supp}(\tilde{\beta}) \backslash S$, then we have
$$\left| \frac{1}{n} (\tilde{\beta}_{S^c} - \check{\beta}_{S^c})^\top X_{S^c}^\top (X\check{\beta} - y) \right| \leq \gamma_n \|\tilde{\beta}_{S^c} - \check{\beta}_{S^c}\|_1 = \gamma_n \|\tilde{\beta}_V\|_1.$$

where we used the assumption $V \subseteq S^c$ and $\check{\beta}_V = 0$.

Combining these inequalities and the assumption $\lambda_n \geq \gamma_n$, we have that

$$\left| \frac{1}{n}(X\check{\beta} - y)^\top X(\tilde{\beta} - \check{\beta}) \right| \leq \lambda_n \|\tilde{\beta}_V\|_1. \tag{D.2}$$

As for the regularization term, we evaluate each term of $\lambda_n(\|\beta(h)\|_1 - \|\tilde{\beta}\|_1) + \frac{\lambda_n}{2}(|\beta(h)|^\top R|\beta(h)| - |\tilde{\beta}|^\top R|\tilde{\beta}|)$ in the following.

(i) Evaluation of $\|\beta(h)\|_1 - \|\tilde{\beta}\|_1$. Because of the definition of $\beta(h)$, it holds that

$$\begin{aligned} \|\beta(h)\|_1 - \|\tilde{\beta}\|_1 &= \|\tilde{\beta} + h(\check{\beta} - \tilde{\beta})\|_1 - \|\tilde{\beta}\|_1 \\ &= \|\tilde{\beta}_S + h(\check{\beta}_S - \tilde{\beta}_S)\|_1 - \|\tilde{\beta}_S\|_1 + \|\tilde{\beta}_V + h(\check{\beta}_V - \tilde{\beta}_V)\|_1 - \|\tilde{\beta}_V\|_1 \\ &= \|\tilde{\beta}_S + h(\check{\beta}_S - \tilde{\beta}_S)\|_1 - \|\tilde{\beta}_S\|_1 + (1-h)\|\tilde{\beta}_V\|_1 - \|\tilde{\beta}_V\|_1 \\ &\leq h\|\check{\beta}_S - \tilde{\beta}_S\|_1 - h\|\tilde{\beta}_V\|_1. \end{aligned} \tag{D.3}$$

(ii) Evaluation of $|\beta(h)|^\top R|\beta(h)| - |\tilde{\beta}|^\top R|\tilde{\beta}|$. Note that

$$\begin{aligned} &|\beta(h)_j|R_{jk}|\beta(h)_k| - |\tilde{\beta}_j|R_{jk}|\tilde{\beta}_k| \\ &= |(1-h)\tilde{\beta}_j + h\check{\beta}_j|R_{jk}|(1-h)\tilde{\beta}_k + h\check{\beta}_k| - |\tilde{\beta}_j|R_{jk}|\tilde{\beta}_k| \\ &\leq (1-h)^2|\tilde{\beta}_j|R_{jk}|\tilde{\beta}_k| + h(1-h)(|\check{\beta}_j|R_{jk}|\tilde{\beta}_k| + |\tilde{\beta}_j|R_{jk}|\check{\beta}_k|) \\ &\quad + h^2|\check{\beta}_j|R_{jk}|\check{\beta}_k| - |\tilde{\beta}_j|R_{jk}|\tilde{\beta}_k| \\ &= -2h|\tilde{\beta}_j|R_{jk}|\tilde{\beta}_k| + h(|\check{\beta}_j|R_{jk}|\tilde{\beta}_k| + |\tilde{\beta}_j|R_{jk}|\check{\beta}_k|) + O(h^2) \\ &= h[(|\check{\beta}_j| - |\tilde{\beta}_j|)R_{jk}|\tilde{\beta}_k| + |\tilde{\beta}_j|R_{jk}(|\check{\beta}_k| - |\tilde{\beta}_k|)] + O(h^2). \end{aligned} \tag{D.4}$$

If $j, k \in S$, then the right hand side of Eq. (D.4) is bounded by

$$\begin{aligned} &h(|\check{\beta}_j - \tilde{\beta}_j|R_{jk}|\check{\beta}_k - \tilde{\beta}_k| + |\check{\beta}_j - \tilde{\beta}_j|R_{jk}|\check{\beta}_k - \tilde{\beta}_k|) \\ &\quad + h(|\check{\beta}_j - \tilde{\beta}_j|R_{jk}|\check{\beta}_k| + |\check{\beta}_j|R_{jk}|\check{\beta}_k - \tilde{\beta}_k|) + O(h^2). \end{aligned}$$

If $j \in V$ and $k \in S$, then the right hand side of Eq. (D.4) is bounded by

$$h|\tilde{\beta}_j|R_{jk}(|\check{\beta}_k| - |\tilde{\beta}_k|) + O(h^2) \leq h|\tilde{\beta}_j|R_{jk}|\check{\beta}_k - \tilde{\beta}_k| + O(h^2).$$

