*Econometrica Supplementary Material*

# SUPPLEMENT TO "PROGRAM EVALUATION AND CAUSAL INFERENCE WITH HIGH-DIMENSIONAL DATA"

(*Econometrica*, Vol. 85, No. 1, January 2017, 233–298)

BY A. BELLONI, V. CHERNOZHUKOV, I. FERNÁNDEZ-VAL, AND C. HANSEN

The supplementary material contains 10 appendices with additional results and some omitted proofs. Appendices G–K include additional results for Sections 2–7, respectively. Appendix L gathers auxiliary results on algebra of covering entropies. Appendices M and N contain the proofs of Sections 4 and 5 omitted from the main text. Appendix O contains the proofs of Sections 6 omitted from the main text, together with the proofs of the additional results for Section 6 in Appendix J. Appendix P reports the results of a simulation experiment.

## APPENDIX G: ADDITIONAL RESULTS FOR SECTION 2

### G.1. *Causal Interpretations for Structural Parameters*

THE QUANTITIES DISCUSSED in Sections 2.2 and 2.3 are well-defined and have causal interpretation under standard conditions. We briefly recall these conditions, using the potential outcomes notation. Let $Y_{u1}$ and $Y_{u0}$ denote the potential outcomes under the treatment states 1 and 0. These outcomes are not observed jointly, and we instead observe $Y_u = DY_{u1} + (1-D)Y_{u0}$, where $D \in \mathcal{D} = \{0, 1\}$ is the random variable indicating program participation or treatment state. Under exogeneity, $D$ is assigned independently of the potential outcomes conditional on covariates $X$, that is, $(Y_{u1}, Y_{u0}) \perp\!\!\!\perp D|X$ a.s., where $\perp\!\!\!\perp$ denotes statistical independence.

Exogeneity fails when $D$ depends on the potential outcomes. For example, people may drop out of a program if they think the program will not benefit them. In this case, instrumental variables are useful in creating quasi-experimental fluctuations in $D$ that may identify useful effects. Let $Z$ be a binary instrument, such as an offer of participation, that generates potential participation decisions $D_1$ and $D_0$ under the instrument states 1 and 0, respectively. As with the potential outcomes, the potential participation decisions under both instrument states are not observed jointly. The realized participation decision is then given by $D = ZD_1 + (1-Z)D_0$. We assume that $Z$ is assigned randomly with respect to potential outcomes and participation decisions conditional on $X$, that is, $(Y_{u0}, Y_{u1}, D_0, D_1) \perp\!\!\!\perp Z|X$ a.s.

There are many causal quantities of interest for program evaluation. Chief among these are various structural averages: $d \longmapsto \mathrm{E}_P[Y_{ud}]$, the causal ASF; $d \longmapsto \mathrm{E}_P[Y_{ud}|D=1]$, the causal ASF-T; $d \longmapsto \mathrm{E}_P[Y_{ud}|D_1 > D_0]$, the causal LASF; and $d \longmapsto \mathrm{E}_P[Y_{ud}|D_1 > D_0, D = 1]$, the causal LASF-T; as well as effects derived from them such as $\mathrm{E}_P[Y_{u1} - Y_{u0}]$, the causal ATE; $\mathrm{E}_P[Y_{u1} - Y_{u0}|D=1]$, the causal ATE-T; $\mathrm{E}_P[Y_{u1} - Y_{u0}|D_1 > D_0]$, the causal LATE; and $\mathrm{E}_P[Y_{u1} - Y_{u0}|D_1 > D_0, D = 1]$, the causal LATE-T. These causal quantities are the same as the structural parameters defined in Sections 2.2–2.3 under the following well-known sufficient condition.

ASSUMPTION G.1—Assumptions for Causal/Structural Interpretability: *The following conditions hold $P$-almost surely*: (*Exogeneity*) $((Y_{u1}, Y_{u0})_{u\in\mathcal{U}}, D_1, D_0) \perp\!\!\!\perp Z|X$; (*First Stage*) $\mathrm{E}_P[D_1|X] \neq \mathrm{E}_P[D_0|X]$; (*Non-Degeneracy*) $\mathrm{P}_P(Z=1|X) \in (0,1)$; (*Monotonicity*) $\mathrm{P}_P(D_1 \geq D_0|X) = 1$.

This condition due to Imbens and Angrist (1994) and Abadie (2003) is much-used in the program evaluation literature. It has an equivalent formulation in terms of a simultaneous

 DOI: 10.3982/ECTA12723

equation model with a binary endogenous variable; see Vytlacil (2002) and Heckman and Vytlacil (1999). For a thorough discussion of this assumption, we refer to Imbens and Angrist (1994). Using this assumption, we present an identification lemma which follows from results of Abadie (2003) and Hong and Nekipelov (2010) that both in turn build upon Imbens and Angrist (1994). The lemma shows that the parameters $\theta_{Y_u}$ and $\vartheta_{Y_u}$ defined in Sections 2.2 and 2.3 have a causal interpretation under Assumption G.1. Therefore, our referring to them as structural/causal is justified under this condition.

LEMMA G.1—Identification of Causal Effects: *Under Assumption G.1, for each $d \in \mathcal{D}$,*

$$\mathrm{E}_P[Y_{ud}|D_1 > D_0] = \theta_{Y_u}(d), \quad \mathrm{E}_P[Y_{ud}|D_1 > D_0, D = 1] = \vartheta_{Y_u}(d).$$

*Furthermore, if $D$ is exogenous, namely $D \equiv Z$ a.s., then*

$$\mathrm{E}_P[Y_{ud}|D_1 > D_0] = \mathrm{E}_P[Y_{ud}],$$
$$\mathrm{E}_P[Y_{ud}|D_1 > D_0, D = 1] = \mathrm{E}_P[Y_{ud}|D = 1].$$

## APPENDIX H: ADDITIONAL RESULTS FOR SECTION 3

COMMENT H.1—Another Strategy for Estimating $m_Z$ and $g_V$: An alternative to the strategy for modeling and estimating $m_Z$ and $g_V$ is to treat $m_Z$ as in the text via (3.7) while modeling $g_V$ through its disaggregation

$$g_V(z, x) = \sum_{d=0}^{1} e_V(d, z, x) l_D(d, z, x), \tag{H.1}$$

where the regression functions $e_V$ and $l_D$ map the support of $(D, Z, X)$, $\mathcal{DZX}$, to the real line and are defined by

$$e_V(d, z, x) := \mathrm{E}_P[V|D = d, Z = z, X = x] \quad \text{and} \tag{H.2}$$
$$l_D(d, z, x) := \mathrm{P}_P[D = d|Z = z, X = x]. \tag{H.3}$$

We will denote other potential values for the functions $e_V$ and $l_D$ by the parameters $e$ and $l$. In this alternative approach, we can again use high-dimensional methods for modeling and estimating $e_V$ and $l_D$ using the same approach as in the main paper, and we can then use the relation (H.1) to estimate $g_V$.[1] Specifically, we model the conditional expectation of $V$ given $D$, $Z$, and $X$ by

$$e_V(d, z, x) =: \Gamma_V\big[f(d, z, x)'\theta_V\big] + \varrho_V(d, z, x), \tag{H.4}$$
$$f(d, z, x) := \big((1 - d)f(z, x)', df(z, x)'\big)', \tag{H.5}$$
$$\theta_V := \big(\theta_V(0, 0)', \theta_V(0, 1)', \theta_V(1, 0)', \theta_V(1, 1)'\big)'. \tag{H.6}$$

We model the conditional probability of $D$ taking on 1 or 0, given $Z$ and $X$, by

$$l_D(1, z, x) =: \Gamma_D\big[f(z, x)'\theta_D\big] + \varrho_D(z, x), \tag{H.7}$$

[1] Upon conditioning on $D = d$, some parts become known; for example, $e_{1_d(D)Y}(d', x, z) = 0$ if $d \neq d'$ and $e_{1_d(D)}(d', x, z) = 1$ if $d = d'$.

$$l_D(0,z,x) = 1 - \Gamma_D\big[f(z,x)'\theta_D\big] - \varrho_D(z,x), \tag{H.8}$$

$$f(z,x) := \big((1-z)f(x)', zf(x)'\big)', \tag{H.9}$$

$$\theta_D := \big(\theta_D(0)', \theta_D(1)'\big)'. \tag{H.10}$$

Here $\varrho_V(d,z,x)$ and $\varrho_D(z,x)$ are approximation errors, and the functions $\Gamma_V(f(d,z,x)'\theta_V)$ and $\Gamma_D(f(z,x)'\theta_D)$ are generalized linear approximations to the target functions $e_V(d,z,x)$ and $l_D(1,z,x)$. The functions $\Gamma_V$ and $\Gamma_D$ are taken again to be known link functions from the set $\mathcal{L} = \{\mathrm{Id}, \Phi, 1-\Phi, \Lambda_0, 1-\Lambda_0\}$ defined following equation (3.7).

As in the strategy in the main text, we maintain approximate sparsity. We assume that there exist $\beta_Z$, $\theta_V$, and $\theta_D$ such that, for all $V \in \mathcal{V}$,

$$\|\theta_V\|_0 + \|\theta_D\|_0 + \|\beta_Z\|_0 \le s. \tag{H.11}$$

That is, there are at most $s = s_n \ll n$ components of $\theta_V$, $\theta_D$, and $\beta_Z$ with nonzero values in the approximations to $e_V$, $l_D$, and $m_Z$.

$$\big\{\mathrm{E}_P\big[\varrho_V^2(D,Z,X)\big]\big\}^{1/2} + \big\{\mathrm{E}_P\big[\varrho_D^2(Z,X)\big]\big\}^{1/2} + \big\{\mathrm{E}_P\big[r_Z^2(X)\big]\big\}^{1/2} \lesssim \sqrt{s/n}. \tag{H.12}$$

Note that the size of the approximating model $s = s_n$ can grow with $n$ just as in standard series estimation as long as $s^2\log^2(p\vee n)\log^2(n)/n \to 0$.

We proceed with the estimation of $e_V$ and $l_D$ analogously to the approach outlined in the main text. The Lasso estimator $\hat{\theta}_V$ and Post-Lasso estimator $\tilde{\theta}_V$ are defined analogously to $\hat{\beta}_V$ and $\tilde{\beta}_V$ using the data $(\tilde{Y}_i, \tilde{X}_i)_{i=1}^n = (V_i, f(D_i, Z_i, X_i))_{i=1}^n$ and the link function $\Lambda = \Gamma_V$. The estimator $\hat{e}_V(D,Z,X) = \Gamma_V[f(D,Z,X)'\bar{\theta}_V]$, with $\bar{\theta}_V = \hat{\theta}_V$ or $\bar{\theta}_V = \tilde{\theta}_V$, has the near oracle rate of convergence $\sqrt{(s\log p)/n}$ and other desirable properties. The Lasso estimator $\hat{\theta}_D$ and Post-Lasso estimators $\tilde{\theta}_D$ are also defined analogously to $\hat{\beta}_V$ and $\tilde{\beta}_V$ using the data $(\tilde{Y}_i, \tilde{X}_i)_{i=1}^n = (D_i, f(Z_i, X_i))_{i=1}^n$ and the link function $\Lambda = \Gamma_D$. Again, the estimator $\hat{l}_D(Z,X) = \Gamma_D[f(Z,X)'\bar{\theta}_D]$ of $l_D(Z,X)$, where $\bar{\theta}_D = \hat{\theta}_D$ or $\bar{\theta}_D = \tilde{\theta}_D$, has good theoretical properties including the near oracle rate of convergence, $\sqrt{(s\log p)/n}$. The resulting estimator for $g_V$ is then

$$\hat{g}_V(z,x) = \sum_{d=0}^{1} \hat{e}_V(d,z,x)\hat{l}_D(d,z,x). \tag{H.13}$$

The remaining estimation steps are the same as with the strategy given in the main text.

## APPENDIX I: ADDITIONAL RESULTS FOR SECTION 4

ASSUMPTION I.1—Approximate Sparsity for the Strategy of Section H.1: *Under each $P \in \mathcal{P}_n$ and for each $n \ge n_0$, uniformly for all $V \in \mathcal{V}$:* (i) *The approximations* (H.4)–(H.10) *and* (3.7) *apply with the link functions $\Gamma_V$, $\Gamma_D$, and $\Lambda_Z$ belonging to the set $\mathcal{L}$, the sparsity condition $\|\theta_V\|_0 + \|\theta_D\|_0 + \|\beta_Z\|_0 \le s$ holding, the approximation errors satisfying $\|\varrho_D\|_{P,2} + \|\varrho_V\|_{P,2} + \|r_Z\|_{P,2} \le \delta_n n^{-1/4}$ and $\|\varrho_D\|_{P,\infty} + \|\varrho_V\|_{P,\infty} + \|r_Z\|_{P,\infty} \le \epsilon_n$, and the sparsity index $s$ and the number of terms $p$ in the vector $f(X)$ obeying $s^2\log^2(p\vee n)\log^2 n \le \delta_n n$.* (ii) *There are estimators $\bar{\theta}_V$, $\bar{\theta}_D$, and $\bar{\beta}_Z$ such that, with probability no less than $1-\Delta_n$, the estimation errors satisfy $\|f(D,Z,X)'(\bar{\theta}_V - \theta_V)\|_{\mathbb{P}_n,2} + \|f(Z,X)'(\bar{\theta}_D - \theta_D)\|_{\mathbb{P}_n,2} + \|f(X)'(\bar{\beta}_Z -$*

*$\beta_Z)\|_{\mathbb{P}_n,2} \leq \delta_n n^{-1/4}$ and $K_n\|\bar\theta_V - \theta_V\|_1 + K_n\|\bar\theta_D - \theta_D\|_1 + K_n\|\bar\beta_Z - \beta_Z\|_1 \leq \epsilon_n$; the estimators are sparse such that $\|\bar\theta_V\|_0 + \|\bar\theta_D\|_0 + \|\bar\beta_Z\|_0 \leq Cs$; and the empirical and population norms induced by the Gram matrix formed by $(f(X_i))_{i=1}^n$ are equivalent on sparse subsets, $\sup_{\|\delta\|_0 \leq \ell_n s} |\|f(X)'\delta\|_{\mathbb{P}_n,2}/\|f(X)'\delta\|_{P,2} - 1| \leq \epsilon_n$.* (iii) *The following boundedness conditions hold*: $\|\|f(X)\|_\infty\|_{P,\infty} \leq K_n$ *and* $\|V\|_{P,\infty} \leq C$.

Under the stated assumptions, the empirical reduced-form process $\hat Z_{n,P} = \sqrt{n}(\hat\rho - \rho)$ defined by (3.16), but constructed using the alternative strategy for estimating $m_Z$ and $g_V$ of Comment H.1, follows a functional central limit theorem and a functional central limit theorem for the multiplier bootstrap. Theorem I.1 states these results. We omit the proof because it is analogous to the proofs of Theorems 4.1–4.2.

THEOREM I.1: *Under Assumption I.1, the results stated in Theorems* 4.1–4.2 *in the main text apply to the alternative strategy for estimating $m_Z$ and $g_V$ of Comment H.1.*

## APPENDIX J: ADDITIONAL RESULTS FOR SECTION 6: FINITE SAMPLE RESULTS OF A CONTINUUM OF LASSO AND POST-LASSO ESTIMATORS FOR FUNCTIONAL RESPONSES

### J.1. *Assumptions*

We consider the following high-level conditions which are implied by the primitive Assumptions 6.1 and 6.2. For each $n \geq 1$, there is a sequence of independent random variables $(W_i)_{i=1}^n$, defined on the probability space $(\Omega, \mathcal{A}_\Omega, \mathrm{P}_P)$ such that model (6.1) holds with $\mathcal{U} \subset [0,1]^{d_u}$. Let $d_\mathcal{U}$ be a metric on $\mathcal{U}$ (and note that the results cover the case where $d_u$ is a function of $n$). Throughout this section, we assume that the variables $(X_i, (Y_{ui}, \zeta_{ui} := Y_{ui} - \mathrm{E}_P[Y_{ui}|X_i])_{u\in\mathcal{U}})$ are generated as suitably measurable transformations of $W_i$ and $u \in \mathcal{U}$. Furthermore, this section uses the notation $\bar{\mathrm{E}}_P[\cdot] = \frac{1}{n}\sum_{i=1}^n \mathrm{E}_P[\cdot]$, because we allow for independent non-identically distributed (i.n.i.d.) data.

Consider fixed sequences of positive numbers $\delta_n \searrow 0$, $\epsilon_n \searrow 0$, and $\Delta_n \searrow 0$ at a speed at most polynomial in $n$, $\ell_n = \log n$, and $1 \leq K_n < \infty$; and positive constants $c$ and $C$ which will not vary with $P$.

CONDITION WL: *Suppose that for some $\epsilon > 0$ there is a $N_n$ such that*: (i) *we have* $\log N(\epsilon, \mathcal{U}, d_\mathcal{U}) \leq N_n$; (ii) *uniformly over $u \in \mathcal{U}$, we have that* $\max_{j\leq p} \frac{\{\bar{\mathrm{E}}_P[|f_j(X)\zeta_u|^3]\}^{1/3}}{\{\bar{\mathrm{E}}_P[|f_j(X)\zeta_u|^2]\}^{1/2}} \Phi^{-1}(1 - 1/\{2pN_n n\}) \leq \delta_n n^{1/6}$ *and* $0 < c \leq \bar{\mathrm{E}}_P[|f_j(X)\zeta_u|^2] \leq C$, $j = 1, \ldots, p$; *and* (iii) *with probability* $1 - \Delta_n$, *we have that* $\sup_{u\in\mathcal{U}} \max_{j\leq p} |(\mathbb{E}_n - \bar{\mathrm{E}}_P)[f_j(X)^2\zeta_u^2]| \leq \delta_n$, $\log(p \vee N_n \vee n) \sup_{d_\mathcal{U}(u,u')\leq\epsilon} \max_{j\leq p} \mathbb{E}_n[f_j(X)^2(\zeta_u - \zeta_{u'})^2] \leq \delta_n$, $\sup_{d_\mathcal{U}(u,u')\leq\epsilon} \|\mathbb{E}_n[f(X)(\zeta_u - \zeta_{u'})]\|_\infty \leq \delta_n n^{-1/2}$.

The following technical lemma justifies the choice of penalty level $\lambda$. It is based on self-normalized moderate deviation theory. In what follows, for $u \in \mathcal{U}$ we let $\hat\Psi_{u0}$ denote a diagonal $p \times p$ matrix of "ideal loadings" with diagonal elements given by $\hat\Psi_{u0jj} = \{\mathbb{E}_n[f_j^2(X)\zeta_u^2]\}^{1/2}$ for $j = 1, \ldots, p$.

LEMMA J.1—Choice of $\lambda$: *Suppose Condition WL holds, let $c' > c > 1$ be constants, $\gamma \in [1/n, 1/\log n]$, and $\lambda = c'\sqrt{n}\Phi^{-1}(1 - \gamma/\{2pN_n\})$. Then for $n \geq n_0$ large enough depending only on Condition WL*,

$$\mathrm{P}_P\Big(\lambda/n \geq c \sup_{u\in\mathcal{U}} \big\|\hat\Psi_{u0}^{-1}\mathbb{E}_n\big[f(X)\zeta_u\big]\big\|_\infty\Big) \geq 1 - \gamma - o(1).$$

We note that Condition WL(iii) contains high-level conditions on the process $(Y_u, \zeta_u)_{u\in\mathcal{U}}$. The following lemma provides easy to verify sufficient conditions that imply Condition WL(iii).

LEMMA J.2: *Suppose the i.i.d. sequence* $((Y_{ui}, \zeta_{ui})_{u\in\mathcal{U}}, X_i)$, $i = 1, \dots, n$, *satisfies the following conditions*: (i) $c \leq \max_{j\leq p} \mathrm{E}_P[f_j(X)^2] \leq C$, $\max_{j\leq p} |f_j(X)| \leq K_n$, $\sup_{u\in\mathcal{U}} \max_{i\leq n} |Y_{ui}| \leq B_n$, *and* $c \leq \sup_{u\in\mathcal{U}} \mathrm{E}_P[\zeta_u^2|X] \leq C$, *P-a.s.*; (ii) *for some random variable* $Y$, *we have* $Y_u = G(Y, u)$ *where* $\{G(\cdot, u) : u \in \mathcal{U}\}$ *is a VC-class of functions with VC index equal to* $C'd_u$; (iii) *for some fixed* $\nu > 0$, *we have* $\mathrm{E}_P[|Y_u - Y_{u'}|^2|X] \leq L_n|u - u'|^\nu$ *for any* $u, u' \in \mathcal{U}$, *P-a.s. For* $\tilde{A} := pnK_nB_nn^\nu/L_n$, *we have, with probability* $1 - \Delta_n$,

$$\sup_{d_{\mathcal{U}}(u,u')\leq 1/n} \left\| \mathbb{E}_n\left[f(X)(\zeta_u - \zeta_{u'})\right] \right\|_\infty$$

$$\lesssim \frac{1}{\sqrt{n}} \left\{ \sqrt{\frac{(1 + d_u)L_n \log(\tilde{A})}{n^\nu}} + \frac{(1 + d_u)K_nB_n \log(\tilde{A})}{\sqrt{n}} \right\},$$

$$\sup_{d_{\mathcal{U}}(u,u')\leq 1/n} \max_{j\leq p} \mathbb{E}_n\left[f_j(X)^2(\zeta_u - \zeta_{u'})^2\right]$$

$$\lesssim L_n n^{-\nu} \left\{ 1 + \sqrt{\frac{K_n^2 \log\left(pnK_n^2\right)}{n}} + \frac{K_n^2}{n} \log\left(pnK_n^2\right) \right\},$$

$$\sup_{u\in\mathcal{U}} \max_{j\leq p} \left| (\mathbb{E}_n - \mathrm{E}_P)\left[f_j^2(X)\zeta_u^2\right] \right|$$

$$\lesssim \sqrt{\frac{(1 + d_u) \log(npK_nB_n)}{n}} + \frac{(1 + d_u)K_n^2B_n^2}{n} \log(npB_nK_n),$$

*where* $\Delta_n$ *is a fixed sequence going to zero*.

