Appendix D. Technical Results and Proofs for Logistic Regression

In this section our goal is to establish sparsity and rates of convergence of the Post-Lasso Logistic estimator. Both of these properties require us to also revisit the analysis of the $\ell_1$-penalize logistic regression (Lasso-Logistic) estimator. In what follows we use a more compact notation, specifically $\eta = (\alpha, \beta)'$, $\tilde{x}_i = (d_i, x_i')'$, $\eta_0 = (\alpha_0, \beta_0)'$. Thus the Lasso-Logistic estimator is defined as any vector $\hat{\eta}$ such that

$$\hat{\eta} \in \arg \min_{\eta} \Lambda(\eta) + \frac{\lambda}{n} \|\eta\|_1. \quad (D.53)$$

We will also consider the post-model selection Logistic estimator associated with a support $\hat{T}^* \subset \{1, \ldots, p\}$ defined as

$$\tilde{\eta} \in \arg \min_{\eta} \Lambda(\eta) : \text{support}(\eta) \subseteq \hat{T}^*. \quad (D.54)$$

D.1. Design conditions and Relations. Next we collect relevant quantities associated with the design matrix $E_n[\tilde{x}_i \tilde{x}_i']$ and the weighted counterpart $E_n[w_i \tilde{x}_i \tilde{x}_i']$ where $w_i = G_i(1 - G_i) \in [0,1]$, $G_i = G(\tilde{x}_i' \eta_0)$, $i = 1, \ldots, n$, is the conditional variance of the outcome variable $y_i$. The non-weighted quantities are well studied in the literature (namely restricted eigenvalue, minimum and maximal sparse eigenvalues).

Definition 1. For $T = \text{support}(\eta_0)$, $|T| \geq 1$, the (logistic) restricted eigenvalue is defined as

$$\kappa_c := \min_{\|\delta_T\|_1 \leq c \|\delta_T\|} \frac{\|\sqrt{w_i \tilde{x}_i' \delta}\|_{2,n}}{\|\tilde{x}_i' \delta\|_{2,n}}.$$

Definition 2. For a subset $A \subset \mathbb{R}^p$ let the non-linear impact coefficient be defined as

$$\bar{q}_A = \inf_{\delta \in A} E_n \left[ w_i |\tilde{x}_i' \delta|^2 \right]^{3/2} / \mathbb{E}_n \left[ w_i |\tilde{x}_i' \delta|^3 \right].$$

In this work we will apply this for $A = \Delta_c$ and $A = \{\delta \in \mathbb{R}^p : \|\delta\|_0 \leq Cs\}$.

The definitions above differ from their counterpart in the analysis of $\ell_1$-penalized least squares estimators by the weighting $0 \leq w_i \leq 1$. Thus it will be relevant to understand their relations through the quantities

$$\psi_{(r)}(c) := \min_{\|\delta_T\|_1 \leq c \|\delta_T\|} \frac{\|\sqrt{w_i \tilde{x}_i' \delta}\|_{2,n}}{\|\tilde{x}_i' \delta\|_{2,n}} \quad \text{and} \quad \psi_{(s)}(m) := \min_{1 \leq \|\delta\|_0 \leq m} \frac{\|\sqrt{w_i \tilde{x}_i' \delta}\|_{2,n}}{\|\tilde{x}_i' \delta\|_{2,n}}$$

Lemma 6 provides three relationships between the weighted versions and the non-weighted versions. Neither dominates the other. Most papers in the literature focus on the first pair of relations which entails assuming that $\min_{i \leq n} w_i$ is bounded away from zero uniformly in $n$. The second and third pairs of relations allow for better control in the presence of a few small weights. The second pair states that if the average harmonic mean of the weights is bounded the ratio between the weighted and non-weighted quantities is controlled by the intrinsic sparsity.
Lemma 6 (Relating weighted and non-weighted design quantities). Letting $w_i = G_i(1 - G_i)$ we have the following inequalities $\psi(r)(c) \geq \min_{i \leq n} \sqrt{w_i}$ and $\psi(s)(m) \geq \min_{i \leq n} \sqrt{w_i}$:

$$\psi(r)(c) \geq \frac{\kappa_n^u \{ E_n[1/w_i]\}^{-1/2}}{\sqrt{\delta} (1 + c) \max_{i \leq n} \| \bar{x}_i \|_\infty} \quad \text{and} \quad \psi(s)(m) \geq \frac{\sqrt{\phi_{\min}(m) \{ E_n[1/w_i]\}^{-1/2}}}{\sqrt{m} \max_{i \leq n} \| \bar{x}_i \|_\infty},$$

where $\kappa_n^u$ is the original (non-weighted) restricted eigenvalue. Moreover, for any $\epsilon \in (0, 1]$ we have

$$\psi(r)(c) \geq \sqrt{\epsilon} \kappa_n^u \left( 1 - E_n[1\{w_i \leq \epsilon\}] \right) \frac{(1 + c)^2 \max_{i \leq n} \| \bar{x}_i \|_\infty^2}{\kappa_n^u^2}$$

and

$$\psi(s)(m) \geq \sqrt{\epsilon} \phi_{\min}(m) \left( 1 - E_n[1\{w_i \leq \epsilon\}] \right) \frac{m \max_{i \leq n} \| \bar{x}_i \|_\infty^2}{\phi_{\min}(m)}.$$ 