If $j \in V$ and $k \in V$, then the right hand side of Eq. (D.4) is bounded by

$$0 + O(h^2) = O(h^2).$$

Based on these evaluations, we have

$$\begin{aligned} &|\beta(h)|^\top R|\beta(h)| - |\tilde{\beta}|^\top R|\tilde{\beta}| \\ \leq& 2h\left(|\check{\beta}_S - \tilde{\beta}_S|^\top R_{SS}|\check{\beta}_S - \tilde{\beta}_S| + |\check{\beta}_S - \tilde{\beta}_S|^\top R_{SS}|\check{\beta}_S| + |\tilde{\beta}_V|^\top R_{VS}|\check{\beta}_S - \tilde{\beta}_S|\right) + O(h^2) \\ \leq& 2h\left(|\check{\beta} - \tilde{\beta}|^\top R|\check{\beta} - \tilde{\beta}| + |\check{\beta}_S - \tilde{\beta}_S|^\top R_{SS}|\check{\beta}_S|\right) + O(h^2) \\ \leq& 2h\bar{D}(\|\check{\beta} - \tilde{\beta}\|_2^2 + \|\check{\beta}\|_2\|\check{\beta}_S - \tilde{\beta}_S\|_2) + O(h^2). \end{aligned}$$

Here, we will show later in Eq. (D.6) that $\|\check{\beta} - \beta^*\|_2 \leq \sqrt{s}\lambda_n/\phi$, and thus it follows that

$$\|\check{\beta}\|_2 \leq \|\beta^*\|_2 + \sqrt{s}\lambda_n/\phi.$$

Therefore, we obtain that

$$\begin{aligned} &|\beta(h)|^\top R|\beta(h)| - |\tilde{\beta}|^\top R|\tilde{\beta}| \\ &\leq 2h\bar{D}\left(\|\check{\beta} - \tilde{\beta}\|_2^2 + (\|\beta^*\|_2 + \sqrt{s}\lambda_n/\phi)\|\check{\beta}_S - \tilde{\beta}_S\|_2\right) + O(h^2). \end{aligned} \tag{D.5}$$

Applying the inequalities (D.2), (D.3) and (D.5) to (D.1) yields that

$$
\begin{aligned}
& L_{\lambda_n}(\beta(h)) - L_{\lambda_n}(\tilde{\beta}) \\
\leq & h\Big\{ -\frac{1}{n}\|X(\check{\beta}-\tilde{\beta})\|_2^2 + \lambda_n\|\tilde{\beta}_S - \check{\beta}_S\|_1 - (\lambda_n - \gamma_n)\|\tilde{\beta}_V\|_1 \\
& + \lambda_n\alpha\bar{D}[\|\check{\beta}-\tilde{\beta}\|_2^2 + (\|\beta^*\|_2 + \sqrt{s}\lambda_n/\phi)\|\check{\beta}_S - \tilde{\beta}_S\|_2]\Big\} + O(h^2) \\
\leq & h\Big\{ -\phi\|\check{\beta}-\tilde{\beta}\|_2^2 + \lambda_n\|\tilde{\beta}_S - \check{\beta}_S\|_1 \\
& + \lambda_n\alpha\bar{D}[\|\check{\beta}-\tilde{\beta}\|_2^2 + (\|\beta^*\|_2 + \sqrt{s}\lambda_n/\phi)\|\check{\beta}_S - \tilde{\beta}_S\|_2]\Big\} + O(h^2) \\
\leq & h\Bigg\{ \left(-\phi + \lambda_n\alpha\bar{D}\right)\|\check{\beta}-\tilde{\beta}\|_2^2 \\
& + \lambda_n\left(\|\tilde{\beta}_S - \check{\beta}_S\|_1 + \alpha\bar{D}(\|\beta^*\|_2 + \sqrt{s}\lambda_n/\phi)\|\check{\beta}_S - \tilde{\beta}_S\|_2\right)\Bigg\} + O(h^2),
\end{aligned}
$$

where we used the assumption $\lambda_n > \gamma_n$ in the second inequality.

Since we have assumed $\alpha < \min\left\{\frac{\sqrt{s}}{2\bar{D}\|\beta^*\|_2}, \frac{\phi}{2\bar{D}\lambda_n}\right\}$, the right hand side is further bounded by

$$
h\left\{-\frac{\phi}{2}\|\check{\beta}-\tilde{\beta}\|_2^2 + 2\lambda_n\sqrt{s}\|\check{\beta}_S - \tilde{\beta}_S\|_2\right\} + O(h^2).
$$