Lemma J.2 allows for several different cases including cases where $Y_u$ is generated by a non-smooth transformation of a random variable $Y$. For example, if $Y_u = 1\{Y \leq u\}$ where $Y$ has bounded conditional probability density function, we have $d_u = 1$, $B_n = 1$, $\nu = 1$, $L_n = \sup_y f_{Y|X}(y|x)$. A similar result holds for independent non-identically distributed data.

In what follows for a vector $\delta \in \mathbb{R}^p$, and a set of indices $T \subseteq \{1, \dots, p\}$, we denote by $\delta_T \in \mathbb{R}^p$ the vector such that $(\delta_T)_j = \delta_j$ if $j \in T$ and $(\delta_T)_j = 0$ if $j \notin T$. For a set $T$, $|T|$ denotes the cardinality of $T$. Moreover, let

$$\Delta_{\mathbf{c},u} := \left\{ \delta \in \mathbb{R}^p : \|\delta_{T_u^c}\|_1 \leq \mathbf{c}\|\delta_{T_u}\|_1 \right\}.$$

## J.2. *Finite Sample Results: Linear Case*

For the model described in (6.1) with $\Lambda(t) = t$ and $M(y, t) = \frac{1}{2}(y - t)^2$, we will study the finite sample properties of the associated Lasso and Post-Lasso estimators of $(\theta_u)_{u\in\mathcal{U}}$ defined in relations (6.2) and (6.3).

The analysis relies on $T_u = \mathrm{supp}(\theta_u)$, $s_u := \|\theta_u\|_0 \leq s$, with $s \geq 1$, and on the restricted eigenvalues

$$\kappa_{\mathbf{c}} = \inf_{u\in\mathcal{U}} \min_{\delta\in\Delta_{\mathbf{c},u}} \frac{\left\| f(X)'\delta \right\|_{\mathbb{P}_n,2}}{\|\delta_{T_u}\|}, \tag{J.1}$$

and maximum and minimum sparse eigenvalues

$$\phi_{\min}(m) = \min_{1\le\|\delta\|_0\le m} \frac{\|f(X)'\delta\|_{\mathbb{P}_n,2}^2}{\|\delta\|^2} \quad \text{and}$$

$$\phi_{\max}(m) = \max_{1\le\|\delta\|_0\le m} \frac{\|f(X)'\delta\|_{\mathbb{P}_n,2}^2}{\|\delta\|^2}.$$

Next we present technical results on the performance of the estimators generated by Lasso that are used in the proof of Theorem 6.1.

LEMMA J.3—Rates of Convergence for Lasso: *The events* $c_r \ge \sup_{u\in\mathcal{U}} \|r_u\|_{\mathbb{P}_n,2}$, $\ell\hat{\Psi}_{u0} \le \hat{\Psi}_u \le L\hat{\Psi}_{u0}$, $u \in \mathcal{U}$, *and* $\lambda/n \ge c\sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty$, *for* $c > 1/\ell$, *imply that, uniformly in* $u \in \mathcal{U}$,

$$\|f(X)'(\hat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \le 2c_r + \frac{2\lambda\sqrt{s}\left(L+\frac{1}{c}\right)}{n\kappa_{\tilde{\mathbf{c}}}}\|\hat{\Psi}_{u0}\|_\infty,$$

$$\|\hat{\theta}_u - \theta_u\|_1 \le 2(1+2\tilde{\mathbf{c}})\left\{\frac{\sqrt{s}c_r}{\kappa_{2\tilde{\mathbf{c}}}} + \frac{\lambda s\left(L+\frac{1}{c}\right)}{n\kappa_{\tilde{\mathbf{c}}}\kappa_{2\tilde{\mathbf{c}}}}\|\hat{\Psi}_{u0}\|_\infty\right\}$$
$$+\left(1+\frac{1}{2\tilde{\mathbf{c}}}\right)\frac{c\|\hat{\Psi}_{u0}^{-1}\|_\infty}{\ell c - 1}\frac{n}{\lambda}c_r^2,$$

*where* $\tilde{\mathbf{c}} = \sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}^{-1}\|_\infty \|\hat{\Psi}_{u0}\|_\infty (Lc+1)/(\ell c-1)$.

The following lemma summarizes sparsity properties of $(\hat{\theta}_u)_{u\in\mathcal{U}}$.

LEMMA J.4—Sparsity Bound for Lasso: *Consider the Lasso estimator* $\hat{\theta}_u$, *its support* $\hat{T}_u = \operatorname{supp}(\hat{\theta}_u)$, *and let* $\hat{s}_u = \|\hat{\theta}_u\|_0$. *Assume that* $c_r \ge \sup_{u\in\mathcal{U}} \|r_u\|_{\mathbb{P}_n,2}$, $\lambda/n \ge c\sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}^{-1} \times \mathbb{E}_n[f(X)\zeta_u]\|_\infty$, *and* $\ell\hat{\Psi}_{u0} \le \hat{\Psi}_u \le L\hat{\Psi}_{u0}$ *for all* $u \in \mathcal{U}$, *with* $L \ge 1 \ge \ell > 1/c$. *Then, for* $c_0 = (Lc+1)/(\ell c-1)$ *and* $\tilde{\mathbf{c}} = c_0 \sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}\|_\infty \|\hat{\Psi}_{u0}^{-1}\|_\infty$, *we have, uniformly over* $u \in \mathcal{U}$,

$$\hat{s}_u \le 16c_0^2\left(\min_{m\in\mathcal{M}} \phi_{\max}(m)\right)\left[\frac{nc_r}{\lambda} + \frac{\sqrt{s}}{\kappa_{\tilde{\mathbf{c}}}}\|\hat{\Psi}_{u0}\|_\infty\right]^2 \|\hat{\Psi}_{u0}^{-1}\|_\infty^2,$$

*where* $\mathcal{M} = \{m \in \mathbb{N} : m > 32c_0^2\phi_{\max}(m)\sup_{u\in\mathcal{U}}[\frac{nc_r}{\lambda} + \frac{\sqrt{s}}{\kappa_{\tilde{\mathbf{c}}}}\|\hat{\Psi}_{u0}\|_\infty]^2\|\hat{\Psi}_{u0}^{-1}\|_\infty^2\}$.

LEMMA J.5—Rate of Convergence of Post-Lasso: *Under Conditions* WL, *let* $\tilde{\theta}_u$ *be the Post-Lasso estimator based on the support* $\tilde{T}_u$. *Then, with probability* $1-o(1)$, *uniformly over* $u \in \mathcal{U}$, *we have for* $\tilde{s}_u = |\tilde{T}_u|$,

$$\|\mathrm{E}_P[Y_u|X] - f(X)'\tilde{\theta}_u\|_{\mathbb{P}_n,2} \le C\frac{\sqrt{\tilde{s}_u\log(p\vee n^{d_u+1})}}{\sqrt{n\phi_{\min}(\tilde{s}_u)}}\|\hat{\Psi}_{u0}\|_\infty$$
$$+ \min_{\operatorname{supp}(\theta)\subseteq\tilde{T}_u} \|\mathrm{E}_P[Y_u|X] - f(X)'\theta\|_{\mathbb{P}_n,2}.$$

*Moreover, if* $\mathrm{supp}(\hat\theta_u) \subseteq \tilde T_u$ *for every* $u \in \mathcal{U}$*, the following events* $c_r \geq \sup_{u\in\mathcal{U}} \|r_u\|_{\mathbb{P}_n,2}$, $\ell\hat\Psi_{u0} \leq \hat\Psi_u \leq L\hat\Psi_{u0}$, $u \in \mathcal{U}$*, and* $\lambda/n \geq c\sup_{u\in\mathcal{U}} \|\hat\Psi_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty$*, for* $c > 1/\ell$*, imply that*

$$\begin{aligned}&\sup_{u\in\mathcal{U}} \min_{\mathrm{supp}(\theta)\subseteq\tilde T_u} \big\|\mathrm{E}_P[Y_u|X] - f(X)'\theta\big\|_{\mathbb{P}_n,2}\\&\quad\leq 3c_r + \left(L + \frac{1}{c}\right)\frac{2\lambda\sqrt{s}}{n\kappa_{\tilde{\mathbf{c}}}}\sup_{u\in\mathcal{U}}\|\hat\Psi_{u0}\|_\infty.\end{aligned}$$

### J.3. *Finite Sample Results: Logistic Case*

For the model described in (6.1) with $\Lambda(t) = \exp(t)/\{1+\exp(t)\}$ and $M(y,t) = -\{1\{y=1\}\log(\Lambda(t)) + 1\{y=0\}\log(1-\Lambda(t))\}$, we will study the finite sample properties of the associated Lasso and Post-Lasso estimators of $(\theta_u)_{u\in\mathcal{U}}$ defined in relations (6.2) and (6.3). In what follows we use the notation

$$M_u(\theta) = \mathbb{E}_n\big[M\big(Y_u, f(X)'\theta\big)\big].$$

In the finite sample analysis, we will consider not only the design matrix $\mathbb{E}_n[f(X)f(X)']$ but also a weighted counterpart $\mathbb{E}_n[w_u f(X)f(X)']$ where $w_{ui} = \mathrm{E}_P[Y_{ui}|X_i](1-\mathrm{E}_P[Y_{ui}|X_i])$, $i = 1,\dots,n$, $u\in\mathcal{U}$, is the conditional variance of the outcome variable $Y_{ui}$.

For $T_u = \mathrm{supp}(\theta_u)$, $s_u = \|\theta_u\|_0 \leq s$, with $s \geq 1$, the (logistic) restricted eigenvalue is defined as

$$\bar\kappa_{\mathbf{c}} := \inf_{u\in\mathcal{U}}\min_{\delta\in\Delta_{\mathbf{c},u}} \frac{\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}}{\|\delta_{T_u}\|}. \tag{J.2}$$

For a subset $A_u \subset \mathbb{R}^p$, $u\in\mathcal{U}$, let the nonlinear impact coefficient (Belloni and Chernozhukov (2011), Belloni, Chernozhukov, and Wei (2013)) be defined as

$$\bar q_{A_u} := \inf_{\delta\in A_u} \frac{\mathbb{E}_n\big[w_u|f(X)'\delta|^2\big]^{3/2}}{\mathbb{E}_n\big[w_u|f(X)'\delta|^3\big]}. \tag{J.3}$$

Note that $\bar q_{A_u}$ can be bounded as

$$\bar q_{A_u} = \inf_{\delta\in A_u} \frac{\mathbb{E}_n\big[w_u|f(X)'\delta|^2\big]^{3/2}}{\mathbb{E}_n\big[w_u|f(X)'\delta|^3\big]} \geq \inf_{\delta\in A_u}\frac{\mathbb{E}_n\big[w_u|f(X)'\delta|^2\big]^{1/2}}{\max_{i\leq n}\|f(X_i)\|_\infty\|\delta\|_1},$$

which can lead to interesting bounds provided $A_u$ is appropriate (like the restrictive set $\Delta_{\mathbf{c},u}$ in the definition of restricted eigenvalues). In Lemma J.6, we have $A_u = \Delta_{2\tilde{\mathbf{c}},u} \cup \{\delta \in \mathbb{R}^p : \|\delta\|_1 \leq \frac{6c\|\hat\Psi_{u0}^{-1}\|_\infty}{\ell c-1}\frac{n}{\lambda}\|\frac{r_u}{\sqrt{w_u}}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}\}$, for $u\in\mathcal{U}$. For this choice of sets, and provided that with probability $1-o(1)$ we have $\ell c > c' > 1$, $\sup_{u\in\mathcal{U}}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2} \lesssim \sqrt{s\log(p\vee n)/n}$, $\sup_{u\in\mathcal{U}}\|\hat\Psi_{u0}^{-1}\|_\infty \lesssim 1$, and $\sqrt{n\log(p\vee n)} \lesssim \lambda$, we have that uniformly over $u\in\mathcal{U}$, with probability $1-o(1)$,

$$\begin{aligned}\bar q_{A_u} &\geq \frac{1}{\max_{i\leq n}\|f(X)\|_\infty}\left(\frac{\bar\kappa_{2\tilde{\mathbf{c}}}}{\sqrt{s_u}(1+2\tilde{\mathbf{c}})} \wedge \frac{(\lambda/n)(\ell c-1)}{6c\|\hat\Psi_{u0}^{-1}\|_\infty\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}\right)\\&\gtrsim \frac{\bar\kappa_{2\tilde{\mathbf{c}}}}{\sqrt{s}\max_{i\leq n}\|f(X_i)\|_\infty}.\end{aligned} \tag{J.4}$$

The definitions above differ from their counterparts in the analysis of $\ell_1$-penalized least squares estimators by the weighting $0 \leq w_{ui} \leq 1$. Thus it is relevant to understand their relations through the quantities

$$\psi_u(A) := \min_{\delta \in A} \frac{\| \sqrt{w_u} f(X)'\delta \|_{\mathbb{P}_n,2}}{\| f(X)'\delta \|_{\mathbb{P}_n,2}}.$$

Many primitive conditions on the data-generating process will imply $\psi_u(A)$ to be bounded away from zero for the relevant choices of $A$. We refer to Belloni, Chernozhukov, and Wei (2013) for bounds on $\psi_u$. For notational convenience we will also work with a rescaling of the approximation errors $\tilde{r}_u(X)$ defined as

$$\tilde{r}_{ui} = \tilde{r}_u(X_i) = \Lambda^{-1}\big(\Lambda\big(f(X_i)'\theta_u\big) + r_{ui}\big) - f(X_i)'\theta_u, \tag{J.5}$$

which is the unique solution to $\Lambda(f(X_i)'\theta_u + \tilde{r}_u(X_i)) = \Lambda(f(X_i)'\theta_u) + r_u(X_i)$. It follows that $|r_{ui}| \leq |\tilde{r}_{ui}|$ and that[2] $|\tilde{r}_{ui}| \leq |r_{ui}| / \inf_{0 \leq t \leq \tilde{r}_{ui}} \Lambda'(f(X_i'\theta_u) + t) \leq |r_{ui}|/\{w_{ui} - 2|r_{ui}|\}_+$.

Next we derive finite sample bounds provided some crucial events occur.

LEMMA J.6—Rates of Convergence for $\ell_1$-Logistic Estimator: *Assume that*

$$\lambda/n \geq c \sup_{u \in \mathcal{U}} \big\| \hat{\Psi}_{u0}^{-1} \mathbb{E}_n\big[f(X)\zeta_u\big]\big\|_\infty$$

*for $c > 1$. Further, let $\ell \hat{\Psi}_{u0} \leq \hat{\Psi}_u \leq L \hat{\Psi}_{u0}$ for $L \geq 1 \geq \ell > 1/c$, uniformly over $u \in \mathcal{U}$, $\tilde{\mathbf{c}} = (Lc+1)/(\ell c - 1)\sup_{u \in \mathcal{U}} \|\hat{\Psi}_{u0}\|_\infty \|\hat{\Psi}_{u0}^{-1}\|_\infty$, and*

$$\begin{aligned} A_u &= \Delta_{2\tilde{\mathbf{c}},u} \\ &\cup \left\{ \delta : \|\delta\|_1 \leq \frac{6c\|\hat{\Psi}_{u0}^{-1}\|_\infty}{\ell c - 1} \frac{n}{\lambda} \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2} \big\|\sqrt{w_u} f(X)'\delta\big\|_{\mathbb{P}_n,2} \right\}. \end{aligned}$$

*Provided that the nonlinear impact coefficient $\bar{q}_{A_u} > 3\{(L + \frac{1}{c})\|\hat{\Psi}_{u0}\|_\infty \frac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\}$ for every $u \in \mathcal{U}$, we have uniformly over $u \in \mathcal{U}$,*

$$\begin{aligned} &\big\|\sqrt{w_u} f(X)'(\hat{\theta}_u - \theta_u)\big\|_{\mathbb{P}_n,2} \\ &\quad \leq 3\left\{ \left(L + \frac{1}{c}\right) \|\hat{\Psi}_{u0}\|_\infty \frac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2} \right\} \quad \textit{and} \\ &\|\hat{\theta}_u - \theta_u\|_1 \leq 3\left\{ \frac{(1+2\tilde{\mathbf{c}})\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}} + \frac{6c\|\hat{\Psi}_{u0}^{-1}\|_\infty}{\ell c - 1} \frac{n}{\lambda} \left\| \frac{r_u}{\sqrt{w_u}} \right\|_{\mathbb{P}_n,2} \right\} \\ &\qquad \times \left\{ \left(L + \frac{1}{c}\right) \|\hat{\Psi}_{u0}\|_\infty \frac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}} \left\| \frac{\tilde{r}_u}{\sqrt{w_u}} \right\|_{\mathbb{P}_n,2} \right\}. \end{aligned}$$

[2]The last relation follows from noting that, for the logistic function, we have $\inf_{0 \leq t \leq \tilde{r}_{ui}} \Lambda'(f(X_i'\theta_u) + t) = \min\{\Lambda'(f(X_i'\theta_u) + \tilde{r}_{ui}), \Lambda'(f(X_i'\theta_u))\}$ since $\Lambda'$ is unimodal. Moreover, $\Lambda'(f(X_i'\theta_u) + \tilde{r}_{ui}) = w_{ui}$ and $\Lambda'(f(X_i'\theta_u)) = \Lambda(f(X_i'\theta_u))[1 - \Lambda(f(X_i'\theta_u))] = [\Lambda(f(X_i'\theta_u)) + r_{ui} - r_{ui}][1 - \Lambda(f(X_i'\theta_u)) - r_{ui} + r_{ui}] \geq w_{ui} - 2|r_{ui}|$ since $|r_{ui}| \leq 1$.

The following result provides bounds on the number of nonzero coefficients in the $\ell_1$-penalized estimator $\hat{\theta}_u$, uniformly over $u \in \mathcal{U}$.

LEMMA J.7—Sparsity of $\ell_1$-Logistic Estimator: *Assume* $\lambda/n \geq c \sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}^{-1} \times \mathbb{E}_n[f(X)\zeta_u]\|_\infty$ *for* $c > 1$. *Further, let* $\ell\hat{\Psi}_{u0} \leq \hat{\Psi}_u \leq L\hat{\Psi}_{u0}$ *for* $L \geq 1 \geq \ell > 1/c$, *uniformly over* $u \in \mathcal{U}$, $c_0 = (Lc+1)/(\ell c - 1)$, $\tilde{\mathbf{c}} = c_0 \sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}\|_\infty \|\hat{\Psi}_{u0}^{-1}\|_\infty$ *and* $A_u = \Delta_{2\tilde{\mathbf{c}},u} \cup \{\delta : \|\delta\|_1 \leq \frac{6c\|\hat{\Psi}_{u0}^{-1}\|_\infty}{\ell c - 1}\frac{n}{\lambda}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}\}$, *and* $\bar{q}_{A_u} > 3\{(L+\frac{1}{c})\|\hat{\Psi}_{u0}\|_\infty \frac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 9\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\}$ *for every* $u \in \mathcal{U}$. *Then for* $\hat{s}_u = \|\hat{\theta}_u\|_0$, *uniformly over* $u \in \mathcal{U}$,

$$\hat{s}_u \leq \Big(\min_{m\in\mathcal{M}} \phi_{\max}(m)\Big)\bigg[\frac{c_0}{\psi(A_u)}\bigg\{3\|\hat{\Psi}_{u0}\|_\infty \frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 28\tilde{\mathbf{c}}\frac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\bigg\}\bigg]^2,$$

*where* $\mathcal{M} = \{m \in \mathbb{N} : m > 2[\frac{c_0}{\psi(A_u)} \sup_{u\in\mathcal{U}}\{3\|\hat{\Psi}_{u0}\|_\infty \frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 28\tilde{\mathbf{c}}\frac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\}]^2\}$.

*Moreover, if* $\sup_{u\in\mathcal{U}} \max_{i\leq n} |f(X_i)'(\hat{\theta}_u - \theta_u) - \tilde{r}_{ui}| \leq 1$, *we have*

$$\hat{s}_u \leq \Big(\min_{m\in\mathcal{M}} \phi_{\max}(m)\Big)4c_0^2\bigg\{3\|\hat{\Psi}_{u0}\|_\infty \frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 28\tilde{\mathbf{c}}\frac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\bigg\}^2,$$

*where* $\mathcal{M} = \{m \in \mathbb{N} : m > 8c_0^2 \sup_{u\in\mathcal{U}}[3\|\hat{\Psi}_{u0}\|_\infty \frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}} + 28\tilde{\mathbf{c}}\frac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}]^2\}$.

Next we turn to finite sample bounds for the logistic regression estimator where the support was selected based on $\ell_1$-penalized logistic regression. The results will hold uniformly over $u \in \mathcal{U}$ provided the side conditions also hold uniformly over $\mathcal{U}$.