Proof. The first pair of bounds is trivial since $w_i \geq 0$. To show the second pair we have

$$E_n[\| \bar{x}_i \delta\|^2] = E_n[\sqrt{w_i} |\bar{x}_i \delta| / \sqrt{w_i}] \leq \{E_n[w_i|\bar{x}_i \delta|^2]\}^{1/2} \cdot \{E_n[|\bar{x}_i \delta|^2 / w_i]\}^{1/2} \leq \{E_n[w_i|\bar{x}_i \delta|^2]\}^{1/2} \cdot \{E_n[1/w_i]\}^{1/2} \max_{i \leq n} \| \bar{x}_i \|_\infty$$

Therefore, for $\vartheta_\delta = \| \bar{x}_i \delta\|_{2,n}/\| \delta \|_1$ we have

$$\frac{\| \sqrt{w_i} |\bar{x}_i \delta| \|_{2,n}}{\| \bar{x}_i \delta\|_{2,n}} \geq \frac{\| \sqrt{w_i} |\bar{x}_i \delta| \|_{2,n}}{\{E_n[1/w_i]\}^{1/2} \max_{i \leq n} \| \bar{x}_i \|_\infty} \frac{\{E_n[w_i|\bar{x}_i \delta|^2]\}^{1/2}}{\| \bar{x}_i \delta\|_{2,n}} \frac{1}{\max_{i \leq n} \| \bar{x}_i \|_\infty} \frac{\{E_n[1/w_i]\}^{1/2}}{\| \bar{x}_i \delta\|_{2,n}}$$

By cancelling out $\| \sqrt{w_i} |\bar{x}_i \delta| \|_{2,n}/\| \bar{x}_i \delta\|_{2,n}$ and squaring both sides we have

$$\vartheta_\delta / \max_{i \leq n} \| \bar{x}_i \|_\infty \geq \vartheta_\delta / \| \delta \|_1.$$

The result follows by noting that for $\delta \in \Delta_c$ we have $\vartheta_\delta \geq \kappa_n^u / \sqrt{(1 + \epsilon) \delta}$ and for any non-zero $\delta$ with $\| \delta \|_0 \leq m$ we have $\vartheta_\delta \geq \sqrt{\phi_{\min}(m) / \sqrt{m}}$.

The third pair follows from noting that

$$E_n[w_i|\bar{x}_i \delta|^2] = E_n[w_i1\{w_i \geq \epsilon\} |\bar{x}_i \delta|^2] + E_n[w_i1\{w_i \leq \epsilon\} |\bar{x}_i \delta|^2] \geq \epsilon E_n[|\bar{x}_i \delta|^2] - \epsilon E_n[1\{w_i \leq \epsilon\} |\bar{x}_i \delta|^2]$$

Moreover, by definition of $\vartheta_\delta$ we have

$$E_n[1\{w_i \leq \epsilon\} |\bar{x}_i \delta|^2] \leq E_n[1\{w_i \leq \epsilon\}] \max_{i \leq n} \| \bar{x}_i \|_\infty \| \delta \|_2^2 \leq E_n[1\{w_i \leq \epsilon\}] \max_{i \leq n} \| \bar{x}_i \|_\infty \| \bar{x}_i \delta\|_{2,n}^2.$$

The result follows.

D.2. Identification Lemmas. In this section we collect new identification results for Logistic regression that might be of independent interest. We build upon the following technical lemma of [1] which is based on (modified) self-concordant functions. However we will apply it differently than in [1]. We exploit the separability of the objective function across observations and make use of the restricted non-linear impact coefficient [2]. In turn this allows us to weaken requirements of the analysis when compared to the literature.
Lemma 7 (Lemma 1 from [1]). Let $g : \mathbb{R} \to \mathbb{R}$ be a convex three times differentiable function such that for all $t \in \mathbb{R}$, $|g'''(t)| \leq M g''(t)$ for some $M \geq 0$. Then, for all $t \geq 0$ we have

$$
\frac{g''(0)}{M^2} \{\exp(-Mt) + Mt - 1\} \leq g(t) - g(0) - g'(0) t \leq \frac{g''(0)}{M^2} \{\exp(Mt) + Mt - 1\}.
$$

Lemma 8. For $t \geq 0$ we have $\exp(-t) + t - 1 \geq \frac{1}{2} t^2 - \frac{1}{6} t^3$.

**Proof of Lemma 8.** For $t \geq 0$, consider the function $f(t) = \exp(-t) + t^3/6 - t^2/2 + t - 1$. The statement is equivalent to $f(t) \geq 0$ for $t \geq 0$. It follows that $f(0) = 0$, $f'(0) = 0$, and $f''(t) = \exp(-t) + t - 1 \geq 0$ so that $f$ is convex. Therefore $f(t) \geq f(0) + t f'(0) = 0$. 