Because of this, if $\|\check{\beta}-\tilde{\beta}\|_2 > \frac{4\sqrt{s}\lambda_n}{\phi}$, then the first term becomes negative, and we conclude that, for sufficiently small $\eta > 0$, it holds that

$$
L_{\lambda_n}(\beta(h)) < L_{\lambda_n}(\tilde{\beta}),
$$

for all $0 < h < \eta$. In other word, $\tilde{\beta}$ is not a local optimal solution. Therefore, we must have

$$
\|\check{\beta}-\tilde{\beta}\|_2 \leq \frac{4\sqrt{s}\lambda_n}{\phi}
$$

Finally, notice that $\|\tilde{\beta}-\beta^*\|_2^2 \leq (\|\tilde{\beta}-\check{\beta}\|_2 + \|\beta^*-\check{\beta}\|_2)^2$ and

$$
\begin{aligned}
\|\check{\beta}-\beta^*\|_2^2 &= \|(X_S^\top X_S)^{-1}X_S^\top y - \beta_S^*\|_2^2 = \|(X_S^\top X_S)^{-1}X_S^\top(X_S\beta_S^* + \epsilon) - \beta_S^*\|_2^2 \\
&= \|(X_S^\top X_S)^{-1}X_S^\top\epsilon\|_2^2 \leq \phi^{-2}\|\frac{1}{n}X_S^\top\epsilon\|_2^2 \leq \phi^{-2}s\gamma_n^2 \leq \phi^{-2}s\lambda_n^2, \qquad \text{(D.6)}
\end{aligned}
$$

which concludes the assertion. □

E Optimization for Logistic Regression

We derive coordinate descent algorithm of IILasso for the binary objective variable. The objective function is

$$
L(\beta) = -\frac{1}{n}\sum_i\left(y_iX^i\beta - \log(1+\exp(X^i\beta))\right) + \lambda\left(\|\beta\|_1 + \frac{\alpha}{2}|\beta|^\top R|\beta|\right),
$$

where X^i is the i-th row of $X = [1, X_1, \cdots, X_p]$ and $\beta = [\beta_0, \beta_1, \cdots, \beta_p]$. Forming a quadratic approximation with the current estimate $\bar{\beta}$, we have

$$
\bar{L}(\beta) = -\frac{1}{2n}\sum_{i=1}^n w_i(z_i - X^i\beta)^2 + C(\bar{\beta}) + \lambda\left(\|\beta\|_1 + \frac{\alpha}{2}|\beta|^\top R|\beta|\right),
$$

where

$$
\begin{aligned}
z_i &= X^i\bar{\beta} + \frac{y_i - \bar{p}(X^i)}{\bar{p}(X^i)(1-\bar{p}(X^i))}, \\
w_i &= \bar{p}(X^i)(1-\bar{p}(X^i)), \\
\bar{p}(X^i) &= \frac{1}{1+\exp(-X^i\bar{\beta})}.
\end{aligned}
$$

Algorithm E.1 CDA for Logistic IILasso

for $\lambda = \lambda_{\max}, \cdots, \lambda_{\min}$ **do**
  initialize β
  while until convergence **do**
    update the quadratic approximation using the current parameters $\bar{\beta}$
    while until convergence **do**
      for $j = 1, \cdots, p$ **do**
        $\beta_j \leftarrow \frac{1}{\frac{1}{n}\sum_{i=1}^n w_i X_{ij}^2 + \lambda\alpha R_{jj}} S\left(\frac{1}{n}\sum_{i=1}^n w_i \left(z_i - X_{i,-j}\beta_{-j}\right) X_{ij},\ \lambda\left(1 + \alpha R_{j,-j}|\beta_{-j}|\right)\right)$
      end for
    end while
  end while
end for

To derive the update equation, when $\beta_j \neq 0$, differentiating the quadratic objective function with respect to β_j yields

$$\begin{aligned}
\partial_{\beta_j} \bar{L}(\beta) &= -\frac{1}{n}\sum_{i=1}^n w_i (z_i - X^i \beta) X_{ij} + \lambda\left(\operatorname{sgn}(\beta_j) + \alpha R_j^\top |\beta| \operatorname{sgn}(\beta_j)\right) \\
&= -\frac{1}{n}\sum_{i=1}^n w_i \left(z_i - X_{i,-j}\beta_{-j}\right) X_{ij} + \left(\frac{1}{n}\sum_{i=1}^n w_i X_{ij}^2 + \lambda R_{jj}\right)\beta_j + \lambda\left(1 + \alpha R_{j,-j}|\beta_{-j}|\right)\operatorname{sgn}(\beta_j).
\end{aligned}$$

This yields

$$\beta_j \leftarrow \frac{1}{\frac{1}{n}\sum_{i=1}^n w_i X_{ij}^2 + \lambda\alpha R_{jj}} S\left(\frac{1}{n}\sum_{i=1}^n w_i \left(z_i - X_{i,-j}\beta_{-j}\right) X_{ij},\ \lambda\left(1 + \alpha R_{j,-j}|\beta_{-j}|\right)\right).$$

These procedures amount to a sequence of nested loops. The whole algorithm is described in Algorithm E.1.