LEMMA J.8—Rate of Convergence for Post-$\ell_1$-Logistic Estimator: *Consider* $\tilde{\theta}_u$ *defined as the post-model-selection logistic regression with the support* $\tilde{T}_u$ *and let* $\tilde{s}_u := |\tilde{T}_u|$. *Uniformly over* $u \in \mathcal{U}$, *we have*

$$\begin{aligned}
&\big\|\sqrt{w_u}f(X)'(\tilde{\theta}_u - \theta_u)\big\|_{\mathbb{P}_n,2} \\
&\quad \leq \sqrt{3}\sqrt{0 \vee \big\{M_u(\tilde{\theta}_u) - M_u(\theta_u)\big\}} \\
&\qquad + 3\left\{\frac{\sqrt{\tilde{s}_u + s_u}\big\|\mathbb{E}_n\big[f(X)\zeta_u\big]\big\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\tilde{s}_u + s_u)}} + 3\left\|\frac{\tilde{r}_u}{\sqrt{w_u}}\right\|_{\mathbb{P}_n,2}\right\}
\end{aligned}$$

*provided that, for every* $u \in \mathcal{U}$ *and* $A_u = \{\delta \in \mathbb{R}^p : \|\delta\|_0 \leq \tilde{s}_u + s_u\}$,

$$\begin{aligned}
&\bar{q}_{A_u} > 6\left\{\frac{\sqrt{\tilde{s}_u + s_u}\big\|\mathbb{E}_n\big[f(X)\zeta_u\big]\big\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\tilde{s}_u + s_u)}} + 3\left\|\frac{\tilde{r}_u}{\sqrt{w_u}}\right\|_{\mathbb{P}_n,2}\right\} \quad \textit{and} \\
&\bar{q}_{A_u} > 6\sqrt{0 \vee \big\{M_u(\tilde{\theta}_u) - M_u(\theta_u)\big\}}.
\end{aligned}$$

COMMENT J.1: Since, for a sparse vector $\delta$ such that $\|\delta\|_0 = k$, we have $\|\delta\|_1 \leq \sqrt{k}\|\delta\| \leq \sqrt{k}\|f(X)'\delta\|_{\mathbb{P}_n,2}/\sqrt{\phi_{\min}(k)}$, the results above can directly establish bounds on the rate of convergence in the $\ell_1$-norm.

## APPENDIX K: ADDITIONAL RESULTS FOR SECTION 7

In this section, we report additional results to supplement those provided in the main text. Specifically, we provide results with both total wealth and net total financial assets as the outcome variable. We present detailed results for four different sets of controls $f(X)$. The first set uses the indicators of marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status, a linear term for family size, five categories for age, four categories for education, and seven categories for income (Indicator specification). We use the same definitions of categories as in Chernozhukov and Hansen (2004) and note that this is identical to the specification in Chernozhukov and Hansen (2004) and Benjamin (2003). The second through fourth specifications correspond to the Quadratic Spline specification, the Quadratic Spline Plus Interactions specification, and the Quadratic Spline Plus Many Interactions specification described in the main text.

Results for intention to treat effects based on using 401(k) eligibility as the treatment variable are given in Table S.I. In Table S.II, we report results using 401(k) participation as the treatment variable instrumenting with 401(k) eligibility. We plot the QTE and QTE-T, based on using 401(k) eligibility as the treatment variable, in Figures S.1–S.4. Finally, the LQTE and LQTE-T, based on using 401(k) participation as the treatment variability and instrumenting with eligibility, are plotted in Figures S.5–S.8. The results are broadly consistent with the discussion provided in the main text with the selection and no-selection results being similar in the low-dimensional cases and the selection results being substantially more regular in the high-dimensional cases. We also see that the patterns of point estimates for total wealth and net total financial assets are similar, though the total wealth estimates have substantially larger estimated standard errors, especially for high quantiles.

## APPENDIX L: AUXILIARY RESULTS: ALGEBRA OF COVERING ENTROPIES

LEMMA L.1—Algebra for Covering Entropies: *Work with the setup described in Appendix* C *of the main text.*

(1) *Let $\mathcal{F}$ be a VC-subgraph class with a finite VC index $k$ or any other class whose entropy is bounded above by that of such a VC-subgraph class; then the covering entropy of $\mathcal{F}$ obeys*

$$\sup_Q \log N\big(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\big) \lesssim 1 + k\log(1/\epsilon) \vee 0.$$

(2) *Forany measurable classes of functions $\mathcal{F}$ and $\mathcal{F}'$ mapping $\mathcal{W}$ to $\mathbb{R}$,*

$$\begin{aligned}
&\log N\big(\epsilon\|F+F'\|_{Q,2}, \mathcal{F}+\mathcal{F}', \|\cdot\|_{Q,2}\big)\\
&\quad\le \log N\bigg(\frac{\epsilon}{2}\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\bigg) + \log N\bigg(\frac{\epsilon}{2}\|F'\|_{Q,2}, \mathcal{F}', \|\cdot\|_{Q,2}\bigg),\\
&\log N\big(\epsilon\|F\cdot F'\|_{Q,2}, \mathcal{F}\cdot\mathcal{F}', \|\cdot\|_{Q,2}\big)\\
&\quad\le \log N\bigg(\frac{\epsilon}{2}\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\bigg) + \log N\bigg(\frac{\epsilon}{2}\|F'\|_{Q,2}, \mathcal{F}', \|\cdot\|_{Q,2}\bigg),\\
&N\big(\epsilon\|F\vee F'\|_{Q,2}, \mathcal{F}\cup\mathcal{F}', \|\cdot\|_{Q,2}\big)\\
&\quad\le N\big(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\big) + N\big(\epsilon\|F'\|_{Q,2}, \mathcal{F}', \|\cdot\|_{Q,2}\big).
\end{aligned}$$

TABLE S.I

ESTIMATES AND STANDARD ERRORS OF AVERAGE 401(K) ELIGIBILITY EFFECTS[a]

| Specification | | | Exogenous: 401(k) Eligibility | | | Endogenous: 401(k) Participation | | |
|---|---|---|---|---|---|---|---|---|
| Series Approximation | Dimension | Selection | Linear Model | ATE | ATE-T | Linear IV | LATE | LATE-T |
| Indicator | 20 | N | 9122<br>(1343) | 8266<br>(1144)<br>{1163} | 11,357<br>(1567)<br>{1635} | 6454<br>(2117) | 6268<br>(1881)<br>{1929} | 8807<br>(2517)<br>{2403} |
| Indicator | 20 | Y | 9191<br>(1348) | 9634<br>(1180)<br>{1113} | 11,701<br>(1644)<br>{1579} | 6562<br>(2121) | 8453<br>(1903)<br>{1887} | 9672<br>(2587)<br>{2604} |
| Quadratic Spline | 35 (32) | N | 8997<br>(1252) | 8093<br>(1082)<br>{967} | 11,250<br>(1513)<br>{1423} | 6194<br>(2020) | 5943<br>(1800)<br>{1823} | 8710<br>(2428)<br>{2467} |
| Quadratic Spline | 35 (32) | Y | 8967<br>(1270) | 7614<br>(1224)<br>{1234} | 10,257<br>(1776)<br>{1676} | 6293<br>(2047) | 6733<br>(1945)<br>{2002} | 7179<br>(2725)<br>{2817} |
| Quadratic Spline Plus Interactions | 311 (272) | N | 9019<br>(1258) | 11,775<br>(4202)<br>{4202} | 11,740<br>(1779)<br>{1757} | 5988<br>(2033) | 73,109<br>(36,787)<br>{36,697} | 6240<br>(2577)<br>{2650} |
| Quadratic Spline Plus Interactions | 311 (272) | Y | 8307<br>(1313) | 7077<br>(1358)<br>{1237} | 8830<br>(2133)<br>{2105} | 4775<br>(2005) | 6177<br>(1894)<br>{1908} | 7130<br>(2651)<br>{2700} |
| Quadratic Spline Plus Many Interactions | 1756 (1526) | N | 8860<br>(1358) | –<br>–<br>– | –<br>–<br>– | 5933<br>(2097) | –<br>–<br>– | –<br>–<br>– |
| Quadratic Spline Plus Many Interactions | 1756 (1526) | Y | 8536<br>(1321) | 7848<br>(1317)<br>{1334} | 9602<br>(2047)<br>{1894} | 5084<br>(1998) | 5881<br>(1912)<br>{1852} | 7142<br>(2876)<br>{2809} |

[a]The sample is drawn from the 1991 SIPP and consists of 9915 observations. All the specifications control for age, income, family size, education, marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status. Indicators specification uses a linear term for family size, five categories for age, four categories for education, and seven categories for income. Quadratic Spline uses indicators for marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status; a third-order polynomial in age; second-order polynomials in education and family size; and a piecewise quadratic polynomial in income with six break points. The "Quadratic Spline Plus Interactions" specification include all first-order interactions between the income variables and the remaining variables. The specification denoted "Quadratic Spline Plus Many Interactions" takes all first-order interactions between all non-income variables and then fully interacts these interactions as well as the main effects with all income variables. Analytic standard errors are given in parentheses. Bootstrap standard errors based on 500 repetitions with Mammen (1993) multipliers are given in braces.

TABLE S.II

ESTIMATES AND STANDARD ERRORS OF AVERAGE 401(K) PARTICIPATION EFFECTS[a]

| Specification | | | Exogenous: 401(k) Eligibility | | | Endogenous: 401(k) Participation | | |
|---|---|---|---|---|---|---|---|---|
| Series Approximation | Dimension | Selection | Linear Model | ATE | ATE-T | Linear IV | LATE | LATE-T |
| Indicator | 20 | N | 13,102 | 11,833 | 16,120 | 9307 | 8972 | 12,500 |
| | | | (1922) | (1638) | (2224) | (3038) | (2692) | (3572) |
| | | | | {1448} | {2201} | | {2572} | {3248} |
| Indicator | 20 | Y | 13,150 | 13,915 | 16,608 | 9323 | 12,210 | 13,729 |
| | | | (1929) | (1704) | (2333) | (3042) | (2749) | (3672) |
| | | | | {1684} | {2417} | | {2835} | {3616} |
| Quadratic Spline | 35 (32) | N | 12,926 | 11,579 | 15,969 | 8910 | 8503 | 12,363 |
| | | | (1796) | (1548) | (2148) | (2901) | (2575) | (3446) |
| | | | | {1413} | {2195} | | {2837} | {3611} |
| Quadratic Spline | 35 (32) | Y | 12,890 | 10,937 | 14,560 | 9079 | 9672 | 10,189 |
| | | | (1821) | (1758) | (2520) | (2941) | (2794) | (3869) |
| | | | | {1709} | {2576} | | {2880} | {3657} |
| Quadratic Spline Plus Interactions | 311 (272) | N | 12,973 | 17,529 | 16,664 | 8599 | 109,160 | 8857 |
| | | | (1804) | (6256) | (2526) | (2923) | (54,927) | (3658) |
| | | | | {6249} | {2558} | | {56,974} | {3784} |
| Quadratic Spline Plus Interactions | 311 (272) | Y | 11,784 | 10,168 | 12,533 | 6964 | 8874 | 10,120 |
| | | | (1995) | (1952) | (3027) | (2935) | (2721) | (3763) |
| | | | | {1963} | {2818} | | {2733} | {3636} |
| Quadratic Spline Plus Many Interactions | 1756 (1526) | N | 12,827 | – | – | 8601 | – | – |
| | | | (1960) | – | – | (3031) | – | – |
| | | | | – | – | | – | – |
| Quadratic Spline Plus Many Interactions | 1756 (1526) | Y | 10,671 | 11,267 | 13,629 | 4620 | 8443 | 10,137 |
| | | | (2001) | (1890) | (2906) | (2928) | (2744) | (4083) |
| | | | | {1835} | {2862} | | {2719} | {4022} |

[a]The sample is drawn from the 1991 SIPP and consists of 9915 observations. All the specifications control for age, income, family size, education, marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status. Indicators specification uses a linear term for family size, five categories for age, four categories for education, and seven categories for income. Quadratic Spline uses indicators for marital status, two-earner status, defined benefit pension status, IRA participation status, and home ownership status; a third-order polynomial in age; second-order polynomials in education and family size; and a piecewise quadratic polynomial in income with six break points. The "Quadratic Spline Plus Interactions" specification include all first-order interactions between the income variables and the remaining variables. The specification denoted "Quadratic Spline Plus Many Interactions" takes all first-order interactions between all non-income variables and then fully interacts these interactions as well as the main effects with all income variables. Analytic standard errors are given in parentheses. Bootstrap standard errors based on 500 repetitions with Mammen (1993) multipliers are given in braces.

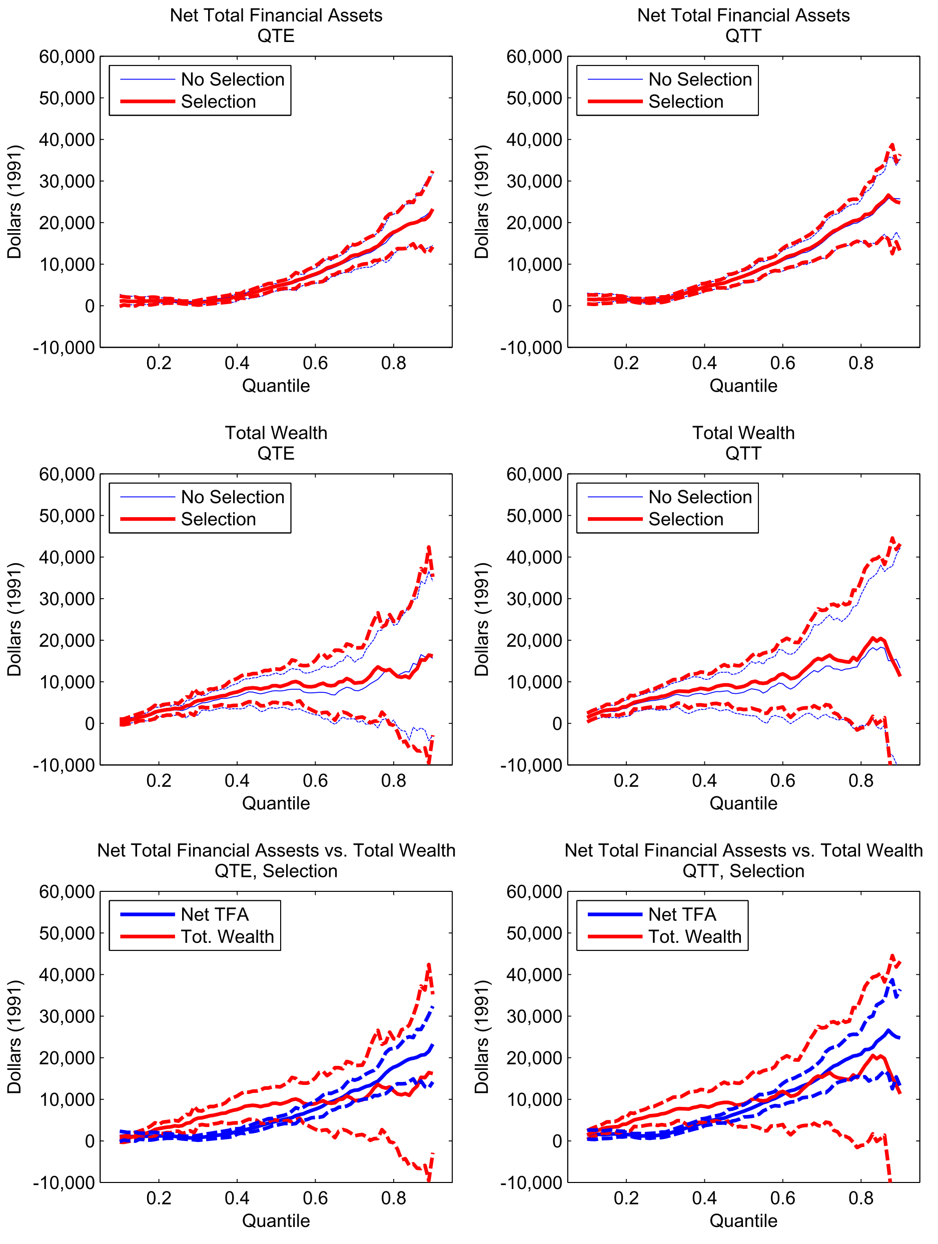


FIGURE S.1.—QTE and QTE-T estimates based on the Indicators specification.

(3) *Given a measurable class $\mathcal{F}$ mapping $\mathcal{W}$ to $\mathbb{R}$ and a random variable $\xi$ taking values in $\mathbb{R}$,*

$$\begin{aligned} &\log \sup_Q N\big(\epsilon \big\| |\xi| F \big\|_{Q,2}, \xi\mathcal{F}, \|\cdot\|_{Q,2}\big) \\ &\quad \leq \log \sup_Q N\big(\epsilon/2 \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\big). \end{aligned}$$

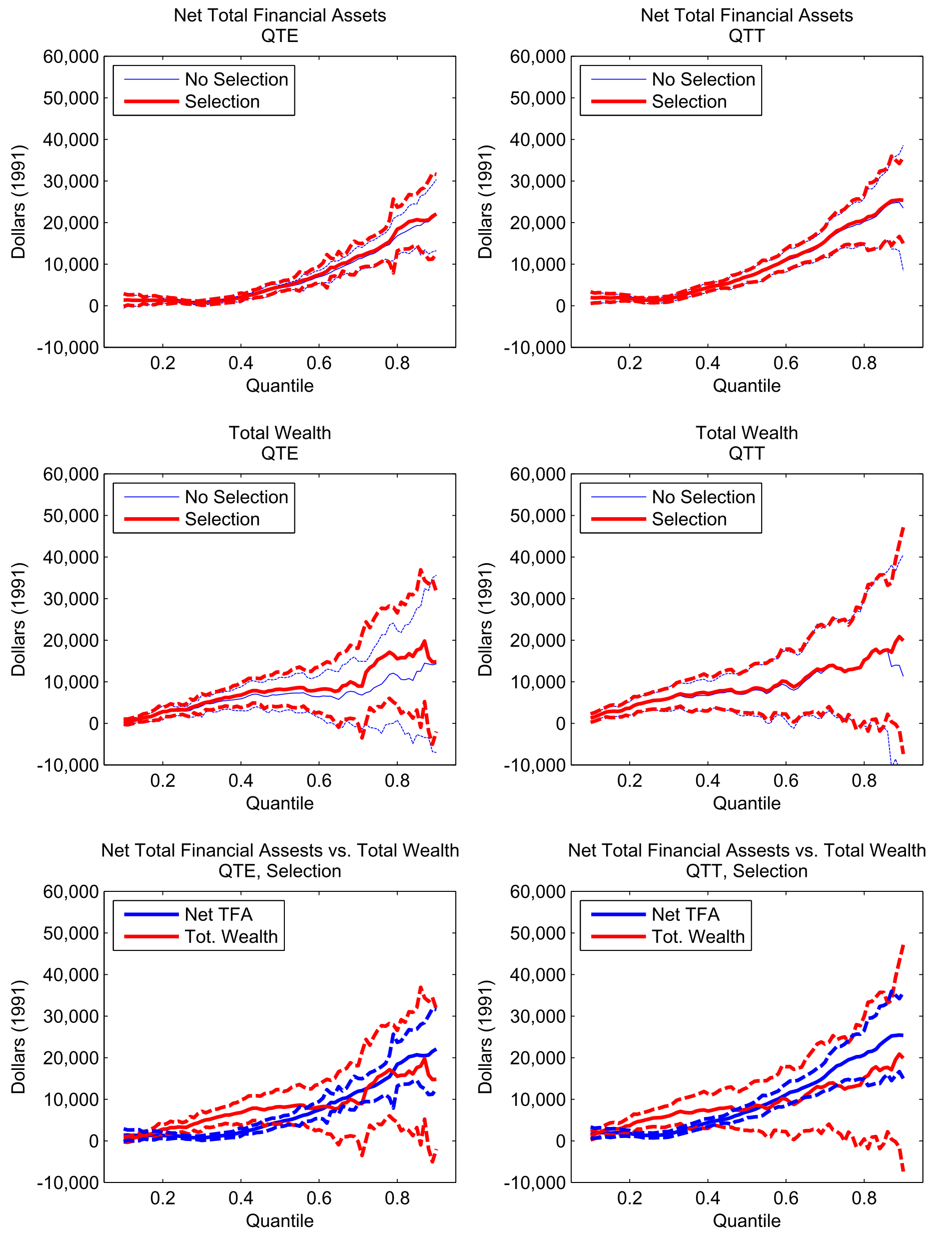


FIGURE S.2.—QTE and QTE-T estimates based on the Quadratic Spline specification.