Lemma 9 (Minoration Lemma). We have that

$$
\Lambda(\eta_0 + \delta) - \Lambda(\eta_0) - \nabla \Lambda(\eta_0)\delta \geq \left\{ \frac{1}{2}\|\sqrt{w_i} \bar{x}_i\delta\|_2^2, n \right\} \wedge \left\{ \frac{\bar{q}_A}{3}\|\sqrt{w_i} \bar{x}_i\delta\|_2, n \right\}.
$$

**Proof.** Step 1. (Minoration). Define the maximal radius over which the following criterion function can be bounded below by a suitable quadratic function

$$
r_A = \sup_r \left\{ r : \Lambda(\eta_0 + \delta) - \Lambda(\eta_0) - \nabla \Lambda(\eta_0)\delta \geq \frac{1}{2}\|\sqrt{w_i} \bar{x}_i\delta\|_2^2, n \right\}
$$

Step 2 below shows that $r_A \geq \bar{q}_A$. By construction of $r_A$ and the convexity of $\Lambda(\eta_0 + \delta) - \Lambda(\eta_0) - \nabla \Lambda(\eta_0)\delta$,

$$
\Lambda(\eta_0 + \delta) - \Lambda(\eta_0) - \nabla \Lambda(\eta_0)\delta \geq \|\sqrt{w_i} \bar{x}_i\delta\|_2^2, n \wedge \inf_{\delta \in A, \|\sqrt{w_i} \bar{x}_i\delta\|_2, n \leq r_A} \Lambda(\eta_0 + \delta) - \Lambda(\eta_0) - \nabla \Lambda(\eta_0)\delta.
$$

Step 2. $(r_A \geq \bar{q}_A)$ Defining $g_i(t) = \log\{1 + \exp(\bar{x}_i'\eta_0 + t\bar{x}_i\delta)\}$ we have

$$
\Lambda(\eta_0 + \delta) - \Lambda(\eta_0) - \nabla \Lambda(\eta_0)\delta =
\begin{align*}
\mathbb{E}_n [\log\{1 + \exp(\bar{x}_i'\eta_0 + \delta)\}] - y_i \bar{x}_i'\eta_0 + \delta] \\
- \mathbb{E}_n [\log\{1 + \exp(\bar{x}_i'\eta_0 - y_i \bar{x}_i'\eta_0)\} - \mathbb{E}_n [\{G_i - y_i\} \bar{x}_i'\delta] \\
\mathbb{E}_n [\log\{1 + \exp(\bar{x}_i'\eta_0 + \delta)\}] - \log\{1 + \exp(\bar{x}_i'\eta_0)\} - G_i \bar{x}_i'\delta] \\
\mathbb{E}_n [g_i(1) - g_i(0) - 1 \cdot g_i'(0)]
\end{align*}
$$

Note that the function $g_i$ is three times differentiable and satisfies, for $G_i(t) := \exp(\bar{x}_i'\eta_0 + t\bar{x}_i\delta)/\{1 + \exp(\bar{x}_i'\eta_0 + t\bar{x}_i\delta)\}$,

$$
g_i'''(t) = (\bar{x}_i\delta)G_i(t), \quad g_i''(t) = (\bar{x}_i\delta)^2G_i(t)[1 - G_i(t)], \quad g_i'''(t) = (\bar{x}_i\delta)^3G_i(t)[1 - G_i(t)][1 - 2G_i(t)].
$$

Thus $|g_i'''(t)| \leq |\bar{x}_i\delta| g_i''(t)$. Therefore, by Lemmas 7 and 8 we have

$$
g_i(1) - g_i(0) - 1 \cdot g_i'(0) \geq \left\{ \frac{(\bar{x}_i\delta)^2 w_i}{(2\delta)^2} \{\exp(-|\bar{x}_i\delta|) + |\bar{x}_i\delta| - 1\} \right\}
$$

Therefore we have

$$
\Lambda(\eta_0 + \delta) - \Lambda(\eta_0) - \nabla \Lambda(\eta_0)\delta \geq \frac{1}{\mathbb{E}_n, n} \left\{ w_i |\bar{x}_i\delta|^2 \right\} - \frac{1}{\mathbb{E}_n, n} \left\{ w_i |\bar{x}_i\delta|^3 \right\}.
$$
Note that for any \( \delta \in A \) such that \( \| \sqrt{w_i} \hat{x}_i \delta \|_{2,n} \leq \bar{q}_A \) we have
\[
\| \hat{x}_i' \delta \|_{2,n} \leq \bar{q}_A \leq \| \sqrt{w_i} \hat{x}_i \|_{2,n}^{2}/\mathbb{E}_n \left[ w_i | \hat{x}_i' \delta |^3 \right],
\]
so that \( \mathbb{E}_n[ w_i | \hat{x}_i' \delta |^3 ] \leq \mathbb{E}_n[ w_i | \hat{x}_i' \delta |^2 ] \). Therefore we have
\[
\Lambda(\eta_0 + \delta) - \Lambda(\eta_0) - \nabla \Lambda(\eta_0)' \delta \geq \frac{1}{2} \mathbb{E}_n \left[ w_i | \hat{x}_i' \delta |^2 \right] - \frac{1}{6} \mathbb{E}_n \left[ w_i | \hat{x}_i' \delta |^3 \right]
\geq \frac{1}{6} \mathbb{E}_n \left[ w_i | \hat{x}_i' \delta |^2 \right]
\]

#### D.3. Penalty Choice and Rate for \( \ell_1 \)-Penalized Logistic Regression.
Next we establish a simple (and known) bound for the choice of the penalty level \( \lambda \) within Lasso-Logistic under standard normalization. Refinements are possible under additional mild assumptions on the covariates.