(4) *Given measurable classes* $\mathcal{F}_j$ *and envelopes* $F_j$, $j = 1, \ldots, k$, *mapping* $\mathcal{W}$ *to* $\mathbb{R}$, *a function* $\phi : \mathbb{R}^k \to \mathbb{R}$ *such that for* $f_j, g_j \in \mathcal{F}_j$, $|\phi(f_1, \ldots, f_k) - \phi(g_1, \ldots, g_k)| \leq \sum_{j=1}^k L_j(x)|f_j(x) - g_j(x)|$, $L_j(x) \geq 0$, *and fixed functions* $\bar{f}_j \in \mathcal{F}_j$, *the class of functions*

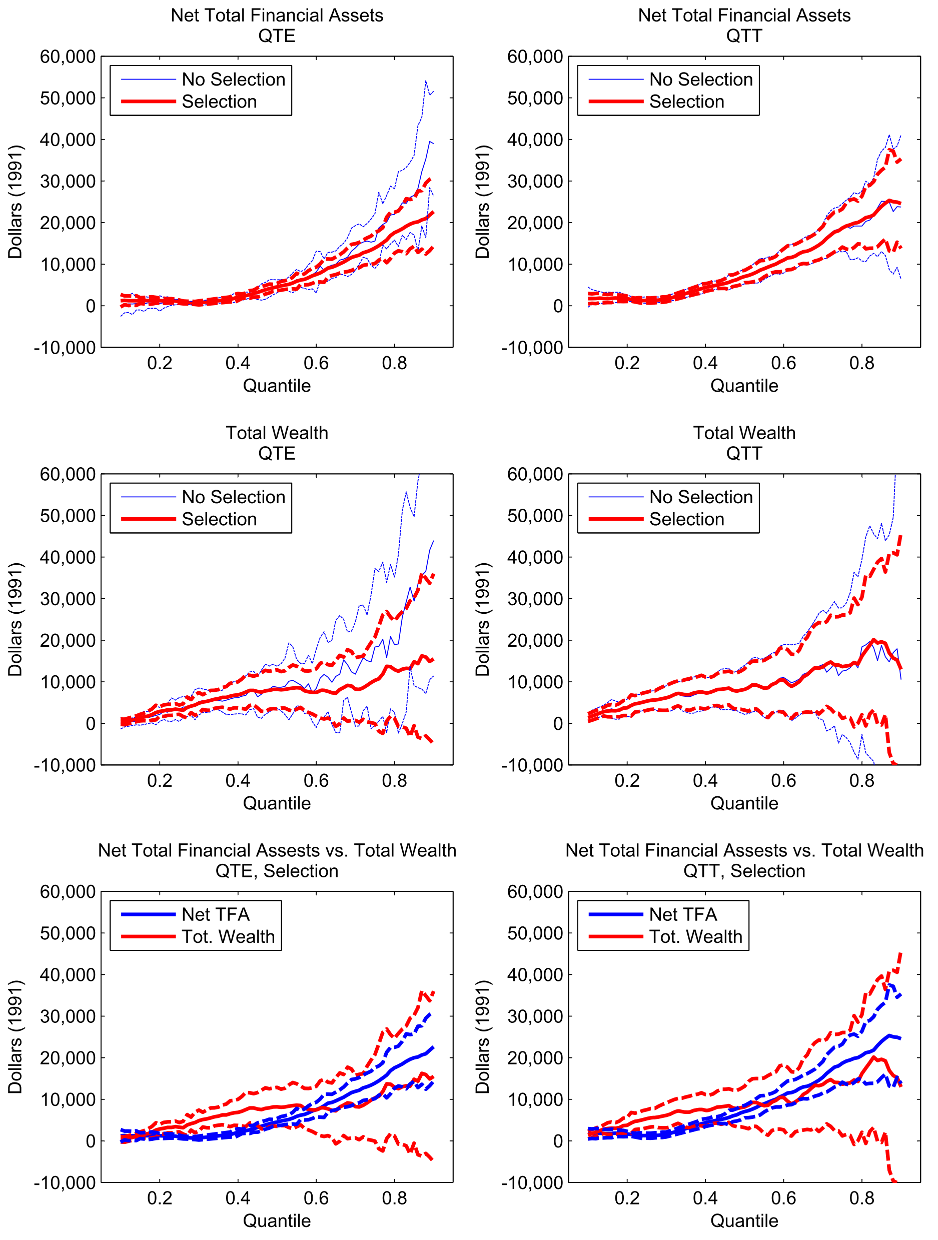


FIGURE S.3.—QTE and QTE-T estimates based on the Quadratic Spline Plus Interactions specification.

$\mathcal{L} = \{\phi(f_1, \dots, f_k) - \phi(\bar{f}_1, \dots, \bar{f}_k) : f_j \in \mathcal{F}_j, j = 1, \dots, k\}$ *satisfies*

$$\log \sup_Q N\left(\epsilon \left\| \sum_{j=1}^{k} L_j F_j \right\|_{Q,2}, \mathcal{L}, \|\cdot\|_{Q,2}\right) \leq \sum_{j=1}^{k} \log \sup_Q N\left(\frac{\epsilon}{k} \|F_j\|_{Q,2}, \mathcal{F}_j, \|\cdot\|_{Q,2}\right).$$

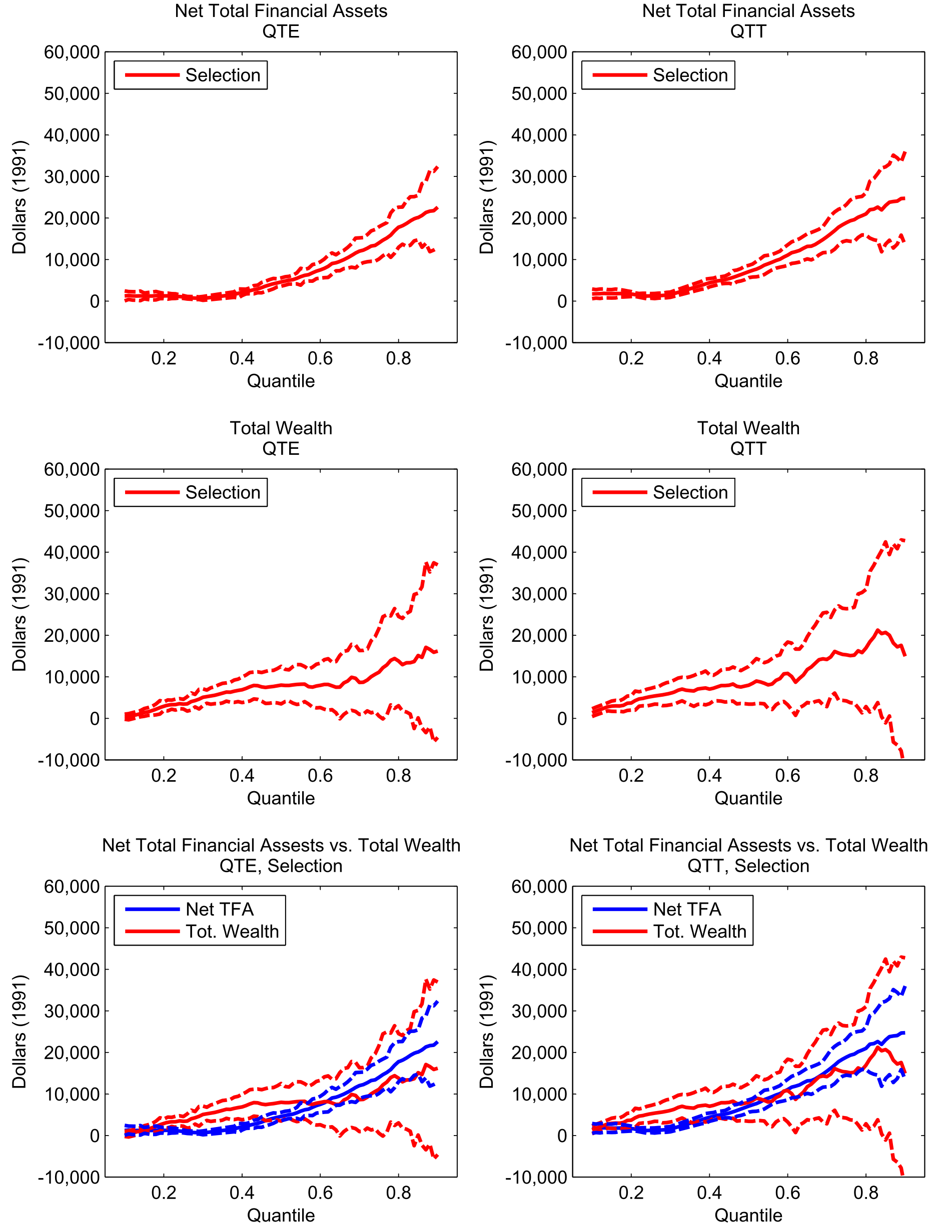


FIGURE S.4.—QTE and QTE-T estimates based on the Quadratic Spline Plus Many Interactions specification.

PROOF: For the proof (1)–(2) see, for example, Andrews (1994), and (3) follows from (2). To show (4), let $f=(f_1,\dots,f_k)$ and $g=(g_1,\dots,g_k)$ where $f_j, g_j \in \mathcal{F}_j$, $j=1,\dots,k$. Then, by the condition on $\phi$, we have

$$\|\phi(f)-\phi(g)\|_{Q,2} \le \left\|\sum_{j=1}^{k} L_j|f_j-g_j|\right\|_{Q,2} \le \sum_{j=1}^{k}\|L_j|f_j-g_j|\|_{Q,2}. \tag{L.1}$$

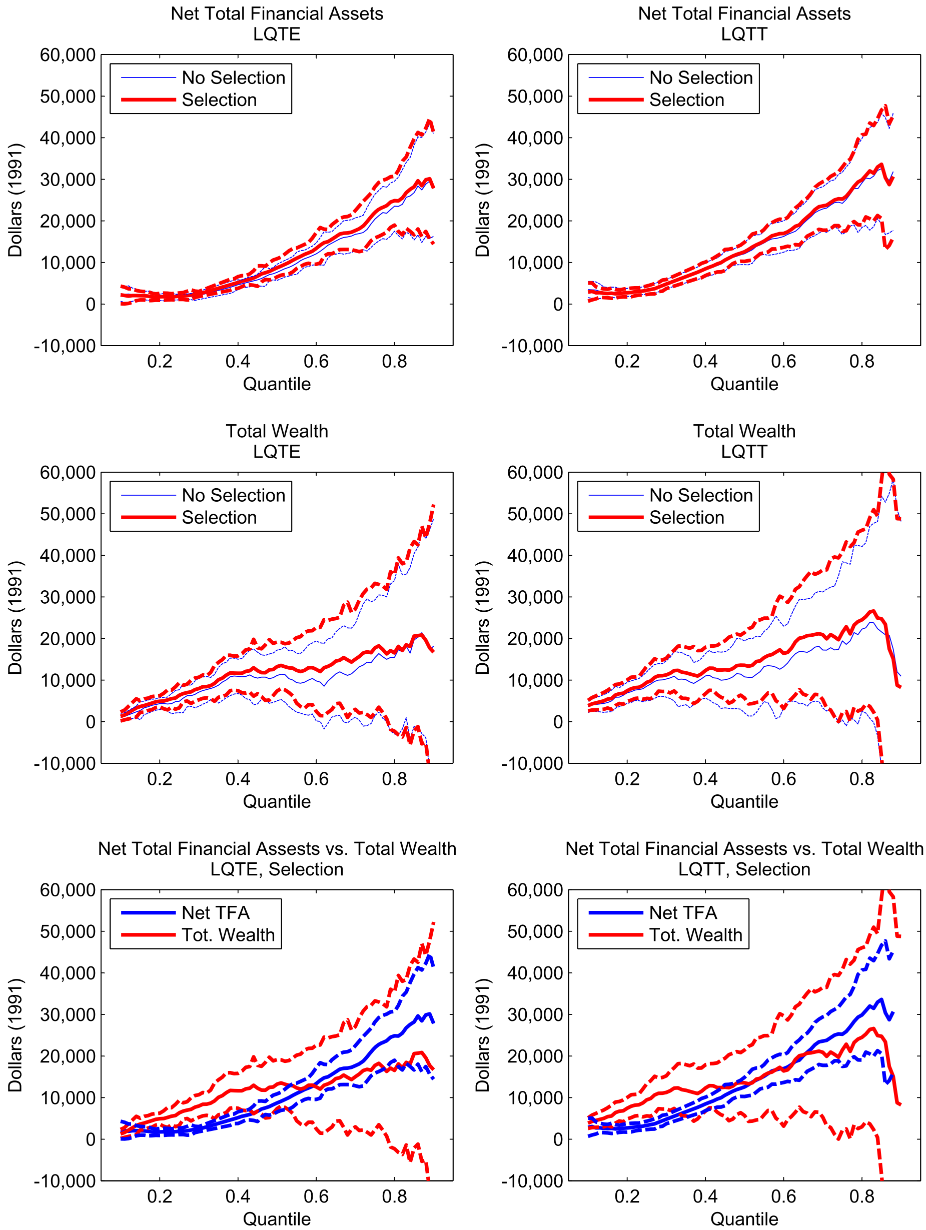


FIGURE S.5.—LQTE and LQTE-T estimates based on the Indicators specification.

Let $\hat{\mathcal{N}}_j$ be a $(\epsilon/k)$-net for $\mathcal{F}_j$ with the measure $\tilde{Q}_j$, where $d\tilde{Q}_j(x) = L_j^2(x)\,dQ(x)$. Then the set $\{\phi(f_1, \dots, f_k) - \phi(\bar{f}_1, \dots, \bar{f}_k) : f_j \in \hat{\mathcal{N}}_j\}$ is an $\epsilon$-net for $\mathcal{L}$ with respect to the measure $Q$ by (L.1). Thus, for any $\epsilon > 0$, we have that

$$\log N\big(\epsilon, \mathcal{L}, \|\cdot\|_{Q,2}\big) \leq \sum_{j=1}^{k} \log N\big(\epsilon/k, \mathcal{F}_j, \|\cdot\|_{\tilde{Q}_j,2}\big).$$

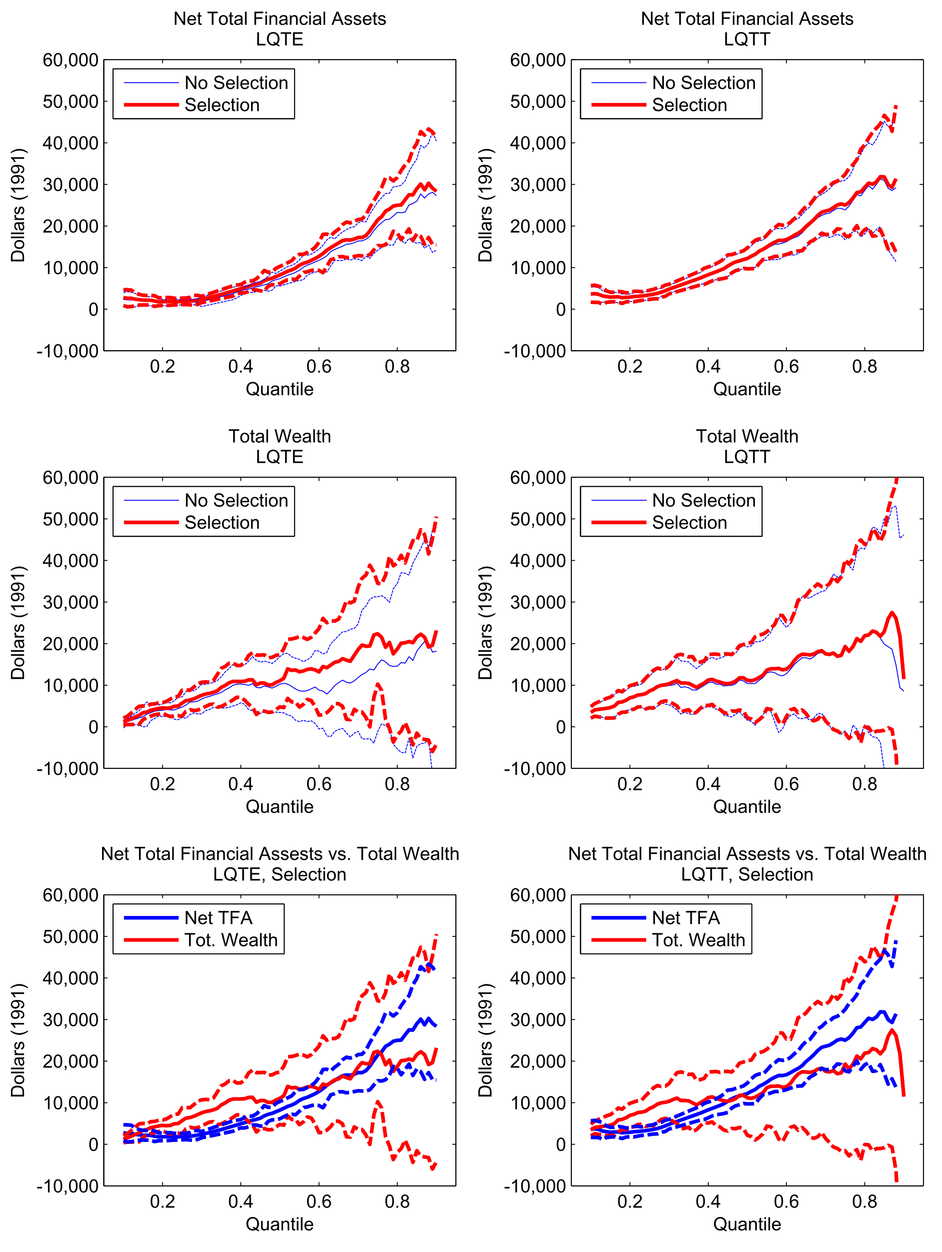


FIGURE S.6.—LQTE and LQTE-T estimates based on the Quadratic Spline specification.

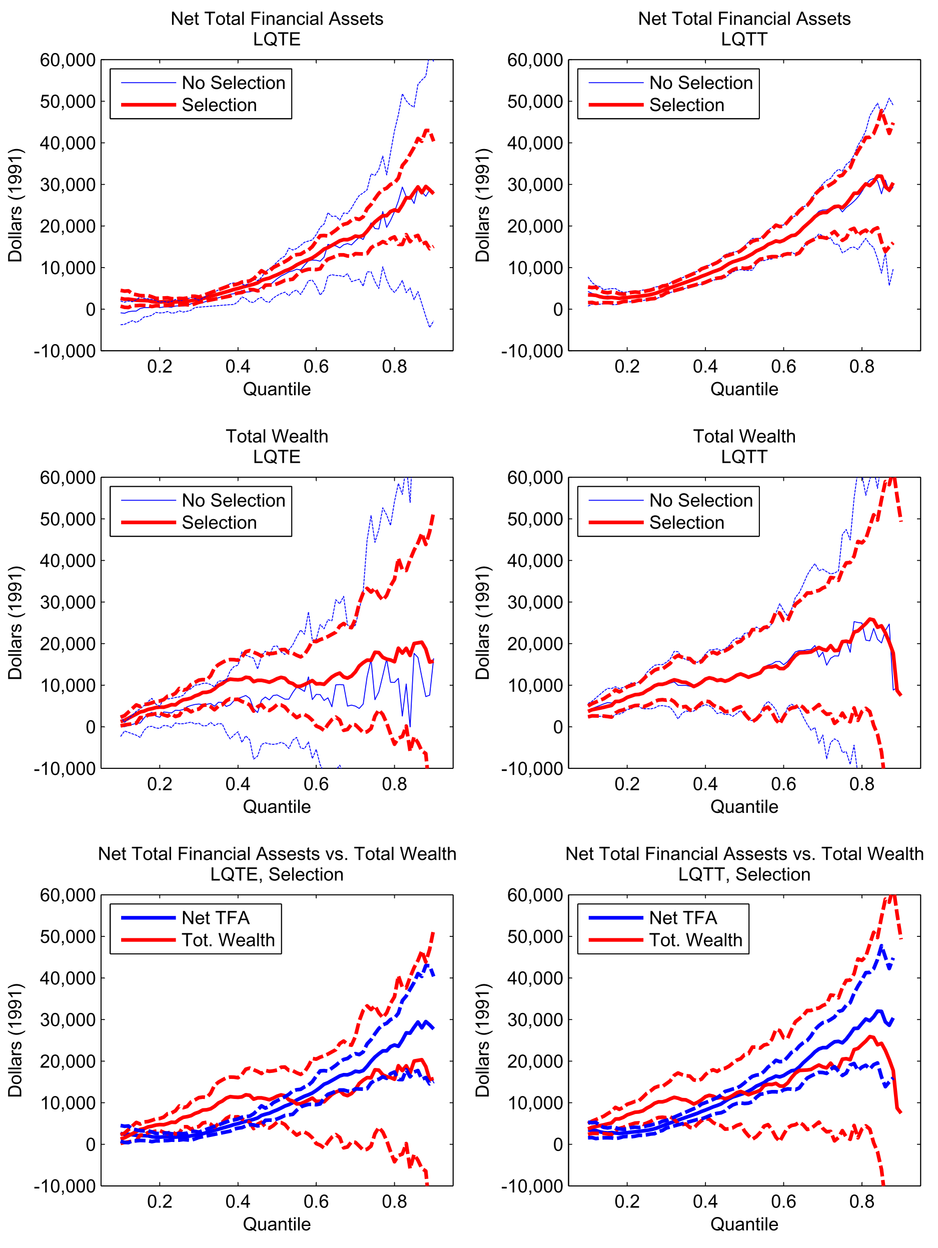


FIGURE S.7.—LQTE and LQTE-T estimates based on the Quadratic Spline Plus Interactions specification.

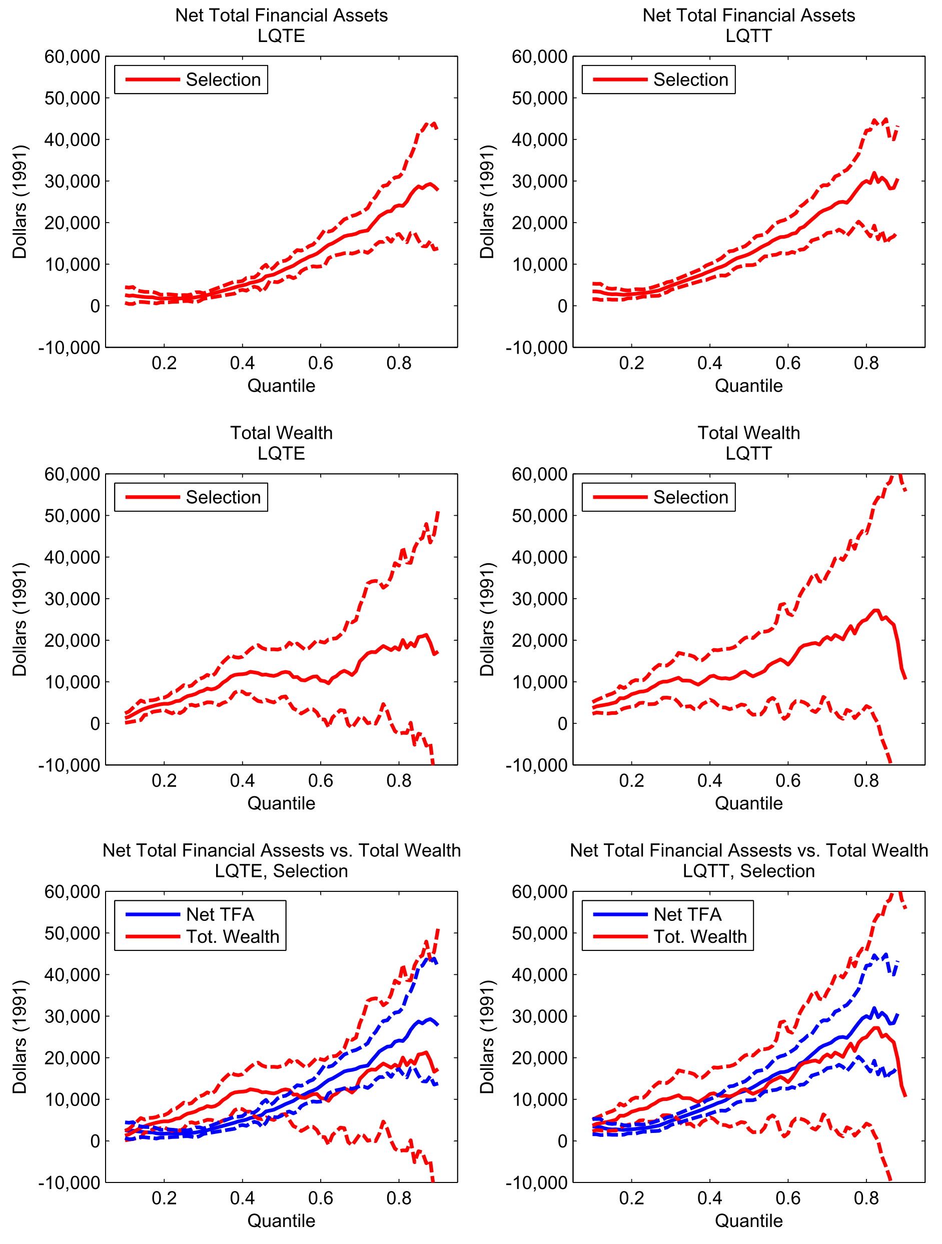


FIGURE S.8.—LQTE and LQTE-T estimates based on the Quadratic Spline Plus Many Interactions specification.

Therefore,

$$
\begin{aligned}
&\log N\left(\epsilon\left\|\sum_{j=1}^{k} L_j F_j\right\|_{Q,2}, \mathcal{L}, \|\cdot\|_{Q,2}\right) \\
&\quad \le \sum_{j=1}^{k} \log N\left(\frac{\epsilon}{k}\left\|\sum_{j=1}^{k} L_j F_j\right\|_{Q,2}, \mathcal{F}_j, \|\cdot\|_{\tilde{Q}_j,2}\right) \\
&\quad \le \sum_{j=1}^{k} \log N\left(\frac{\epsilon}{k}\|L_j F_j\|_{Q,2}, \mathcal{F}_j, \|\cdot\|_{\tilde{Q}_j,2}\right) \\
&\quad = \sum_{j=1}^{k} \log N\left(\frac{\epsilon}{k}\|F_j\|_{\tilde{Q}_j,2}, \mathcal{F}_j, \|\cdot\|_{\tilde{Q}_j,2}\right) \\
&\quad \le \sum_{j=1}^{k} \log \sup_{\tilde{Q}} N\left(\frac{\epsilon}{k}\|F_j\|_{\tilde{Q},2}, \mathcal{F}_j, \|\cdot\|_{\tilde{Q},2}\right),
\end{aligned}
$$

and the result follows since the right-hand side no longer depends on $Q$. *Q.E.D.*

LEMMA L.2—Covering Entropy for Classes Obtained as Conditional Expectations: *Let $\mathcal{F}$ denote a class of measurable functions $f : \mathcal{W} \times \mathcal{Y} \longmapsto \mathbb{R}$ with a measurable envelope $F$. For a given $f \in \mathcal{F}$, let $\bar{f} : \mathcal{W} \longmapsto \mathbb{R}$ be the function $\bar{f}(w) := \int f(w, y)\, d\mu_w(y)$ where $\mu_w$ is a regular conditional probability distribution over $y \in \mathcal{Y}$ conditional on $w \in \mathcal{W}$. Set $\bar{\mathcal{F}} = \{\bar{f} : f \in \mathcal{F}\}$ and let $\bar{F}(w) := \int F(w, y)\, d\mu_w(y)$ be an envelope for $\bar{\mathcal{F}}$. Then, for $r, s \ge 1$,*

$$
\log \sup_{Q} N\big(\epsilon\|\bar{F}\|_{Q,r}, \bar{\mathcal{F}}, \|\cdot\|_{Q,r}\big) \le \log \sup_{\tilde{Q}} N\big((\epsilon/4)^r\|F\|_{\tilde{Q},s}, \mathcal{F}, \|\cdot\|_{\tilde{Q},s}\big),
$$

*where $Q$ belongs to the set of finitely-discrete probability measures over $\mathcal{W}$ such that $0 < \|\bar{F}\|_{Q,r} < \infty$, and $\tilde{Q}$ belongs to the set of finitely-discrete probability measures over $\mathcal{W} \times \mathcal{Y}$ such that $0 < \|F\|_{\tilde{Q},s} < \infty$. In particular, for every $\epsilon > 0$ and any $k \ge 1$,*

$$
\log \sup_{Q} N\big(\epsilon, \bar{\mathcal{F}}, \|\cdot\|_{Q,k}\big) \le \log \sup_{\tilde{Q}} N\big(\epsilon/2, \mathcal{F}, \|\cdot\|_{\tilde{Q},k}\big).
$$

PROOF: The proof generalizes the proof of Lemma A.2 in Ghosal, Sen, and van der Vaart (2000). For $f, g \in \mathcal{F}$ and the corresponding $\bar{f}, \bar{g} \in \bar{\mathcal{F}}$, and any probability measure $Q$ on $\mathcal{W}$, by Jensen's inequality, for any $k \ge 1$,

$$
\begin{aligned}
\mathrm{E}_Q\big[|\bar{f} - \bar{g}|^k\big] &= \mathrm{E}_Q\left[\left|\int (f - g)\, d\mu_w(y)\right|^k\right] \le \mathrm{E}_Q\left[\int |f - g|^k\, d\mu_w(y)\right] \\
&= \mathrm{E}_{\tilde{Q}}\big[|f - g|^k\big],
\end{aligned}
$$

where $d\tilde{Q}(w, y) = dQ(w)\, d\mu_w(y)$. Therefore, for any $\epsilon > 0$,

$$
\sup_{Q} N\big(\epsilon, \bar{\mathcal{F}}, \|\cdot\|_{Q,k}\big) \le \sup_{\tilde{Q}} N\big(\epsilon, \mathcal{F}, \|\cdot\|_{\tilde{Q},k}\big) \le \sup_{\tilde{Q}} N\big(\epsilon/2, \mathcal{F}, \|\cdot\|_{\tilde{Q},k}\big),
$$

where we use Problems 2.5.1–2.5.2 of van der Vaart and Wellner (1996) to replace the supremum over $\bar{Q}$ with the supremum over finitely-discrete probability measures $\tilde{Q}$.

Moreover, $\|\bar{F}\|_{Q,1} = \mathrm{E}_Q[\bar{F}(w)] = \mathrm{E}_Q[\int F(w,y)\,d\mu_w(y)] = \mathrm{E}_{\bar{Q}}[F(w,y)] = \|F\|_{\bar{Q},1}$. Therefore taking $k=1$,

$$\begin{aligned}\sup_Q N\big(\epsilon\|\bar{F}\|_{Q,1},\bar{\mathcal{F}},\|\cdot\|_{Q,1}\big) &\le \sup_{\bar{Q}} N\big(\epsilon\|F\|_{\bar{Q},1},\mathcal{F},\|\cdot\|_{\bar{Q},1}\big)\\ &\le \sup_{\tilde{Q}} N\big((\epsilon/2)\|F\|_{\tilde{Q},1},\mathcal{F},\|\cdot\|_{\tilde{Q},1}\big)\\ &\le \sup_{\tilde{Q}} N\big((\epsilon/2)\|F\|_{\tilde{Q},s},\mathcal{F},\|\cdot\|_{\tilde{Q},s}\big),\end{aligned}$$

where we use Problems 2.5.1–2.5.2 of van der Vaart and Wellner (1996) to replace the supremum over $\bar{Q}$ with the supremum over finitely-discrete probability measures $\tilde{Q}$, and then Problem 2.10.4 of van der Vaart and Wellner (1996) to argue that the last bound is weakly increasing in $s \ge 1$.

Also, by the second part of the proof of Theorem 2.6.7 of van der Vaart and Wellner (1996),

$$\sup_Q N\big(\epsilon\|F\|_{Q,r},\mathcal{F},\|\cdot\|_{Q,r}\big) \le \sup_Q N\big((\epsilon/2)^r\|F\|_{Q,1},\mathcal{F},\|\cdot\|_{Q,1}\big).$$

*Q.E.D.*

COMMENT L.1: Lemma L.2 extends the result in Lemma A.2 in Ghosal, Sen, and van der Vaart (2000) and Lemma 5 in Sherman (1994) which considered integral classes with respect to a fixed measure $\mu$ on $\mathcal{Y}$. In our applications, we need to allow the integration measure to vary with $w$, namely we allow for $\mu_w$ to be a conditional distribution.

## APPENDIX M: PROOFS FOR SECTION 4

### M.1. *Proof of Theorem 4.1*

*Step 0* (Preparation). In the proof $a \lesssim b$ means that $a \le Ab$, where the constant $A$ depends on the constants in Assumptions 4.1 and 4.2 only, but not on $n$ once $n \ge n_0 = \min\{j : \delta_j \le 1/2\}$, and not on $P \in \mathcal{P}_n$. We consider a sequence $P_n$ in $\mathcal{P}_n$, but for simplicity, we write $P = P_n$ throughout the proof, *suppressing* the index $n$. Since the argument is asymptotic, we can assume that $n \ge n_0$ in what follows.

To proceed with the presentation of the proofs, it might be convenient for the reader to have the notation collected in one place. The influence function and low-bias moment functions for $\alpha_V(z)$ for $z \in \mathcal{Z} = \{0,1\}$ are given respectively by

$$\psi^\alpha_{V,z}(W) = \psi^\alpha_{V,z,g_V,m_Z}\big(W,\alpha_V(z)\big),$$
$$\psi^\alpha_{V,z,g,m}(W,\alpha) = \frac{1(Z=z)\big(V-g(z,X)\big)}{m(z,X)} + g(z,X) - \alpha.$$

The influence function and the moment function for $\gamma_V$ are $\psi^\gamma_V(W) = \psi^\gamma_V(W,\gamma_V)$ and $\psi^\gamma_V(W,\gamma) = V - \gamma$. Recall that the estimator of the reduced-form parameters $\alpha_V(z)$ and $\gamma_V$ are solutions $\alpha = \hat{\alpha}_V(z)$ and $\gamma = \hat{\gamma}_V$ to the equations

$$\mathbb{E}_n\big[\psi^\alpha_{V,z,\hat{g}_V,\hat{m}_Z}(W,\alpha)\big] = 0, \quad \mathbb{E}_n\big[\psi^\gamma_V(W,\gamma)\big] = 0,$$

where $\hat{g}_V(z,x) = \Lambda_V(f(z,x)'\bar{\beta}_V)$, $\hat{m}_Z(1,x) = \Lambda_Z(f(x)'\bar{\beta}_Z)$, $\hat{m}_Z(0,x) = 1 - \hat{m}_Z(1,x)$, and $\bar{\beta}_V$ and $\bar{\beta}_Z$ are estimators as in Assumption 4.2. For each variable $V \in \mathcal{V}_u$,

$$\mathcal{V}_u = (V_{uj})_{j=1}^5 = \big(Y_u, \mathbf{1}_0(D)Y_u, \mathbf{1}_0(D), \mathbf{1}_1(D)Y_u, \mathbf{1}_1(D)\big),$$

we obtain the estimator $\hat{\rho}_u = (\{\hat{\alpha}_V(0), \hat{\alpha}_V(1), \hat{\gamma}_V\})_{V\in\mathcal{V}_u}$ of $\rho_u := (\{\alpha_V(0), \alpha_V(1), \gamma_V\})_{V\in\mathcal{V}_u}$. The estimator and the estimand are vectors in $\mathbb{R}^{d_\rho}$ with a fixed finite dimension. We stack these vectors into the processes $\hat{\rho} = (\hat{\rho}_u)_{u\in\mathcal{U}}$ and $\rho = (\rho_u)_{u\in\mathcal{U}}$.

*Step 1* (Linearization). In this step we establish the first claim, namely that

$$\sqrt{n}(\hat{\rho} - \rho) = Z_{n,P} + o_P(1) \quad \text{in} \quad \mathbb{D} = \ell^\infty(\mathcal{U})^{d_\rho}, \tag{M.1}$$

where $Z_{n,P} = (\mathbb{G}_n\psi_u^\rho)_{u\in\mathcal{U}}$ and $\psi_u^\rho = (\{\psi_{V,0}^\alpha, \psi_{V,1}^\alpha, \psi_V^\gamma\})_{V\in\mathcal{V}_u}$. The components $(\sqrt{n}(\hat{\gamma}_{V_{uj}} - \gamma_{V_{uj}}))_{u\in\mathcal{U}}$ of $\sqrt{n}(\hat{\rho} - \rho)$ trivially have the linear representation (with no error) for each $j \in \mathcal{J}$. We only need to establish the claim for the empirical process $(\sqrt{n}(\hat{\alpha}_{V_{uj}}(z) - \alpha_{V_{uj}}(z)))_{u\in\mathcal{U}}$ for $z \in \{0,1\}$ and each $j \in \mathcal{J}$, which we do in the steps below.

(a) We make some preliminary observations. For $t = (t_1, t_2, t_3, t_4) \in \mathbb{R}^2 \times (0,1)^2$, $v \in \mathbb{R}$, and $(z, \bar{z}) \in \{0,1\}^2$, we define the function $(v, z, \bar{z}, t) \longmapsto \varphi(v, z, \bar{z}, t)$ via

$$\varphi(v, z, 1, t) = \frac{1(z=1)(v - t_2)}{t_4} + t_2,$$
$$\varphi(v, z, 0, t) = \frac{1(z=0)(v - t_1)}{t_3} + t_1.$$

The derivatives of this function with respect to $t$ obey, for all $k = (k_j)_{j=1}^4 \in \mathbb{N}^4 : 0 \le |k| \le 3$,

$$\begin{aligned} &\big|\partial_t^k \varphi(v, z, \bar{z}, t)\big| \le L, \\ &\quad \forall (v, \bar{z}, z, t) : |v| \le C, |t_1|, |t_2| \le C, c'/2 \le |t_3|, |t_4| \le 1 - c'/2, \end{aligned} \tag{M.2}$$

where $L$ depends only on $c'$ and $C$, $|k| = \sum_{j=1}^4 k_j$, and $\partial_t^k := \partial_{t_1}^{k_1}\partial_{t_2}^{k_2}\partial_{t_3}^{k_3}\partial_{t_4}^{k_4}$.

(b) Let

$$\hat{h}_V(X) := \big(\hat{g}_V(0,X), \hat{g}_V(1,X), 1 - \hat{m}_Z(1,X), \hat{m}_Z(1,X)\big)',$$
$$h_V(X) := \big(g_V(0,X), g_V(1,X), 1 - m_Z(1,X), m_Z(1,X)\big)',$$
$$f_{\hat{h}_V,V,z}(W) := \varphi\big(V, Z, z, \hat{h}_V(X)\big),$$
$$f_{h_V,V,z}(W) := \varphi\big(V, Z, z, h_V(X)\big).$$

We observe that with probability no less than $1 - \Delta_n$,

$$\hat{g}_V(0,\cdot) \in \mathcal{G}_V(0), \quad \hat{g}_V(1,\cdot) \in \mathcal{G}_V(1),$$
$$\hat{m}_Z(1,\cdot) \in \mathcal{M}(1), \quad \hat{m}_Z(0,\cdot) \in \mathcal{M}(0) = 1 - \mathcal{M}(1),$$

where

$$\mathcal{G}_V(z) := \left\{ \begin{array}{l} x \longmapsto \Lambda_V\big(f(z,x)'\beta\big) : \|\beta\|_0 \le sC \\ \big\|\Lambda_V\big(f(z,X)'\beta\big) - g_V(z,X)\big\|_{P,2} \lesssim \delta_n n^{-1/4} \\ \big\|\Lambda_V\big(f(z,X)'\beta\big) - g_V(z,X)\big\|_{P,\infty} \lesssim \epsilon_n \end{array} \right\},$$

$$\mathcal{M}(1) := \left\{ \begin{array}{l} x \longmapsto \Lambda_Z\big(f(x)'\beta\big) : \|\beta\|_0 \leq sC \\ \big\|\Lambda_Z\big(f(X)'\beta\big) - m_Z(1, X)\big\|_{P,2} \lesssim \delta_n n^{-1/4} \\ \big\|\Lambda_Z\big(f(X)'\beta\big) - m_Z(1, X)\big\|_{P,\infty} \lesssim \epsilon_n \end{array} \right\}.$$

To see this, note that under Assumption 4.2 for all $n \geq \min\{j : \delta_j \leq 1/2\}$,

$$\begin{aligned}
&\big\|\Lambda_Z\big(f(X)'\beta\big) - m_Z(1, X)\big\|_{P,2} \\
&\quad \leq \big\|\Lambda_Z\big(f(X)'\beta\big) - \Lambda_Z\big(f(X)'\beta_Z\big)\big\|_{P,2} + \big\|r_Z(X)\big\|_{P,2} \\
&\quad \lesssim \|\partial\Lambda_Z\|_\infty \big\|f(X)'(\beta - \beta_Z)\big\|_{P,2} + \big\|r_Z(X)\big\|_{P,2} \\
&\quad \lesssim \|\partial\Lambda_Z\|_\infty \big\|f(X)'(\beta - \beta_Z)\big\|_{\mathbb{P}_n,2} + \big\|r_Z(X)\big\|_{P,2} \lesssim \delta_n n^{-1/4}, \\
&\big\|\Lambda_Z\big(f(X)'\beta\big) - m_Z(1, X)\big\|_{P,\infty} \\
&\quad \leq \big\|\Lambda_Z\big(f(X)'\beta\big) - \Lambda_Z\big(f(X)'\beta_Z\big)\big\|_{P,\infty} + \big\|r_Z(X)\big\|_{P,\infty} \\
&\quad \leq \|\partial\Lambda_Z\|_\infty \big\|f(X)'(\beta - \beta_Z)\big\|_{P,\infty} + \big\|r_Z(X)\big\|_{P,\infty} \\
&\quad \lesssim K_n \|\beta - \beta_Z\|_1 + \epsilon_n \leq 2\epsilon_n,
\end{aligned}$$

for $\beta = \bar{\beta}_Z$, with evaluation after computing the norms, and for $\|\partial\Lambda\|_\infty$ denoting $\sup_{l \in \mathbb{R}} |\partial\Lambda(l)|$ here and below. Similarly, under Assumption 4.2,

$$\begin{aligned}
&\big\|\Lambda_V\big(f(Z, X)'\beta\big) - g_V(Z, X)\big\|_{P,2} \\
&\quad \lesssim \|\partial\Lambda_V\|_\infty \big\|f(Z, X)'(\beta - \beta_V)\big\|_{\mathbb{P}_n,2} + \big\|r_V(Z, X)\big\|_{P,2} \lesssim \delta_n n^{-1/4}, \\
&\big\|\Lambda_V\big(f(Z, X)'\beta\big) - g_V(Z, X)\big\|_{P,\infty} \lesssim K_n \|\beta - \beta_V\|_1 + \epsilon_n \leq 2\epsilon_n,
\end{aligned}$$

for $\beta = \bar{\beta}_V$, with evaluation after computing the norms, and noting that, for any $\beta$,

$$\begin{aligned}
&\big\|\Lambda_V\big(f(0, X)'\beta\big) - g_V(0, X)\big\|_{P,2} \vee \big\|\Lambda_V\big(f(1, X)'\beta\big) - g_V(1, X)\big\|_{P,2} \\
&\quad \lesssim \big\|\Lambda_V\big(f(Z, X)'\beta\big) - g_V(Z, X)\big\|_{P,2}
\end{aligned}$$

under condition (iii) of Assumption 4.1, and

$$\begin{aligned}
&\big\|\Lambda_V\big(f(0, X)'\beta\big) - g_V(0, X)\big\|_{P,\infty} \vee \big\|\Lambda_V\big(f(1, X)'\beta\big) - g_V(1, X)\big\|_{P,\infty} \\
&\quad \leq \big\|\Lambda_V\big(f(Z, X)'\beta\big) - g_V(Z, X)\big\|_{P,\infty}
\end{aligned}$$

under condition (iii) of Assumption 4.1.