**Lemma 10** (Choice of Penalty, Hoeffding’s Inequality). Assume that \( \mathbb{E}_n[ \hat{x}_{ij}^2 ] = 1 \). Then, for any \( \gamma \in (0, 1) \) we have
\[
P \left( \| \nabla \Lambda(\eta_0) \|_\infty \leq \sqrt{2 \log(2(p + 1)/\gamma)/n} \right) \leq \gamma.
\]

**Proof.** Let \( G_i = \mathbb{E}[y_i | \hat{x}_i] = \frac{\exp(\hat{x}_i' \eta_0)}{1+\exp(\hat{x}_i' \eta_0)} \), so that \( \| \nabla \Lambda(\eta_0) \|_\infty = \| \mathbb{E}_n[(y_i - G_i) \hat{x}_i] \|_\infty \). Then
\[
P(\| \mathbb{E}_n[(y_i - G_i) \hat{x}_i] \|_\infty \geq t) \leq (p + 1) \max_{j \leq p} P(\| \mathbb{E}_n[(y_i - G_i) \hat{x}_{ij}] \|_\infty \geq t) \leq 2(p + 1) \exp(-t^2 n/2).
\]

**Lemma 11** (Choice of Penalty, Self-Normalized Moderate Deviation Theory). Normalize the covariates so that \( \mathbb{E}_n[ \hat{x}_{ij}^2 ] = 1 \), let \( \tilde{l}_j = \sqrt{\mathbb{E}_n[ w_i \hat{x}_{ij}^2 ]} \), and \( \tilde{l}_j = \sqrt{\mathbb{E}_n[ \hat{w}_i \hat{x}_{ij}^2 ]} \). Assume that \( K_{\bar{l}}^2 \log p \leq n \delta_n \min_j l_j^2 \), \( \Phi^{-1}(1 - 2p/\gamma) \leq \delta_n n^{1/3} \), and \( \| \hat{w}_i - w_i \|_{2,n} K_{\bar{l}} \leq \delta_n \min_j l_j^2 \). Then, setting \( \hat{\Gamma} = \text{diag}(\hat{l}) \), for any \( \gamma \in (0, 1) \) and \( \mu > 0 \), for \( n \) sufficiently large we have
\[
P \left( \| \hat{\Gamma}^{-1} \nabla \Lambda(\eta_0) \|_\infty \leq 1 + (1 + \mu) \Phi^{-1}(1 - \gamma/2p)/\sqrt{n} \right) \leq \gamma + o(1).
\]

**Proof.** Let \( \Gamma = \text{diag}(l) \), \( \tilde{l}_j = \sqrt{\mathbb{E}_n[(y_i - G_i)\hat{x}_{ij}^2]} \), and \( \tilde{l}_j = \sqrt{\mathbb{E}_n[\hat{w}_i \hat{x}_{ij}^2]} \). We have
\[
\| \hat{\Gamma}^{-1} \nabla \Lambda(\eta_0) \|_\infty \leq \| \{ \tilde{l}_j^{-1} - 1 \} \hat{\Gamma}^{-1} \nabla \Lambda(\eta_0) \|_\infty + \| \hat{\Gamma}^{-1} \nabla \Lambda(\eta_0) \|_\infty \leq \{ \| \tilde{l}_j^{-1} \hat{\Gamma} \|_\infty + \| \gamma - 1 \|_\infty + \| \gamma - 1 \|_\infty \| \hat{\Gamma}^{-1} \nabla \Lambda(\eta_0) \|_\infty \} + \| \hat{\Gamma}^{-1} \nabla \Lambda(\eta_0) \|_\infty \]
By Lemma 3 we have
\[
\max_{j \in [p]} |(\mathbb{E}_n - \bar{E})(y_i - G_i)^2 \bar{x}_{ij}| \lesssim p \sqrt{\log p} \max_{j \in [p]} \{\mathbb{E}_n[\bar{x}_{ij}]\}^{1/2}.
\]
Therefore for \(n\) large enough we have \(\max_{j \in [p]} |\tilde{\delta}_{i,j} - \hat{\delta}_{i,j}| \sqrt{\frac{1}{|\hat{l}|}} \lesssim \mu/16\) under the assumed growth conditions with probability \(1 - o(1)\). In the same event we have
\[
\|\tilde{\Gamma}^{-1} \nabla \Lambda(\eta_0)\|_\infty \leq \{1 + \mu/2\}\|\tilde{\Gamma}^{-1} \nabla \Lambda(\eta_0)\|_\infty.
\]
Finally, by self-normalized moderate deviation theory we have
\[
P(\|\tilde{\Gamma}^{-1} \nabla \Lambda(\eta_0)\|_\infty > t) \leq p \max_{j \in [p]} \left( \mathbb{E}_n[(y_i - G_i)^2 \bar{x}_{ij}] \right) > t \leq 2p\Phi^{-1}(1 - \gamma/[2p])\{1 + O(\delta_n)\}
\]

**Remark D.1.** Note that we can replace \((\tilde{w}_i)_{i=1}^\eta\) with \((\bar{w}_i)_{i=1}^\eta\) in Lemma 11 if \(\tilde{w}_i \geq w_i\) by construction. For instance \(w_i \leq \bar{w}_i := 1/4\). Therefore it is valid to use \(\lambda = \frac{c}{\sqrt{n}}\Phi^{-1}(1 - \gamma/[2p])\) and \(\hat{l}_j = 1\) for \(c > 1\).