Hence with probability at least $1 - \Delta_n$,

$$\begin{aligned}
\hat{h}_V \in \mathcal{H}_{V,n} := \big\{h = &\big(\bar{g}(0, \cdot), \bar{g}(1, \cdot), \bar{m}_Z(0, \cdot), \bar{m}_Z(1, \cdot)\big) \\
&\in \mathcal{G}_V(0) \times \mathcal{G}_V(1) \times \mathcal{M}(0) \times \mathcal{M}(1)\big\}.
\end{aligned}$$

(c) We have that

$$\alpha_V(z) = \mathrm{E}_P[f_{h_V,V,z}] \quad \text{and} \quad \hat{\alpha}(z) = \mathbb{E}_n[f_{\hat{h}_V,V,z}],$$

so that

$$\begin{aligned}&\sqrt{n}\big(\hat{\alpha}_V(z)-\alpha_V(z)\big)\\&\quad=\underbrace{\mathbb{G}_n[f_{h_V,V,z}]}_{\mathrm{I}_V(z)}+\underbrace{\mathbb{G}_n[f_{h,V,z}-f_{h_V,V,z}]}_{\mathrm{II}_V(z)}+\underbrace{\sqrt{n}P[f_{h,V,z}-f_{h_V,V,z}]}_{\mathrm{III}_V(z)},\end{aligned}$$

with $h$ evaluated at $h=\hat{h}_V$.

(d) Note that for

$$\begin{aligned}&\Delta_{V,i}:=(\Delta_{1V,i},\Delta_{2V,i},\Delta_{3V,i},\Delta_{4V,i})=h(X_i)-h_V(X_i),\\&\Delta_{V,i}^k:=\Delta_{1V,i}^{k_1}\Delta_{2V,i}^{k_2}\Delta_{3V,i}^{k_3}\Delta_{4V,i}^{k_4},\\&\mathrm{III}_V(z)=\sqrt{n}\sum_{|k|=1}P\big[\partial_t^k\varphi\big(V_i,Z_i,z,h_V(X_i)\big)\Delta_{V,i}^k\big]\\&\qquad+\sqrt{n}\sum_{|k|=2}2^{-1}P\big[\partial_t^k\varphi\big(V_i,Z_i,z,h_V(X_i)\big)\Delta_{V,i}^k\big]\\&\qquad+\sqrt{n}\sum_{|k|=3}6^{-1}\int_0^1P\big[\partial_t^k\varphi\big(V_i,Z_i,z,h_V(X_i)+\lambda\Delta_{V,i}\big)\Delta_{V,i}^k\big]\,d\lambda\\&\qquad=:\mathrm{III}_V^a(z)+\mathrm{III}_V^b(z)+\mathrm{III}_V^c(z),\end{aligned}$$

with $h$ evaluated at $h=\hat{h}$ after computing the expectations under $P$.

By the law of iterated expectations and the orthogonality property of the moment condition for $\alpha_V$,

$$\begin{aligned}&\mathrm{E}_P\big[\partial_t^k\varphi\big(V_i,Z_i,z,h_V(X_i)\big)|X_i\big]=0\quad\forall k\in\mathbb{N}^4:|k|=1\\&\quad\Longrightarrow\quad\mathrm{III}_V^a(z)=0.\end{aligned}$$

Moreover, uniformly for any $h\in\mathcal{H}_{V,n}$, in view of properties noted in Steps (a) and (b),

$$\begin{aligned}&\big|\mathrm{III}_V^b(z)\big|\lesssim\sqrt{n}\|h-h_V\|_{P,2}^2\lesssim\sqrt{n}\big(\delta_n n^{-1/4}\big)^2\le\delta_n^2,\\&\big|\mathrm{III}_V^c(z)\big|\lesssim\sqrt{n}\|h-h_V\|_{P,2}^2\|h-h_V\|_{P,\infty}\lesssim\sqrt{n}\big(\delta_n n^{-1/4}\big)^2\epsilon_n\le\delta_n^2\epsilon_n.\end{aligned}$$

Since $\hat{h}_V\in\mathcal{H}_{V,n}$ for all $V\in\mathcal{V}=\{V_{uj}:u\in\mathcal{U},j\in\mathcal{J}\}$ with probability $1-\Delta_n$, for $n\ge n_0$,

$$\mathrm{P}_P\big(\big|\mathrm{III}_V(z)\big|\lesssim\delta_n^2,\forall z\in\{0,1\},\forall V\in\mathcal{V}\big)\ge1-\Delta_n.$$

(e) Furthermore, with probability $1-\Delta_n$

$$\sup_{V\in\mathcal{V}}\max_{z\in\{0,1\}}\big|\mathrm{II}_V(z)\big|\le\sup_{h\in\mathcal{H}_{V,n},z\in\{0,1\},V\in\mathcal{V}}\big|\mathbb{G}_n[f_{h,V,z}]-\mathbb{G}_n[f_{h_V,V,z}]\big|.$$

The classes of functions

$$\mathcal{V}:=\{V_{uj}:u\in\mathcal{U},j\in\mathcal{J}\}\quad\text{and}\quad\mathcal{V}^*:=\big\{g_{V_{uj}}(Z,X):u\in\mathcal{U},j\in\mathcal{J}\big\},\tag{M.3}$$

viewed as maps from the sample space $\mathcal{W}$ to the real line, are bounded by a constant envelope and obey $\log\sup_Q N(\epsilon, \mathcal{V}, \|\cdot\|_{Q,2}) \lesssim \log(\mathrm{e}/\epsilon) \vee 0$, which holds by Assumption 4.1(ii), and $\log\sup_Q N(\epsilon, \mathcal{V}^*, \|\cdot\|_{Q,2}) \lesssim \log(\mathrm{e}/\epsilon) \vee 0$, which holds by Assumption 4.1(ii) and Lemma L.2. The uniform covering entropy of the function sets

$$\mathcal{B} = \{1(Z = z) : z \in \{0, 1\}\} \quad \text{and} \quad \mathcal{M}^* = \{m_Z(z, X) : z \in \{0, 1\}\}$$

are trivially bounded by $\log(\mathrm{e}/\epsilon) \vee 0$.

The class of functions

$$\mathcal{G} := \{\mathcal{G}_V(z) : V \in \mathcal{V}, z \in \{0, 1\}\}$$

has a constant envelope and is a subset of

$$\begin{aligned} &\{(x, z) \longmapsto \Lambda(f(z, x)'\beta) : \\ &\quad \|\beta\|_0 \le sC, \Lambda \in \mathcal{L} = \{\mathrm{Id}, \Phi, 1 - \Phi, \Lambda_0, 1 - \Lambda_0\}\}, \end{aligned}$$

which is a union of five sets of the form

$$\{(x, z) \longmapsto \Lambda(f(z, x)'\beta) : \|\beta\|_0 \le sC\}$$

with $\Lambda \in \mathcal{L}$ a fixed monotone function for each of the five sets; each of these sets are the unions of at most $\binom{2p}{Cs}$ VC-subgraph classes of functions with VC indices bounded by $C$'s. Note that a fixed monotone transformation $\Lambda$ preserves the VC-subgraph property (van der Vaart and Wellner (1996, Lemma 2.6.18)). Therefore,

$$\log\sup_Q N(\epsilon, \mathcal{G}, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(\mathrm{e}/\epsilon)) \vee 0.$$

Similarly, the class of functions $\mathcal{M} = (\mathcal{M}(1) \cup (1 - \mathcal{M}(1)))$ has a constant envelope, is a union of at most five sets, which are themselves the unions of at most $\binom{p}{Cs}$ VC-subgraph classes of functions with VC indices bounded by $C$'s since a fixed monotone transformation $\Lambda$ preserves the VC-subgraph property. Therefore, $\log\sup_Q N(\epsilon, \mathcal{M}, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(\mathrm{e}/\epsilon)) \vee 0$.

Finally, the set of functions

$$\mathcal{J}_n = \{f_{h,V,z} - f_{h_V,V,z} : z \in \{0, 1\}, V \in \mathcal{V}, h \in \mathcal{H}_{V,n}\}$$

is a Lipschitz transform of function sets $\mathcal{V}$, $\mathcal{V}^*$, $\mathcal{B}$, $\mathcal{M}^*$, $\mathcal{G}$, and $\mathcal{M}$, with bounded Lipschitz coefficients and with a constant envelope. Therefore,

$$\log\sup_Q N(\epsilon, \mathcal{J}_n, \|\cdot\|_{Q,2}) \lesssim (s \log p + s \log(\mathrm{e}/\epsilon)) \vee 0.$$

Applying Lemma C.1 with $\sigma_n = C'\delta_n n^{-1/4}$ and the envelope $J_n = C'$, with probability $1 - \Delta_n$ for some constant $K > e$,

$$\begin{aligned} &\sup_{V \in \mathcal{V}} \max_{z \in \{0,1\}} |\mathrm{II}_V(z)| \\ &\quad \le \sup_{f \in \mathcal{J}_n} |\mathbb{G}_n(f)| \end{aligned}$$

$$
\begin{aligned}
&\lesssim \left(\sqrt{s\sigma_n^2 \log\left(p \vee K \vee \sigma_n^{-1}\right)} + \frac{s}{\sqrt{n}} \log\left(p \vee K \vee \sigma_n^{-1}\right)\right) \\
&\lesssim \left(\sqrt{s\delta_n^2 n^{-1/2} \log(p \vee n)} + \sqrt{s^2 n^{-1} \log^2(p \vee n)}\right) \\
&\lesssim \left(\delta_n \delta_n^{1/4} + \delta_n^{1/2}\right) \lesssim \delta_n^{1/2}.
\end{aligned}
$$

Here we have used some simple calculations, exploiting the boundedness condition in Assumptions 4.1 and 4.2, to deduce that

$$
\sup_{f \in \mathcal{J}_n} \|f\|_{P,2} \lesssim \sup_{h \in \mathcal{H}_{V,n}, V \in \mathcal{V}} \|h - h_V\|_{P,2} \lesssim \delta_n n^{-1/4} \lesssim \sigma_n \leq \|J_n\|_{P,2},
$$

by definition of the set $\mathcal{H}_{V,n}$, so that we can use Lemma C.1. We also note that $\log(1/\delta_n) \lesssim \log(n)$ by the assumption on $\delta_n$ and that $s^2 \log^2(p \vee n) \log^2(n)/n \leq \delta_n$ by Assumption 4.2(i).

(f) The claim of Step 1 follows by collecting Steps (a)–(e).

*Step 2* (Uniform Donskerness). Here we claim that Assumption 4.1 implies that the set of vectors of functions $(\psi_u^\rho)_{u \in \mathcal{U}}$ is $P$-Donsker uniformly in $\mathcal{P}$, namely that

$$
Z_{n,P} \rightsquigarrow Z_P \quad \text{in} \quad \mathbb{D} = \ell^\infty(\mathcal{U})^{d_\rho}, \text{ uniformly in } P \in \mathcal{P},
$$

where $Z_{n,P} = (\mathbb{G}_n \psi_u^\rho)_{u \in \mathcal{U}}$ and $Z_P = (\mathbb{G}_P \psi_u^\rho)_{u \in \mathcal{U}}$. Moreover, $Z_P$ has bounded, uniformly continuous paths uniformly in $P \in \mathcal{P}$:

$$
\sup_{P \in \mathcal{P}} \mathrm{E}_P \sup_{u \in \mathcal{U}} \left\| Z_P(u) \right\| < \infty, \quad \lim_{\varepsilon \searrow 0} \sup_{P \in \mathcal{P}} \mathrm{E}_P \sup_{d_{\mathcal{U}}(u, \tilde{u}) \leq \varepsilon} \left\| Z_P(u) - Z_P(\tilde{u}) \right\| = 0.
$$

To verify these claims we shall invoke Theorem B.1

To demonstrate the claim, it will suffice to consider the set of $\mathbb{R}$-valued functions $\Psi = (\psi_{uk} : u \in \mathcal{U}, k \in [d_\rho])$. Further, we notice that $\mathbb{G}_n \psi_{V,z}^\alpha = \mathbb{G}_n f$, for $f \in \mathcal{F}_z$,

$$
\mathcal{F}_z = \left\{ \frac{1\{Z = z\}\left(V - g_V(z, X)\right)}{m_Z(z, X)} + g_V(z, X), V \in \mathcal{V} \right\}, \quad z = 0, 1,
$$

and that $\mathbb{G}_n \psi_V^\gamma = \mathbb{G}_n f$, for $f = V \in \mathcal{V}$. Hence $\mathbb{G}_n(\psi_{uk}) = \mathbb{G}_n(f)$ for $f \in \mathcal{F}_P = \mathcal{F}_0 \cup \mathcal{F}_1 \cup \mathcal{V}$. We thus need to check that the conditions of Theorem B.1 apply to $\mathcal{F}_P$ uniformly in $P \in \mathcal{P}$.

Observe that $\mathcal{F}_z$ is formed as a uniform Lipschitz transform of the function sets $\mathcal{B}$, $\mathcal{V}$, $\mathcal{V}^*$, and $\mathcal{M}^*$ defined in Step 1(e), where the validity of the Lipschitz property relies on Assumption 4.1(iii) (to keep the denominator away from zero) and on the boundedness conditions in Assumption 4.1(iii) and Assumption 4.2(iii). The function sets $\mathcal{B}$, $\mathcal{V}$, $\mathcal{V}^*$, and $\mathcal{M}^*$ are uniformly bounded classes that have uniform covering entropy bounded by $\log(\mathrm{e}/\epsilon) \vee 0$ up to a multiplicative constant, and so $\mathcal{F}_z$, which is uniformly bounded under Assumption 4.1, the uniform covering entropy bounded by $\log(\mathrm{e}/\epsilon) \vee 0$ up to a multiplicative constant (e.g., van der Vaart and Wellner (1996)). Since $\mathcal{F}_P$ is uniformly bounded and is a finite union of function sets with the uniform entropies obeying the said properties, it also follows that $\mathcal{F}_P$ has this property; namely,

$$
\sup_{P \in \mathcal{P}} \sup_{Q} \log N\left(\epsilon, \mathcal{F}_P, \|\cdot\|_{Q,2}\right) \lesssim \log(\mathrm{e}/\epsilon) \vee 0.
$$

Since $\int_0^\infty \sqrt{\log(\mathrm{e}/\epsilon) \vee 0}\, d\epsilon = \mathrm{e}\sqrt{\pi}/2 < \infty$ and $\mathcal{F}_P$ is uniformly bounded, the first condition in (B.1) and the entropy condition (B.2) in Theorem B.1 hold.

We demonstrate the second condition in (B.1). Consider a sequence of positive constants $\epsilon$ approaching zero, and note that

$$\sup_{d_{\mathcal{U}}(u,\tilde{u})\le\epsilon} \max_{k\le d_\rho} \|\psi_{uk} - \psi_{\tilde{u}k}\|_{P,2} \lesssim \sup_{d_{\mathcal{U}}(u,\tilde{u})\le\epsilon} \|f_u - f_{\tilde{u}}\|_{P,2},$$

where $f_u$ and $f_{\tilde{u}}$ must be of the form

$$\frac{1\{Z=z\}\bigl(U_u - g_{U_u}(z,X)\bigr)}{m_Z(z,X)} + g_{U_u}(z,X),$$
$$\frac{1\{Z=z\}\bigl(U_{\tilde{u}} - g_{U_{\tilde{u}}}(z,X)\bigr)}{m_Z(z,X)} + g_{U_{\tilde{u}}}(z,X),$$

with $(U_u, U_{\tilde{u}})$ equal to either $(Y_u, Y_{\tilde{u}})$ or $(1_d(D)Y_u, 1_d(D)Y_{\tilde{u}})$, for $d=0$ or 1, and $z=0$ or 1. Then

$$\sup_{P\in\mathcal{P}} \|f_u - f_{\tilde{u}}\|_{P,2} \lesssim \sup_{P\in\mathcal{P}} \|Y_u - Y_{\tilde{u}}\|_{P,2} \to 0,$$

as $d_{\mathcal{U}}(u,\tilde{u}) \to 0$ by Assumption 4.1(ii). Indeed, $\sup_{P\in\mathcal{P}} \|f_u - f_{\tilde{u}}\|_{P,2} \lesssim \sup_{P\in\mathcal{P}} \|Y_u - Y_{\tilde{u}}\|_{P,2}$ follows from a sequence of inequalities holding uniformly in $P\in\mathcal{P}$: (1)

$$\|f_u - f_{\tilde{u}}\|_{P,2} \lesssim \|U_u - U_{\tilde{u}}\|_{P,2} + \bigl\|g_{U_u}(z,X) - g_{U_{\tilde{u}}}(z,X)\bigr\|_{P,2},$$

which we deduce using the triangle inequality and the fact that $m_Z(z,X)$ is bounded away from zero, (2) $\|U_u - U_{\tilde{u}}\|_{P,2} \le \|Y_u - Y_{\tilde{u}}\|_{P,2}$, which we deduce using the Holder inequality, and (3)

$$\bigl\|g_{U_u}(z,X) - g_{U_{\tilde{u}}}(z,X)\bigr\|_{P,2} \le \|U_u - U_{\tilde{u}}\|_{P,2},$$

which we deduce by the definition of $g_{U_u}(z,X) = \mathrm{E}_P[U_u|X, Z=z]$ and the contraction property of the conditional expectation. *Q.E.D.*

### M.2. *Proof of Theorem 4.2*

The proof will be similar to the proof of Theorem 4.1.

*Step 0* (Preparation). In the proof $a \lesssim b$ means that $a \le Ab$, where the constant $A$ depends on the constants in Assumptions 4.1 and 4.2 only, but not on $n$ once $n \ge n_0 = \min\{j : \delta_j \le 1/2\}$, and not on $P \in \mathcal{P}_n$. We consider a sequence $P_n$ in $\mathcal{P}_n$, but for simplicity, we write $P = P_n$ throughout the proof, suppressing the index $n$. Since the argument is asymptotic, we can assume that $n \ge n_0$ in what follows. Let $\mathbb{P}_n$ denote the measure that puts mass $n^{-1}$ on points $(\xi_i, W_i)$ for $i = 1, \dots, n$. Let $\mathbb{E}_n$ denote the expectation with respect to this measure, so that $\mathbb{E}_n f = n^{-1}\sum_{i=1}^n f(\xi_i, W_i)$, and $\mathbb{G}_n$ denote the corresponding empirical process $\sqrt{n}(\mathbb{E}_n - P)$, that is,

$$\begin{aligned}\mathbb{G}_n f &= \sqrt{n}(\mathbb{E}_n f - Pf)\\ &= n^{-1/2}\sum_{i=1}^n \biggl(f(\xi_i, W_i) - \int f(s,w)\, dP_\xi(s)\, dP(w)\biggr).\end{aligned}$$

Recall that we define the bootstrap draw as

$$Z^*_{n,P} = \sqrt{n}\big(\hat\rho^* - \hat\rho\big) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \hat\psi^\rho_u(W_i)\right)_{u\in\mathcal{U}} = \big(\mathbb{G}_n \xi \hat\psi^\rho_u\big)_{u\in\mathcal{U}},$$

since $P[\xi\hat\psi^\rho_u] = 0$ because $\xi$ is independent of $W$ and has zero mean. Here $\hat\psi^\rho_u = (\hat\psi^\rho_V)_{V\in\mathcal{V}_u}$, where $\hat\psi^\rho_V(W) = \{\psi^\alpha_{V,0,\hat g_V,\hat m_Z}(W, \hat\alpha_V(0)), \psi^\alpha_{V,1,\hat g_V,\hat m_Z}(W, \hat\alpha_V(1)), \psi^\gamma_V(W, \hat\gamma_V)\}$, is a plug-in estimator of the influence function $\psi^\rho_u$.

*Step 1* (Linearization). In this step we establish that

$$\text{(M.4)}\qquad \begin{aligned} &\zeta^*_{n,P} := Z^*_{n,P} - G^*_{n,P} = o_P(1), \quad \text{for} \\ &G^*_{n,P} := \big(\mathbb{G}_n \xi \psi^\rho_u\big)_{u\in\mathcal{U}} \text{ in } \mathbb{D} = \ell^\infty(\mathcal{U})^{d_\rho}, \end{aligned}$$

where $\zeta^*_{n,P} = \zeta_{n,P}(D_n, B_n)$ is a linearization error, arising completely due to estimation of the influence function; if the influence function were known, this term would be zero.

For the components $(\sqrt{n}(\hat\gamma^*_V - \hat\gamma_V))_{V\in\mathcal{V}}$ of $\sqrt{n}(\hat\rho^* - \hat\rho)$, the linearization follows by the representation

$$\sqrt{n}\big(\hat\gamma^*_V - \hat\gamma_V\big) = \mathbb{G}_n\xi\psi^\gamma_V - \underbrace{(\hat\gamma_V - \gamma_V)\mathbb{G}_n\xi}_{\mathrm{I}^*_V},$$

for all $V \in \mathcal{V}$, and noting that $\sup_{V\in\mathcal{V}} |\mathrm{I}^*_V| = \sup_{V\in\mathcal{V}} |(\hat\gamma_V - \gamma_V)||\mathbb{G}_n\xi| = O_P(n^{-1/2})$, for $\mathcal{V}$ defined in (M.3) by Theorem 4.1 and by $|\mathbb{G}_n\xi| = O_P(1)$.

It remains to establish the claim for the empirical process $(\sqrt{n}(\hat\alpha^*_{V_{uj}}(z) - \hat\alpha_{V_{uj}}(z)))_{u\in\mathcal{U}}$ for $z \in \{0,1\}$ and $j \in \mathcal{J}$. As in the proof of Theorem 4.1, we have that with probability at least $1 - \Delta_n$,

$$\begin{aligned} \hat h_V \in \mathcal{H}_{V,n} := \big\{&h = \big(\bar g_V(0,\cdot), \bar g_V(1,\cdot), \bar m_Z(0,\cdot), \bar m_Z(1,\cdot)\big) \\ &\in \mathcal{G}_V(0)\times\mathcal{G}_V(1)\times\mathcal{M}(0)\times\mathcal{M}(1)\big\}. \end{aligned}$$

We have the representation

$$\begin{aligned} &\sqrt{n}\big(\hat\alpha^*_V(z) - \hat\alpha_V(z)\big) \\ &\quad = \mathbb{G}_n\xi\psi^\alpha_{V,z} + \underbrace{\mathbb{G}_n[\xi f_{\hat h_V,V,z} - \xi f_{h_V,V,z}] - \big(\hat\alpha_V(z) - \alpha_V(z)\big)\mathbb{G}_n\xi}_{\mathrm{II}^*_V(z)}, \end{aligned}$$

where $\sup_{V\in\mathcal{V}, z\in\{0,1\}}(\hat\alpha_V(z) - \alpha_V(z)) = O_P(n^{-1/2})$ by Theorem 4.1.

Hence to establish $\sup_{V\in\mathcal{V}} |\mathrm{II}^*_V(z)| = o_P(1)$, it remains to show that with probability $1 - \Delta_n$,

$$\sup_{z\in\{0,1\}, V\in\mathcal{V}} \big|\mathbb{G}_n[\xi f_{\hat h_V,V,z} - \xi f_{h_V,V,z}]\big| \le \sup_{f\in\xi\mathcal{J}_n} \big|\mathbb{G}_n(f)\big| = o_P(1),$$

where

$$\mathcal{J}_n = \big\{f_{h,V,z} - f_{h_V,V,z} : z \in \{0,1\}, V \in \mathcal{V}, h \in \mathcal{H}_{V,n}\big\}.$$

By the calculations in Step 1(e) of the proof of Theorem 4.1, $\mathcal{J}_n$ obeys $\log\sup_Q N(\epsilon, \mathcal{J}_n, \|\cdot\|_{Q,2}) \lesssim (s\log p + s\log(e/\epsilon)) \vee 0$. By Lemma L.1, multiplication of this class by $\xi$ does not change the entropy bound modulo an absolute constant, namely,

$$\log\sup_Q N\big(\epsilon\|J_n\|_{Q,2}, \xi\mathcal{J}_n, \|\cdot\|_{Q,2}\big) \lesssim \big(s\log p + s\log(e/\epsilon)\big) \vee 0,$$

where the envelope $J_n$ for $\xi\mathcal{J}_n$ is $|\xi|$ times a constant. Also, $\mathrm{E}[\exp(|\xi|)] < \infty$ implies that $(\mathrm{E}[\max_{i\le n}|\xi_i|^2])^{1/2} \lesssim \log n$. Thus, applying Lemma C.1 with $\sigma = \sigma_n = C'\delta_n n^{-1/4}$ and the envelope $J_n = C'|\xi|$, for some constant $K > e$,

$$\begin{aligned}\sup_{f\in\xi\mathcal{J}_n}\big|\mathbb{G}_n(f)\big| &\lesssim \left(\sqrt{s\sigma_n^2\log\big(p\vee K\vee\sigma_n^{-1}\big)} + \frac{s\log n}{\sqrt{n}}\log\big(p\vee K\vee\sigma_n^{-1}\big)\right)\\ &\lesssim \Big(\sqrt{s\delta_n^2 n^{-1/2}\log(p\vee n)} + \sqrt{s^2 n^{-1}\log^2(p\vee n)\log^2(n)}\Big)\\ &\lesssim \big(\delta_n\delta_n^{1/4} + \delta_n^{1/2}\big) \lesssim \big(\delta_n^{1/2}\big) = o_P(1),\end{aligned}$$

for $\sup_{f\in\xi\mathcal{J}_n}\|f\|_{P,2} = \sup_{f\in\mathcal{J}_n}\|f\|_{P,2} \lesssim \sigma_n$, where the details of calculations are the same as in Step 1(e) of the proof of Theorem 4.1.

Finally, we conclude that

$$\big\|\zeta^*_{n,P}\big\|_{\mathbb{D}} \lesssim \sup_{V\in\mathcal{V}}\big|\mathrm{I}^*_V\big| + \sup_{V\in\mathcal{V}, z\in\{0,1\}}\big|\mathrm{II}^*_V\big| = o_P(1).$$

*Step* 2. Here we are claiming that $Z^*_{n,P} \rightsquigarrow_B Z_P$ in $\mathbb{D}$, under any sequence $P = P_n \in \mathcal{P}_n$, where $Z_P = (\mathbb{G}_P\psi^\rho_u)_{u\in\mathcal{U}}$. We have that

$$\begin{aligned}&\sup_{h\in\mathrm{BL}_1(\mathbb{D})}\big|\mathrm{E}_{B_n}h\big(Z^*_{n,P}\big) - \mathrm{E}_P h(Z_P)\big|\\ &\quad\le \sup_{h\in\mathrm{BL}_1(\mathbb{D})}\big|\mathrm{E}_{B_n}h\big(G^*_{n,P}\big) - \mathrm{E}_P h(Z_P)\big| + \mathrm{E}_{B_n}\big(\big\|\zeta^*_{n,P}\big\|_{\mathbb{D}}\wedge 2\big),\end{aligned}$$

where the first term is $o^*_P(1)$, since $G^*_{n,P} \rightsquigarrow_B Z_P$ by Theorem B.2, and the second term is $o_P(1)$ because $\|\zeta^*_{n,P}\|_{\mathbb{D}} = o_P(1)$ implies that $\mathrm{E}_P(\|\zeta^*_{n,P}\|_{\mathbb{D}}\wedge 2) = \mathrm{E}_P\mathrm{E}_{B_n}(\|\zeta^*_{n,P}\|_{\mathbb{D}}\wedge 2) \to 0$, which in turn implies that $\mathrm{E}_{B_n}(\|\zeta^*_{n,P}\|_{\mathbb{D}}\wedge 2) = o_P(1)$ by the Markov inequality. *Q.E.D.*

### M.3. *Proof of Corollary 4.1*

This is an immediate consequence of Theorems 4.1, 4.2, B.3 and B.4. *Q.E.D.*

## APPENDIX N: OMITTED PROOFS FOR SECTION 5

LEMMA N.1—Donsker Theorem for Classes Changing With $n$: *Work with the setup described in Appendix* B *of the main text. Suppose that for some fixed constant $q > 2$ and every sequence $\delta_n \searrow 0$,*

$$\|F_n\|_{P_n,q} = O(1), \qquad \sup_{d_T(s,t)\le\delta_n}\|f_{n,s} - f_{n,t}\|_{P_n,2} \to 0,$$

$$\int_0^{\delta_n}\sup_Q\sqrt{\log N\big(\epsilon\|F_n\|_{Q,2}, \mathcal{F}_n, \|\cdot\|_{Q,2}\big)}\,d\epsilon \to 0.$$

(a) *Then the empirical process* $(\mathbb{G}_n f_{n,t})_{t\in T}$ *is asymptotically tight in* $\ell^\infty(T)$. (b) *For any subsequence such that the covariance function* $P_n f_{n,s} f_{n,t} - P_n f_{n,s} P_n f_{n,t}$ *converges pointwise on* $T \times T$, $(\mathbb{G}_n f_{n,t})_{t\in T}$ *converges in* $\ell^\infty(T)$ *to a Gaussian process with covariance function given by the limit of the covariance function along that subsequence*.

PROOF: The proof that follows is similar to the proof of Theorem 2.11.22 in van der Vaart and Wellner (1996, pp. 220–221), except that the probability law is allowed to depend on $n$. Indeed, the use of Theorem 2.11.1 in van der Vaart and Wellner (1996), which does allow for the probability space to depend on $n$, allows us to establish claim (a), whereas the proof of claim (b) follows by a standard argument.

The random distance given in Theorem 2.11.1 in van der Vaart and Wellner (1996) (Lemma N.2 below) reduces to $d_n^2(s,t) = \frac{1}{n}\sum_{i=1}^n (f_{n,s} - f_{n,t})^2(W_i) = \mathbb{P}_n(f_{n,s} - f_{n,t})^2$. It follows that $N(\varepsilon, T, d_n) = N(\varepsilon, \mathcal{F}_n, L_2(\mathbb{P}_n))$, for every $\varepsilon > 0$. If $F_n$ is replaced by $F_n \vee 1$, then the conditions of the lemma still hold. Hence, assume without loss of generality that $F_n \geq 1$. Insert the bound on the covering numbers and next make a change of variables to bound the entropy integral $\int_0^{\delta_n} \sqrt{\log N(\varepsilon, \mathcal{F}_n, d_n)}\, d\varepsilon$ in Lemma N.2 by $\int_0^{\delta_n} \sqrt{\log N(\varepsilon \|F_n\|_{\mathbb{P}_n,2}, \mathcal{F}_n, L_2(\mathbb{P}_n))}\, d\varepsilon \|F_n\|_{\mathbb{P}_n,2}$. This converges to zero in probability for every $\delta_n \downarrow 0$ by the conditions of the lemma. Apply Lemma N.2 to obtain the result. *Q.E.D.*

LEMMA N.2—van der Vaart and Wellner (1996, Theorem 2.11.1): *For each* $n$, *let* $Z_{n1}, \dots, Z_{n,m_n}$ *be independent stochastic processes, defined on the product probability space* $\prod_{i=1}^{m_n} (\mathcal{W}_{ni}, \mathcal{A}_{ni}, P_{ni})$, *with each* $Z_{ni} = Z_{ni}(f, w)$ *depending on the ith coordinate of* $w = (w_1, \dots, w_{m_n})$, *and indexed by a totally bounded semimetric space* $(T, \rho)$. *Assume that the sums* $\sum_{i=1}^{m_n} e_i Z_{ni}$ *are measurable in the sense that every one of the maps*

$$w \longmapsto \sup_{\rho(f,g)<\delta} \left| \sum_{i=1}^{m_n} e_i \big(Z_{ni}(f) - Z_{ni}(g)\big) \right|,$$

$$w \longmapsto \sup_{\rho(f,g)<\delta} \left| \sum_{i=1}^{m_n} e_i \big(Z_{ni}(f) - Z_{ni}(g)\big)^2 \right|$$

*is measurable, for every* $\delta > 0$, *every vector* $(e_1, \dots, e_{m_n}) \in \{-1, 0, 1\}^{m_n}$, *and every natural number* $n$. *Also, for every* $\eta > 0$ *and every* $\delta_n \downarrow 0$,

$$\sum_{i=1}^{m_n} \mathrm{E}^* \|Z_{ni}\|_{\mathcal{F}_n}^2 \big\{ \|Z_{ni}\|_{\mathcal{F}_n} > \eta \big\} + \sup_{\rho(s,t)<\delta_n} \sum_{i=1}^{m_n} \mathrm{E}\big(Z_{ni}(f) - Z_{ni}(g)\big)^2 \to 0,$$

*and* $\int_0^{\delta_n} \sqrt{\log N(\varepsilon, \mathcal{F}_n, d_n)}\, d\varepsilon \xrightarrow{\mathrm{P}^*} 0$, *where* $d_n$ *is the random semimetric*

$$d_n^2(f,g) = \sum_{i=1}^{m_n} \big(Z_{ni}(f) - Z_{ni}(g)\big)^2.$$

*Then the sequence* $\sum_{i=1}^{m_n} (Z_{ni} - \mathrm{E} Z_{ni})$ *is asymptotically* $\rho$*-equicontinuous*.

## APPENDIX O: PROOFS FOR SECTION 6 AND APPENDIX J

PROOF OF THEOREM 6.1: In order to establish the result uniformly in $P \in \mathcal{P}_n$, it suffices to establish the result under the probability measure induced by any sequence $P = P_n \in \mathcal{P}_n$. In the proof we shall use $P$, suppressing the dependency of $P_n$ on the sample size $n$. To prove this result, we invoke Lemmas J.3–J.5 in Appendix J. These lemmas rely on specific events (described below) and Condition WL which is also stated in Appendix J. We will show that Assumption 6.1 implies that the required events occur with probability $1-o(1)$ and also implies Condition WL.

Let $\hat{\Psi}_{u0,jj} = \{\mathbb{E}_n[|f_j(X)\zeta_u|^2]\}^{1/2}$ denote the ideal penalty loadings. The three events required to occur with probability $1-o(1)$ are the following: $E_1 := \{c_r \geq \sup_{u\in\mathcal{U}} \|r_u\|_{\mathbb{P}_n,2}\}$, and where $c_r := C\sqrt{s\log(p\vee n)/n}$; $E_2 := \{\lambda/n \geq \sqrt{c}\sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)]\|_\infty\}$, $E_3 := \{\ell\hat{\Psi}_{u0} \leq \hat{\Psi}_u \leq L\hat{\Psi}_{u0}\}$, for some $1/\sqrt{c} < 1/\sqrt[4]{c} < \ell$ and $L$ uniformly bounded for the penalty loading $\hat{\Psi}_u$ in all iterations $k \leq K$ for $n$ sufficiently large.

By Assumption 6.1(iv)(b), $E_1$ holds with probability $1-o(1)$.

Next we verify that Condition WL holds. Condition WL(i) is implied by the approximate sparsity condition in Assumption 6.1(i) and the covering condition in Assumption 6.1(ii). By Assumption 6.1 we have that $d_u$ is fixed and the Algorithm sets $\gamma \in [1/n, \min\{\log^{-1} n, pn^{d_u-1}\}]$ so that $\gamma = o(1)$ and $\Phi^{-1}(1-\gamma/\{2pn^{d_u}\}) \leq C\log^{1/2}(np) \leq C\delta_n n^{1/6}$ by Assumption 6.1(i). Since it is assumed that $\mathrm{E}_P[|f_j(X)\zeta_u|^2] \geq c$ and $\mathrm{E}_P[|f_j(X) \times \zeta_u|^3] \leq C$ uniformly in $j \leq p$ and $u \in \mathcal{U}$, Condition WL(ii) holds. Condition WL(iii) follows from Assumption 6.1(iv).

Since Condition WL holds, by Lemma J.1, the event $E_2$ occurs with probability $1-o(1)$.

Next we proceed to verify occurrence of $E_3$. In the first iteration, the penalty loadings are defined as $\hat{\Psi}_{ujj} = \{\mathbb{E}_n[|f_j(X)Y_u|^2]\}^{1/2}$ for $j = 1, \dots, p$, $u \in \mathcal{U}$. By Assumption 6.1, $\underline{c} \leq \mathrm{E}_P[|f_j(X)\zeta_u|^2] \leq \mathrm{E}_P[|f_j(X)Y_u|^2] \leq C$ uniformly over $u \in \mathcal{U}$ and $j = 1, \dots, p$. Moreover, Assumption 6.1(iv)(b) yields

$$\sup_{u\in\mathcal{U}}\max_{j\leq p}\left|(\mathbb{E}_n - \mathrm{E}_P)\left[\left|f_j(X)Y_u\right|^2\right]\right| \leq \delta_n \quad \text{and}$$

$$\sup_{u\in\mathcal{U}}\max_{j\leq p}\left|(\mathbb{E}_n - \mathrm{E}_P)\left[\left|f_j(X)\zeta_u\right|^2\right]\right| \leq \delta_n$$

with probability $1-\Delta_n$. In turn, this shows that for $n$ large so that $\delta_n \leq \underline{c}/4$, we have[3]

$$\begin{aligned}(1-2\delta_n/\underline{c})\mathbb{E}_n\left[\left|f_j(X)\zeta_u\right|^2\right] &\leq \mathbb{E}_n\left[\left|f_j(X)Y_u\right|^2\right] \\ &\leq \left(\{C+\delta_n\}/\{\underline{c}-\delta_n\}\right)\mathbb{E}_n\left[\left|f_j(X)\zeta_u\right|^2\right]\end{aligned}$$

with probability $1-\Delta_n$ so that $\ell\hat{\Psi}_{u0} \leq \hat{\Psi}_u \leq L\hat{\Psi}_{u0}$ for some uniformly bounded $L$ and $\ell > 1/\sqrt[4]{c}$. Moreover, $\tilde{\mathbf{c}} = \{(L\sqrt{c}+1)/(\sqrt{c}\ell-1)\}\sup_{u\in\mathcal{U}}\|\hat{\Psi}_{u0}^{-1}\|_\infty\|\hat{\Psi}_{u0}\|_\infty$ is uniformly bounded for $n$ large enough which implies that $\kappa_{2\tilde{\mathbf{c}}}$ as defined in (J.1) in Appendix J.2 is bounded away from zero with probability $1-\Delta_n$ by the condition on sparse eigenvalues of order $s\ell_n$ (see Bickel, Ritov, and Tsybakov (2009, Lemma 4.1(ii))).

[3]Indeed, using that $\underline{c} \leq \mathrm{E}_P[|f_j(X)\zeta_u|^2] \leq \mathrm{E}_P[|f_j(X)Y_u|^2] \leq C$, we have $(1-2\delta_n/\underline{c})\mathbb{E}_n[|f_j(X)\zeta_u|^2] \leq (1-2\delta_n/\underline{c})\{\delta_n + \mathrm{E}_P[|f_j(X)\zeta_u|^2]\} \leq \mathrm{E}_P[|f_j(X)\zeta_u|^2] - \delta_n \leq \mathrm{E}_P[|f_j(X)Y_u|^2] - \delta_n \leq \mathbb{E}_n[|f_j(X)Y_u|^2]$. Similarly, $\mathbb{E}_n[|f_j(X)Y_u|^2] \leq \delta_n + \mathrm{E}_P[|f_j(X)Y_u|^2] \leq \delta_n + C \leq (\{\delta_n + C\}/\{\underline{c}-\delta_n\})\mathbb{E}_n[|f_j(X)\zeta_u|^2]$.

By Lemma J.3, since $\lambda \in [cn^{1/2}\log^{1/2}(p \vee n), Cn^{1/2}\log^{1/2}(p \vee n)]$ by the choice of $\gamma$ and $d_u$ fixed, $c_r \le C\sqrt{s\log(p \vee n)/n}$, $\sup_{u\in\mathcal{U}}\|\hat{\Psi}_{u0}\|_\infty \le C$, we have

$$\sup_{u\in\mathcal{U}}\big\|f(X)'(\hat{\theta}_u - \theta_u)\big\|_{\mathbb{P}_n,2} \le C'\sqrt{\frac{s\log(p\vee n)}{n}} \quad \text{and}$$

$$\sup_{u\in\mathcal{U}}\|\hat{\theta}_u - \theta_u\|_1 \le C'\sqrt{\frac{s^2\log(p\vee n)}{n}}.$$

In the application of Lemma J.4, by Assumption 6.1(iv)(c), we have that $\min_{m\in\mathcal{M}}\phi_{\max}(m)$ is uniformly bounded for $n$ large enough with probability $1-o(1)$. Thus, with probability $1-o(1)$, by Lemma J.4 we have

$$\sup_{u\in\mathcal{U}}\hat{s}_u \le C\left[\frac{nc_r}{\lambda} + \sqrt{s}\right]^2 \le C's.$$

Therefore, by Lemma J.5 the Post-Lasso estimators $(\tilde{\theta}_u)_{u\in\mathcal{U}}$ satisfy, with probability $1-o(1)$,

$$\sup_{u\in\mathcal{U}}\big\|f(X)'(\tilde{\theta}_u - \theta_u)\big\|_{\mathbb{P}_n,2} \le \bar{C}\sqrt{\frac{s\log(p\vee n)}{n}} \quad \text{and}$$

$$\sup_{u\in\mathcal{U}}\|\tilde{\theta}_u - \theta_u\|_1 \le \bar{C}\sqrt{\frac{s^2\log(p\vee n)}{n}}$$

for some $\bar{C}$ independent of $n$, since uniformly in $u\in\mathcal{U}$ we have a sparsity bound $\|(\tilde{\theta}_u - \theta_u)\|_0 \le C''s$ and that ensures that a bound on the prediction rate yields a bound on the $\ell_1$-norm rate through the relations $\|v\|_1 \le \sqrt{\|v\|_0}\|v\| \le \sqrt{\|v\|_0}\|f(X)'v\|_{\mathbb{P}_n,2}/\sqrt{\phi_{\min}(\|v\|_0)}$.