**Lemma 12.** Assume \(\lambda/n \geq c\|\nabla \Lambda(\eta_0)\|_\infty\), \(c > 1\) and let \(c = (c + 1)/(c - 1)\). Provided that \(\tilde{\gamma}_\Delta > 3(1 + \frac{1}{c})\lambda\sqrt{s}/(n\kappa_c)\)
\[
\|\sqrt{\tilde{w}_i} \tilde{x}'_i(\tilde{\eta} - \eta_0)\|_{2,n} \leq 3(1 + \frac{1}{c})\lambda\sqrt{s}/(n\kappa_c) \quad \text{and} \quad \|\tilde{\eta} - \eta_0\|_1 \leq 3\left(\frac{1 + c)(1 + c)}{c}\right)\frac{\lambda s}{n\kappa_c^2}
\]

**Proof.** Let \(\delta = \tilde{\eta} - \eta_0\). By definition of \(\tilde{\eta}\) in (D.53) we have \(\Lambda(\tilde{\eta}) = \frac{\lambda}{n}\|\tilde{\eta}\|_1 \leq \Lambda(\eta_0) = \frac{\lambda}{n}\|\eta_0\|_1\). Thus,
\[
\Lambda(\tilde{\eta}) - \Lambda(\eta_0) \leq \frac{\lambda}{n}\|\eta_0\|_1 - \frac{\lambda}{n}\|\tilde{\eta}\|_1 \\
\leq \frac{\lambda}{n}\|\delta_T\|_1 - \frac{\lambda}{n}\|\delta_{T^c}\|_1
\]

However, by convexity of \(\Lambda(\cdot)\) and Holder inequality we have
\[
\Lambda(\tilde{\eta}) - \Lambda(\eta_0) \geq -\|\nabla \Lambda(\eta_0)\|_\infty \|\delta\|_1 \\
\geq -\frac{1}{\kappa_c}||\delta_T\|_1 - \frac{1}{\kappa_c}||\delta_{T^c}\|_1
\]
Combining these relations we have
\[
\frac{1}{\kappa_c}||\delta_T\|_1 - \frac{1}{\kappa_c}||\delta_{T^c}\|_1 \leq \frac{\lambda}{n\kappa_c^2}||\delta_T\|_1 - \frac{1}{\kappa_c}||\delta_{T^c}\|_1
\]
which leads to \(||\delta_{T^c}\|_1 \leq c||\delta_T\|_1\).

By Lemma 9 with \(A = \Delta_c\) and the reasoning above we have
\[
\frac{1}{3}||\sqrt{\tilde{w}_i} \tilde{x}'_i\delta\|_{2,n}^2 \wedge \left\{ \frac{2\lambda}{c}\|\sqrt{\tilde{w}_i} \tilde{x}'_i\delta\|_{2,n} \right\} \leq \Lambda(\tilde{\eta}) - \Lambda(\eta_0) - \nabla \Lambda(\eta_0)'\delta \\
\leq \frac{1}{\kappa_c}||\delta_T\|_1 - \frac{1}{\kappa_c}||\delta_{T^c}\|_1 + \||\nabla \Lambda(\eta_0)\|_\infty \|\delta\|_1 \\
\leq (1 + \frac{1}{c})\frac{\lambda}{c}\||\delta_T\|_1 - (1 + \frac{1}{c})\frac{\lambda s}{n\kappa_c^2}||\delta_T\|_1 \\
\leq (1 + \frac{1}{c})\frac{\lambda s}{n\kappa_c^2}||\sqrt{\tilde{w}_i} \tilde{x}'_i\delta\|_{2,n}/\kappa_c
\]
Provided that \(\tilde{\gamma}_A > 3(1 + \frac{1}{c})\lambda\sqrt{s}/(\kappa_c n)\), so that the minimum on the LHS needs to be the quadratic term, we have
\[
\|\sqrt{\tilde{w}_i} \tilde{x}'_i\delta\|_{2,n} \leq 3(1 + \frac{1}{c})\frac{\lambda s}{n\kappa_c^2}
\]


D.4. **Sparsity of Lasso-Logistic.** We begin by establishing sparsity bounds which do not rely on large penalty choices nor on the irrepresentability condition. The (data-driven) sparsity is fundamental for the analysis of the rate of convergence of the Post-Lasso-Logistic estimator. The following lemma is useful.

**Lemma 13.** The logistic link function satisfies \(|G(t + t_0) - G(t_0)| \leq G'(t_0)\{\exp(|t|) - 1\}. If \(|t| \leq 1\) we have \(\exp(|t|) - 1 \leq 2|t|\).

**Proof.** Note that \(|G''(s)| \leq |G'(s)|\) for all \(s\). So that \(-1 \leq \frac{d}{ds} \log(G'(s)) = \frac{G''(s)}{G'(s)} \leq 1\). Suppose \(s \geq 0\). Therefore

\[-s \leq \log(G'(s + t_0)) - \log(G'(t_0)) \leq s.

In turn this implies \(G'(t_0) \exp(-s) \leq G'(s + t_0) \leq G'(t_0) \exp(s)\). Integrating one more time from 0 to \(t\),

\[G'(t_0)\{1 - \exp(-t)\} \leq G(t + t_0) - G(t_0) \leq G'(t_0)\{\exp(t) - 1\}.

The first result follows by noting that \(1 - \exp(-t) \leq \exp(t) - 1\). The second follows by verification. \(\blacksquare\)

**Lemma 14 (Sparsity).** Consider \(\hat{\eta}\) as defined in (D.53), let \(\hat{s} = |\text{support}(\hat{\eta})|\) and suppose \(\lambda/n \geq c\|\nabla \Lambda(\eta_0)\|_\infty\). Then

\[\hat{s} \leq \frac{c^2(n/\lambda)^2}{(c-1)^2} \phi_{\max}(\hat{s}) \|\hat{x}'(\hat{\eta} - \eta_0)\|_2^2/n.

Provided that \(q_{\Delta_c} > 3(1 + \frac{1}{c})\lambda \sqrt{\pi}/(n\kappa_c)\) we have

\[\hat{s} \leq s \cdot \phi_{\max}(\hat{s}) \frac{9c^2}{(\psi(r)(c))^2\kappa_c^2}.

Moreover, if \(\frac{3(1+c)(1+c)}{c} \frac{\lambda s}{n\kappa_c} \max_{i \leq n} \|\hat{x}_i\|_\infty \leq 1\) we have

\[\sqrt{s} \leq 6c\sqrt{\phi_{\max}(\hat{s})} \sqrt{s} \quad \text{and} \quad \hat{s} \leq s \cdot 36c^2 \min_{m \in \mathcal{M}} \frac{\phi_{\max}(m)}{\kappa_c^2}

where \(\mathcal{M} = \{m \in \mathbb{N} : m > 72c^2\phi_{\max}(m)/\kappa_c^2\}\)

**Proof.** Let \(\hat{T} = \text{support}(\hat{\eta}), \hat{s} = |\hat{T}|, \delta = \hat{\eta} - \eta_0,\) and \(\hat{G}_i = \exp(\hat{x}'_i \hat{\eta})/\{1 + \exp(\hat{x}'_i \hat{\eta})\}\). For any \(j \in \hat{T}\) we have \(|\nabla \Lambda(\hat{\eta})| = |\mathbb{E}_n[(y_i - \hat{G}_i)x_{ij}]| = \lambda/n\).

The first relation follows from

\[\frac{\Delta}{\sqrt{s}} = \frac{\|\mathbb{E}_n[(y_i - \hat{G}_i)x_{ij}]\|_2}{\sqrt{s}} \leq \frac{\|\mathbb{E}_n[(y_i - \hat{G}_i)x_{ij}]\|_2 + \|\mathbb{E}_n[(\hat{G}_i - G_i)x_{ij}]\|_2}{\sqrt{s}} \leq \frac{\|\mathbb{E}_n[(y_i - \hat{G}_i)x_{ij}]\|_\infty + \|\mathbb{E}_n[\hat{x}'_i \delta x_{ij}]\|_2}{\sqrt{s}} \leq \frac{\Delta}{\sqrt{s}} + \sqrt{\phi_{\max}(\hat{s})} \|\hat{x}'_i \delta x_{ij}\|_2

The second follows from the first, the definition of \(\psi(r)(c)\), and Lemma 12 so that

\[\hat{s} \leq \frac{c^2(n/\lambda)^2}{(c-1)^2} \phi_{\max}(\hat{s}) \|\hat{x}'_i \delta x_{ij}\|_2 \leq \frac{c^2(n/\lambda)^2}{(c-1)^2} \phi_{\max}(\hat{s}) \frac{\sqrt{\mathbb{E}_n[\hat{x}'_i \delta x_{ij}]^2}}{\psi(r)(c)^2} \leq s \cdot \phi_{\max}(\hat{s}) \frac{9c^2}{\psi(r)(c)^2\kappa_c^2}\]

The irrepresentability condition is the assumption that \(\|\mathbb{E}_n[\hat{x}'_i \delta x_{ij}]\{\mathbb{E}_n[\hat{x}'_i x_{ij}]\}^{-1}\text{sign}(\eta_0T)\|_\infty < 1\).
The third relation follows from

\[ \frac{1}{\sqrt{n}} \sqrt{s} = \| E_n[(y_i - \hat{G}_i)\hat{x}_i] \|_2 \]
\[ \leq \| E_n[(y_i - G_i)\hat{x}_i] \|_2 + \| E_n[(\hat{G}_i - G_i)\hat{x}_i] \|_2 \]
\[ \leq \sqrt{\| E_n[(y_i - G_i)\hat{x}_i] \|_2} \sup_{\| \theta \|_0 \leq \hat{p}_i} 1 = E_n[|\hat{G}_i - G_i| \cdot |\hat{x}'_i\theta|] \]
\[ \leq \frac{1}{\sqrt{n}} \sqrt{s} + 2\sqrt{\phi_{\text{max}}(s)}\| \acute{w}_i\hat{x}_i\delta \|_2,n \]

where we used Lemma 13 so that \( |\hat{G}_i - G_i| \leq w_i |\hat{x}'_i\delta| \) since by Lemma 12 we have \( \| \delta \|_1 \leq 3(1+\epsilon) \frac{\lambda_0}{n^2} \) so that \( \max_{i \leq n} \| \hat{x}_i \|_{\infty} \| \delta \|_1 \leq 1 \) by the assumed condition.

Therefore, by the \( \| \cdot \|_2,n \) bound in Lemma 12 we have

\[ (1 - \frac{1}{c}) \frac{1}{\sqrt{n}} \sqrt{s} \leq 6\sqrt{\phi_{\text{max}}(s)}(1+\epsilon) \frac{\lambda_\sqrt{s}}{n^2} \]

which implies \( \sqrt{s} \leq 6c\sqrt{\phi_{\text{max}}(s)}(1+\epsilon) \frac{\lambda_\sqrt{s}}{n^2} \).

The last relation follows by the previous result and the fact that sparse eigenvalues are sublinear functions.

D.5. Post model selection Logistic regression rate.

Lemma 15. Consider \( \tilde{\eta} \) as defined in (D.54). Let \( \tilde{s}^* := |\tilde{T}^*| \). We have

\[ \| \sqrt{w_i}\tilde{x}'_i(\tilde{\eta} - \eta_0) \|_2,n \leq \sqrt{3} \max \{0, \Lambda(\tilde{\eta}) - \Lambda(\eta_0)\} \]
\[ + 3\sqrt{\tilde{s}^* + s} \| \nabla \Lambda(\eta_0) \|_\infty / \sqrt{\phi_{\text{min}}(\tilde{s}^* + s)} \]

provided that \( \tilde{q}_A/6 > \sqrt{\tilde{s}^* + s} \| \nabla \Lambda(\eta_0) \|_\infty / \sqrt{\phi_{\text{min}}(\tilde{s}^* + s)} \) and \( \tilde{q}_A/6 > \sqrt{\max \{0, \Lambda(\tilde{\eta}) - \Lambda(\eta_0)\} \} \) for \( A = \{ \delta \in \mathbb{R}^p : \| \delta \|_0 \leq \tilde{s}^* + s \} \).

Proof. Let \( \tilde{\delta} = \tilde{\eta} - \eta_0 \) and \( \tilde{t}_{2,n} = \| \sqrt{w_i}\tilde{x}'_i\tilde{\delta} \|_2,n \). By Lemma 9 with \( A = \{ \delta \in \mathbb{R}^p : \| \delta \|_0 \leq \tilde{s}^* + s \} \), we have

\[ \frac{1}{\sqrt{n}} \tilde{t}_{2,n} \wedge \{ \frac{4}{\sqrt{n}} \tilde{t}_{2,n} \} \leq \Lambda(\tilde{\eta}) - \Lambda(\eta_0) - \nabla \Lambda(\eta_0)\tilde{\delta} \]
\[ \leq \Lambda(\tilde{\eta}) - \Lambda(\eta_0) + \| \nabla \Lambda(\eta_0) \|_\infty \| \tilde{\delta} \|_1 \]
\[ \leq \max \{0, \Lambda(\tilde{\eta}) - \Lambda(\eta_0)\} + \tilde{t}_{2,n} \| \nabla \Lambda(\eta_0) \|_\infty \sqrt{\phi_{\text{min}}(\tilde{s}^* + s)} \]

Provided that \( \tilde{q}_A/6 > \sqrt{\tilde{s}^* + s} \| \nabla \Lambda(\eta_0) \|_\infty / \sqrt{\phi_{\text{min}}(\tilde{s}^* + s)} \) and \( \tilde{q}_A/6 > \sqrt{\max \{0, \Lambda(\tilde{\eta}) - \Lambda(\eta_0)\} \} \), if the minimum on the LHS is the linear term, we have \( \tilde{t}_{2,n} \leq \sqrt{\max \{0, \Lambda(\tilde{\eta}) - \Lambda(\eta_0)\} \} \) which implies the result.

Otherwise, since for positive numbers \( a^2 \leq b + ac \) implies \( a \leq \sqrt{b} + c \), we have

\[ \tilde{t}_{2,n} \leq \sqrt{3} \max \{0, \Lambda(\tilde{\eta}) - \Lambda(\eta_0)\} + 3\sqrt{\tilde{s}^* + s} \| \nabla \Lambda(\eta_0) \|_\infty / \sqrt{\phi_{\text{min}}(\tilde{s}^* + s)} \].

Appendix E. Additional Monte Carlo

E.1. Monte Carlo for Approximately Sparse Models. In this section we provide further simulations to illustrate the performance of the proposed methods. In particular we illustrate the performance of the method when applied to approximately sparse models. We consider a similar design to the one used in Section 4 of the main text, namely

\[ E[y \mid d, x] = G(da_0 + x'c_{\nu_y}), \quad d = x'c_{\nu_d} + v. \]
However, the vectors $\nu_y$ and $\nu_d$ are set to

$$
\nu_{yj} = 1/j^2, \nu_{dj} = 1/j^2,
$$

so they are approximately sparse. Again we let $x = (1, z')'$ consists of an intercept and covariates $z \sim N(0, \Theta)$, and the error $v$ is i.i.d. as $N(0, 1)$. The dimension $p$ of the covariates $x$ is 250, and the sample size $n$ is 200. The regressors are correlated with $\Theta_{ij} = \rho|i-j|$ and $\rho = 0.5$. As before the coefficient $c_d$ is used to control the $R^2$ of the reduced form equation, $c_y$ is set similarly and in every repetition, we draw new errors $v_i$’s and controls $x_i$’s. The figures display the results over 100 different designs where $\alpha_0 = 0.5$ and the values of $c_y$ and $c_d$ are set to achieve $R^2 = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ for each equation. There were 1000 replications for each of the 100 designs.

Figure 6 reveals that the performance of the method for this approximately sparse design is very similar to the performance obtained with the sparse designs considered in Section 4. Again the double selection estimator arise as a more reliable estimator.

**References**

[1] Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[2] A. Belloni and V. Chernozhukov. $\ell_1$-penalized quantile regression for high dimensional sparse models. *Annals of Statistics*, 39(1):82–130, 2011.

[3] A. Belloni, V. Chernozhukov, and C. Hansen. Lasso methods for gaussian instrumental variables models. arXiv: [math.ST], http://arxiv.org/abs/1012.1297, 2010.

[4] A. Belloni, V. Chernozhukov, and K. Kato. Robust inference in high-dimensional approximately sparse quantile regression models. arXiv, (1312.7186), 2013.

[5] A. Belloni, V. Chernozhukov, and K. Kato. Uniform post selection inference for LAD regression models and other Z-estimators. *Biometrika*, (102):77–94, 2015.

[6] Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

[7] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Ying Wei. Uniformly valid post-regularization confidence regions for many functional parameters in Z-estimation framework. (arXiv:1512.07619), 2015.

[8] Alexandre Belloni, Victor Chernozhukov, Iván Fernández-Val, and Chris Hansen. Program evaluation with high-dimensional data. *arXiv preprint arXiv:1311.2645, forthcoming Econometrica*, 2013.

[9] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference methods for high-dimensional sparse econometric models. *Advances in Economics and Econometrics, 10th World Congress of Econometric Society*, Volume III, Econometrics, Edited by Daron Acemoglu, Manuel Arellano and Ede Dialektik, 245–295, 2013.

[10] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

[11] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[12] F. Bunea. Honest variable selection in linear and logistic regression models via $\ell_1$ and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2:11531194, 2008.

[13] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.

[14] Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Series in Statistics. Springer, Berlin, 2008.

[15] M. Kwemou. Non-asymptotic oracle inequalities for the lasso and group lasso in high dimensional logistic model. *arXiv preprint*, (arXiv:1206.0710), 2012.
[16] M. Ledoux and M. Talagrand. *Probability in Banach Spaces (Isoperimetry and processes)*. Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer-Verlag, 1991.

[17] Hannes Leeb and Benedikt M. Pötscher. Model selection and inference: facts and fiction. *Economic Theory*, 21:21–59, 2005.

[18] Hannes Leeb and Benedikt M. Pötscher. Can one estimate the conditional distribution of post-model-selection estimator? *The Annals of Statistics*, 34(5):2554–2591, 2006.

[19] Hannes Leeb and Benedikt M. Pötscher. Sparse estimators and the oracle property, or the return of Hodges’ estimator. *J. Econometrics*, 142(1):201–211, 2008.

[20] L. Meier, V. Van der Geer, and P. Bühlmann. The group lasso for logistic regression. *J. R. Statist. Soc.: Series B (Statist. Methodol.)*, 70(1):5371, 2008.

[21] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[22] J. Neyman. Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander, editor, *Probability and Statistics, the Harold Cramer Volume*. New York: John Wiley and Sons, Inc., 1959.

[23] J. Neyman. $c(\alpha)$ tests and their use. *Sankhyā*, 41:1–21, 1979.

[24] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482 – 494, Jan. 2013.

[25] Benedikt M. Pötscher. Confidence sets based on sparse estimators are necessarily large. *Sankhyā*, 71(1, Ser. A):1–18, 2009.

[26] Benedikt M. Pötscher and Hannes Leeb. On the distribution of penalized maximum likelihood estimators: the LASSO, SCAD, and thresholding. *J. Multivariate Anal.*, 100(9):2065–2082, 2009.

[27] P. Ravikumar, M. Wainwright, and J. Lafferty. High-dimensional ising model selection using -regularized logistic regression. *Ann. Statist.*, 38(2):1287–1319, 2010.

[28] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *ArXiv:1106.1151*, 2011.

[29] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:1025-1045, 2008.

[30] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58:267–288, 1996.

[31] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.

[32] Sara van de Geer, Peter Bühlmann, Ya’acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202, 2014.

[33] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.

[34] Aad W. van der Vaart and Jon A. Wellner. Empirical process indexed by estimated functions. *IMS Lecture Notes-Monograph Series*, 55:234–252, 2007.

[35] Lie Wang. $L_1$ penalized LAD estimator for high dimensional linear regression. *J. Multivariate Anal.*, 120:135–151, 2013.

[36] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. *J. R. Statist. Soc. B*, 76:217–242, 2014.
Figure 6. For the approximately sparse model defined by (E.55), the figures display the rp(0.05) of the naive post selection estimator and the proposed confidence regions based on optimal instrument (CRD and CIT) and double selection (CRDS). There are a total of 100 different designs with α₀ = 0.5. The results are based on 1000 replications for each design.