In the $k$th iteration, the penalty loadings are constructed based on $(\tilde{\theta}_u^{(k)})_{u\in\mathcal{U}}$, defined as $\hat{\Psi}_{ujj} = \{\mathbb{E}_n[|f_j(X)\{Y_u - f(X)'\tilde{\theta}_u^{(k)}\}|^2]\}^{1/2}$ for $j=1,\dots,p$, $u\in\mathcal{U}$. We assume $(\tilde{\theta}_u^{(k)})_{u\in\mathcal{U}}$ satisfy the rates above uniformly in $u\in\mathcal{U}$. Then with probability $1-o(1)$, we have uniformly in $u\in\mathcal{U}$ and $j=1,\dots,p$

$$\begin{aligned}
|\hat{\Psi}_{ujj} - \hat{\Psi}_{u0jj}| &\le \big\{\mathbb{E}_n\big[\big|f_j(X)\big\{f(X)'(\tilde{\theta}_u - \theta_u)\big\}\big|^2\big]\big\}^{1/2} \\
&\quad + \big\{\mathbb{E}_n\big[\big|f_j(X)r_u\big|^2\big]\big\}^{1/2} \\
&\le K_n\big\|f(X)'(\tilde{\theta}_u - \theta_u)\big\|_{\mathbb{P}_n,2} + K_n\|r_u\|_{\mathbb{P}_n,2} \\
&\le \bar{C}K_n\sqrt{\frac{s\log(p\vee n)}{n}} \le \bar{C}\delta_n^{1/2} \le \hat{\Psi}_{u0jj}\big(2\bar{C}\delta_n^{1/2}/\underline{c}\big),
\end{aligned}$$

where we used that $\max_{i\le n, j\le p}|f_j(X_i)| \le K_n$ a.s., and $K_n^2 s\log(p\vee n) \le \delta_n n$ by Assumption 6.1(iv)(a), and that $\inf_{u\in\mathcal{U}, j\le p}\hat{\Psi}_{u0jj} \ge \underline{c}/2$ with probability $1-o(1)$ for $n$ large so that $\delta_n \le \underline{c}/2$. Further, for $n$ large so that $(2\bar{C}\delta_n^{1/2}/\underline{c}) < 1 - 1/\sqrt[4]{c}$, this establishes that the event of the penalty loadings for the $(k+1)$th iteration also satisfy $\ell\hat{\Psi}_{u0}^{-1} \le \hat{\Psi}_u^{-1} \le L\hat{\Psi}_{u0}^{-1}$ for a uniformly bounded $L$ and some $\ell > 1/\sqrt[4]{c}$ with probability $1-o(1)$ uniformly in $u\in\mathcal{U}$.

This leads to the stated rates of convergence and sparsity bound. *Q.E.D.*

PROOF OF THEOREM 6.2: In order to establish the result uniformly in $P \in \mathcal{P}_n$, it suffices to establish the result under the probability measure induced by any sequence $P = P_n \in \mathcal{P}_n$. In the proof we shall use $P$, suppressing the dependency of $P_n$ on the sample size $n$. The proof is similar to the proof of Theorem 6.1. We invoke Lemmas J.6, J.7, and J.8 which require Condition WL and some events to occur. We show that Assumption 6.2 implies Condition WL and that the required events occur with probability at least $1-o(1)$.

Let $\hat{\Psi}_{u0,jj} = \{\mathbb{E}_n[|f_j(X)\zeta_u|^2]\}^{1/2}$ denote the ideal penalty loadings, $w_{ui} = \mathrm{E}_P[Y_{ui}|X_i](1-\mathrm{E}_P[Y_{ui}|X_i])$ the conditional variance of $Y_{ui}$ given $X_i$, and $\tilde{r}_{ui} = \tilde{r}_u(X_i)$ the rescaled approximation error as defined in (J.5). The three events required to occur with probability $1-o(1)$ are as follows: $E_1 := \{c_r \geq \sup_{u\in\mathcal{U}} \|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\}$ for $c_r := C'\sqrt{s\log(p\vee n)/n}$ where $C'$ is large enough; $E_2 := \{\lambda/n \geq \sqrt{c}\sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)]\|_\infty\}$; and $E_3 := \{\ell\hat{\Psi}_{u0} \leq \hat{\Psi}_u \leq L\hat{\Psi}_{u0}\}$, for $\ell > 1/\sqrt[4]{c}$ and $L$ uniformly bounded, for the penalty loading $\hat{\Psi}_u$ in all iterations $k \leq K$ for $n$ sufficiently large.

Regarding $E_1$, by Assumption 6.2(iii), we have $\underline{c}(1-\underline{c}) \leq w_{ui} \leq 1/4$. Since $|r_u(X_i)| \leq \delta_n$ a.s. uniformly on $u \in \mathcal{U}$ for $i = 1, \dots, n$, we have that the rescaled approximation error defined in (J.5) satisfies $|\tilde{r}_u(X_i)| \leq |r_u(X_i)|/\{\underline{c}(1-\underline{c}) - 2\delta_n\}_+ \leq \tilde{C}|r_u(X_i)|$ for $n$ large enough so that $\delta_n \leq \underline{c}(1-\underline{c})/4$. Thus $\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2} \leq \tilde{C}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}$. Assumption 6.2(iv)(b) yields $\sup_{u\in\mathcal{U}} \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2} \leq C\sqrt{s\log(p\vee n)/n}$ with probability $1-o(1)$, so $E_3$ occurs with probability $1-o(1)$.

To apply Lemma J.1 to show that $E_2$ occurs with probability $1-o(1)$, we need to verify Condition WL. Condition WL(i) is implied by the sparsity in Assumption 6.2(i) and the covering condition in Assumption 6.2(ii). By Assumption 6.2 we have that $d_u$ is fixed and the Algorithm sets $\gamma \in [1/n, \min\{\log^{-1} n, pn^{d_u-1}\}]$ so that $\gamma = o(1)$ and $\Phi^{-1}(1-\gamma/\{2pn^{d_u}\}) \leq C\log^{1/2}(np) \leq C\delta_n n^{1/6}$ by Assumption 6.2(i). Since it is assumed that $\mathrm{E}_P[|f_j(X)\zeta_u|^2] \geq c$ and $\mathrm{E}_P[|f_j(X)\zeta_u|^3] \leq C$ uniformly in $j \leq p$ and $u \in \mathcal{U}$, Condition WL(ii) holds. Condition WL(iii) follows from Assumption 6.1(iv). Then, by Lemma J.1, the event $E_2$ occurs with probability $1-o(1)$.

Next we verify the occurrence of $E_3$. In the initial iteration, the penalty loadings are defined as $\hat{\Psi}_{ujj} = \frac{1}{2}\{\mathbb{E}_n[|f_j(X)|^2]\}^{1/2}$ for $j = 1, \dots, p$, $u \in \mathcal{U}$. Assumption 6.2(iv)(c) for the sparse eigenvalues implies that for $n$ large enough, $c' \leq \mathbb{E}_n[|f_j(X)|^2] \leq C'$ for all $j = 1, \dots, p$, with probability $1-o(1)$.

Moreover, Assumption 6.2(iv)(b) yields

$$\sup_{u\in\mathcal{U}}\max_{j\leq p}\left|(\mathbb{E}_n - \mathrm{E}_P)\left[\left|f_j(X)\zeta_u\right|^2\right]\right| \leq \delta_n \tag{O.1}$$

with probability $1-\Delta_n$, so that $\hat{\Psi}_{u0jj}$ is bounded away from zero and from above uniformly over $j = 1, \dots, p$, $u \in \mathcal{U}$, with the same probability because $\mathrm{E}_P[|f_j(X)\zeta_u|^2]$ is bounded away from zero and above. By (O.1) and $\mathrm{E}_P[|f_j(X)\zeta_u|^2] \leq \frac{1}{4}\mathrm{E}_P[|f_j(X)|^2]$, for $n$ large enough, we have $\ell\hat{\Psi}_{u0} \leq \hat{\Psi}_u \leq L\hat{\Psi}_{u0}$ for some uniformly bounded $L$ and $\ell > 1/\sqrt[4]{c}$ with probability $1-\Delta_n$.

Thus, $\tilde{\mathbf{c}} = \{(L\sqrt{c}+1)/(\ell\sqrt{c}-1)\}\sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}^{-1}\|_\infty\|\hat{\Psi}_{u0}\|_\infty$ is uniformly bounded. In turn, since $\inf_{u\in\mathcal{U}}\min_{i\leq n} w_{ui} \geq \underline{c}(1-\underline{c})$ is bounded away from zero, we have $\bar{\kappa}_{2\tilde{\mathbf{c}}} \geq \sqrt{\underline{c}(1-\underline{c})}\kappa_{2\tilde{\mathbf{c}}}$ by their definitions in (J.1) and (J.2). It follows that $\kappa_{2\tilde{\mathbf{c}}}$ is bounded away from zero by the condition on $s\ell_n$ sparse eigenvalues stated in Assumption 6.2(iv)(c); see Bickel, Ritov, and Tsybakov (2009, Lemma 4.1(ii)).

By the choice of $\gamma$ and $d_u$ fixed, $\lambda \in [cn^{1/2}\log^{1/2}(p\vee n), Cn^{1/2}\log^{1/2}(p\vee n)]$. By relation (J.4) and Assumption 6.2(iv)(a), $\inf_{u\in\mathcal{U}} \bar{q}_{A_u} \geq c'\bar{\kappa}_{2\tilde{\mathbf{c}}}/\{\sqrt{s}K_n\}$. Under the condition

$K_n^2 s^2 \log^2(p \vee n) \leq \delta_n n$, the side condition in Lemma J.6 holds with probability $1 - o(1)$, and the lemma yields

$$\sup_{u \in \mathcal{U}} \left\| f(X)'(\hat{\theta}_u - \theta_u) \right\|_{\mathbb{P}_n,2} \leq C' \sqrt{\frac{s \log(p \vee n)}{n}} \quad \text{and}$$
$$\sup_{u \in \mathcal{U}} \| \hat{\theta}_u - \theta_u \|_1 \leq C' \sqrt{\frac{s^2 \log(p \vee n)}{n}}.$$

In turn, under Assumption 6.2(iv)(c) and $K_n^2 s^2 \log^2(p \vee n) \leq \delta_n n$, with probability $1 - o(1)$, Lemma J.7 implies

$$\sup_{u \in \mathcal{U}} \hat{s}_u \leq C'' \left[ \frac{n c_r}{\lambda} + \sqrt{s} \right]^2 \leq C''' s,$$

since $\min_{m \in \mathcal{M}} \phi_{\max}(m)$ is uniformly bounded. The rate of convergence for $\tilde{\theta}_u$ is given by Lemma J.8, namely, with probability $1 - o(1)$,

$$\sup_{u \in \mathcal{U}} \left\| f(X)'(\tilde{\theta}_u - \theta_u) \right\|_{\mathbb{P}_n,2} \leq \bar{C} \sqrt{\frac{s \log(p \vee n)}{n}} \quad \text{and}$$
$$\sup_{u \in \mathcal{U}} \| \tilde{\theta}_u - \theta_u \|_1 \leq \bar{C} \sqrt{\frac{s^2 \log(p \vee n)}{n}}$$

for some $\bar{C}$ independent of $n$, since by (O.16) we have, uniformly in $u \in \mathcal{U}$,

$$\begin{aligned} M_u(\tilde{\theta}_u) - M_u(\theta_u) &\leq M_u(\hat{\theta}_u) - M_u(\theta_u) \leq \frac{\lambda}{n} \| \hat{\Psi}_u \theta_u \|_1 - \frac{\lambda}{n} \| \hat{\Psi}_u \hat{\theta}_u \|_1 \\ &\leq \frac{\lambda}{n} \left\| \hat{\Psi}_u (\hat{\theta}_{uT_u} - \theta_u) \right\|_1 \leq \bar{C}' s \log(p \vee n)/n, \end{aligned}$$

$\sup_{u \in \mathcal{U}} \| \mathbb{E}_n[f(X) \zeta_u] \|_\infty \leq C \sqrt{\log(p \vee n)/n}$ by Lemma J.1, $\phi_{\min}(\hat{s}_u + s_u)$ is bounded away from zero (by Assumption 6.2(iv)(c) and $\hat{s}_u \leq C''' s$), $\inf_{u \in \mathcal{U}} \psi_u(\{\delta \in \mathbb{R}^p : \|\delta\|_0 \leq \hat{s}_u + s_u\})$ is bounded away from zero (because $\inf_{u \in \mathcal{U}} \min_{i \leq n} w_{ui} \geq \underline{c}(1 - \underline{c})$), and $\sup_{u \in \mathcal{U}} \| \hat{\Psi}_{u0} \|_\infty \leq C$ with probability $1 - o(1)$.

In the $k$th iteration, the penalty loadings are constructed based on $(\tilde{\theta}_u^{(k)})_{u \in \mathcal{U}}$, defined as $\hat{\Psi}_{ujj} = \{\mathbb{E}_n[|f_j(X)\{Y_u - \Lambda(f(X)'\tilde{\theta}_u^{(k)})\}|^2]\}^{1/2}$ for $j = 1, \dots, p$, $u \in \mathcal{U}$. We assume $(\tilde{\theta}_u^{(k)})_{u \in \mathcal{U}}$ satisfy the rates above uniformly in $u \in \mathcal{U}$. Then

$$\begin{aligned} &|\hat{\Psi}_{ujj} - \hat{\Psi}_{u0jj}| \\ &\quad \leq \left\{ \mathbb{E}_n\left[ \left| f_j(X) \left\{ \Lambda\left(f(X)'\tilde{\theta}_u^{(k)}\right) - \Lambda\left(f(X)'\theta_u\right) \right\} \right|^2 \right] \right\}^{1/2} \\ &\qquad + \left\{ \mathbb{E}_n\left[ \left| f_j(X) r_u \right|^2 \right] \right\}^{1/2} \\ &\quad \leq \left\{ \mathbb{E}_n\left[ \left| f_j(X) \left\{ f(X)'\left(\tilde{\theta}_u^{(k)} - \theta_u\right) \right\} \right|^2 \right] \right\}^{1/2} + \left\{ \mathbb{E}_n\left[ \left| f_j(X) r_u \right|^2 \right] \right\}^{1/2} \\ &\quad \leq K_n \left\| f(X)'\left(\tilde{\theta}_u^{(k)} - \theta_u\right) \right\|_{\mathbb{P}_n,2} + K_n \| r_u \|_{\mathbb{P}_n,2} \lesssim_P K_n \sqrt{\frac{s \log(p \vee n)}{n}} \\ &\quad \leq C \delta_n \leq (2C\delta_n/\underline{c}) \hat{\Psi}_{u0jj}, \end{aligned}$$

and therefore, provided that $(2C\delta_n/\underline{c}) < 1 - 1/\sqrt[4]{c}$, uniformly in $u \in \mathcal{U}$, $\ell\hat{\Psi}_{u0} \leq \hat{\Psi}_u \leq L\hat{\Psi}_{u0}$ for $\ell > 1/\sqrt[4]{c}$ and $L$ uniformly bounded with probability $1 - o(1)$. Then the same proof for the initial penalty loading choice applies to the iterate $(k+1)$. *Q.E.D.*

### O.1. *Proofs for Lasso With Functional Response: Penalty Level*

PROOF OF LEMMA J.1: By the triangle inequality,

$$
\begin{aligned}
&\sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty \\
&\quad \leq \max_{u\in\mathcal{U}^\epsilon} \|\hat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty \\
&\qquad + \sup_{u\in\mathcal{U}^\epsilon, u'\in\mathcal{U}, d_\mathcal{U}(u,u')\leq\epsilon} \|\hat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u] - \hat{\Psi}_{u'0}^{-1}\mathbb{E}_n[f(X)\zeta_{u'}]\|_\infty,
\end{aligned}
$$

where $\mathcal{U}^\epsilon$ is a minimal $\epsilon$-net of $\mathcal{U}$. We will set $\epsilon = 1/n$ so that $|\mathcal{U}^\epsilon| \leq n^{d_u}$.

The proofs in this section rely on the following result due to Jing, Shao, and Wang (2003).

LEMMA O.1—Moderate Deviations for Self-Normalized Sums: *Let $Z_1, \ldots, Z_n$ be independent, zero-mean random variables and $\mu \in (0, 1]$. Let $S_{n,n} = \sum_{i=1}^n Z_i$, $V_{n,n}^2 = \sum_{i=1}^n Z_i^2$,*

$$
M_n = \left\{\frac{1}{n}\sum_{i=1}^n \mathrm{E}[Z_i^2]\right\}^{1/2} \Big/ \left\{\frac{1}{n}\sum_{i=1}^n \mathrm{E}[|Z_i|^{2+\mu}]\right\}^{1/\{2+\mu\}} > 0,
$$

*and $0 < \ell_n \leq n^{\mu/(2(2+\mu))}M_n$. Then for some absolute constant $A$,*

$$
\left|\frac{\mathrm{P}(|S_{n,n}/V_{n,n}| \geq x)}{2(1-\Phi(x))} - 1\right| \leq \frac{A}{\ell_n^{2+\mu}}, \quad 0 \leq x \leq n^{\mu/(2(2+\mu))}\frac{M_n}{\ell_n} - 1.
$$

For each $j = 1, \ldots, p$, and each $u \in \mathcal{U}^\epsilon$, we will apply Lemma O.1 with $Z_i := f_j(X_i)\zeta_{ui}$, and $\mu = 1$. Then, by Lemma O.1, the union bound, and $|\mathcal{U}^\epsilon| \leq N_n$, we have

$$
\begin{aligned}
\text{(O.2)}\qquad &\mathrm{P}_P\left(\sup_{u\in\mathcal{U}^\epsilon}\max_{j\leq p}\left|\frac{\sqrt{n}\mathbb{E}_n[f_j(X)\zeta_u]}{\sqrt{\mathbb{E}_n[f_j(X)^2\zeta_u^2]}}\right| > \Phi^{-1}\left(1 - \frac{\gamma}{2pN_n}\right)\right) \\
&\quad \leq 2pN_n(\gamma/2pN_n)\{1+o(1)\} \leq \gamma\{1+o(1)\},
\end{aligned}
$$

provided that $\max_{u,j}\{\bar{\mathrm{E}}_P[|f_j(X)\zeta_u|^3]^{1/3}/\bar{\mathrm{E}}_P[|f_j(X)\zeta_u|^2]^{1/2}\}\Phi^{-1}(1-\gamma/2pN_n) \leq \delta_n n^{1/6}$, which holds by Condition WL since $\gamma \geq 1/n$ (under this condition, there is $\ell_n \to \infty$ obeying conditions of Lemma O.1).

Moreover, by the triangle inequality, we have

$$
\begin{aligned}
\text{(O.3)}\qquad &\sup_{u\in\mathcal{U}^\epsilon, u'\in\mathcal{U}, d_\mathcal{U}(u,u')\leq\epsilon} \|\hat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u] - \hat{\Psi}_{u'0}^{-1}\mathbb{E}_n[f(X)\zeta_{u'}]\|_\infty \\
&\quad \leq \sup_{u\in\mathcal{U}^\epsilon, u'\in\mathcal{U}, d_\mathcal{U}(u,u')\leq\epsilon} \|(\hat{\Psi}_{u0}^{-1} - \hat{\Psi}_{u'0}^{-1})\hat{\Psi}_{u0}\|_\infty \|\hat{\Psi}_{u0}^{-1}\mathbb{E}_n[f(X)\zeta_u]\|_\infty \\
&\qquad + \sup_{u,u'\in\mathcal{U}, d_\mathcal{U}(u,u')\leq\epsilon} \|\mathbb{E}_n[f(X)(\zeta_u - \zeta_{u'})]\|_\infty \|\hat{\Psi}_{u'0}^{-1}\|_\infty.
\end{aligned}
$$

To control the first term in (O.3), we note that by Condition WL, $\hat{\Psi}_{u0jj}$ is bounded away from zero with probability $1-o(1)$ uniformly over $u \in \mathcal{U}$ and $j = 1, \ldots, p$. Thus we have, uniformly over $u \in \mathcal{U}$ and $j = 1, \ldots, p$,

$$\left|\left(\hat{\Psi}_{u0jj}^{-1} - \hat{\Psi}_{u'0jj}^{-1}\right)\hat{\Psi}_{u0jj}\right| = |\hat{\Psi}_{u0jj} - \hat{\Psi}_{u'0jj}|/\hat{\Psi}_{u'0jj} \leq C|\hat{\Psi}_{u0jj} - \hat{\Psi}_{u'0jj}| \tag{O.4}$$

with the same probability. Moreover, we have

$$\begin{aligned} &\sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leq \epsilon} \max_{j \leq p} \left| \left\{\mathbb{E}_n\left[f_j(X)^2 \zeta_u^2\right]\right\}^{1/2} - \left\{\mathbb{E}_n\left[f_j(X)^2 \zeta_{u'}^2\right]\right\}^{1/2} \right| \\ &\quad \leq \sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leq \epsilon} \max_{j \leq p} \left\{\mathbb{E}_n\left[f_j(X)^2 (\zeta_u - \zeta_{u'})^2\right]\right\}^{1/2}. \end{aligned} \tag{O.5}$$

Thus, relations (O.4) and (O.5) imply that, with probability $1-o(1)$,

$$\begin{aligned} &\sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leq \epsilon} \left\| \left(\hat{\Psi}_{u0}^{-1} - \hat{\Psi}_{u'0}^{-1}\right)\hat{\Psi}_{u0} \right\|_\infty \\ &\quad \lesssim \sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leq \epsilon} \max_{j \leq p} \left\{\mathbb{E}_n\left[f_j(X)^2 (\zeta_u - \zeta_{u'})^2\right]\right\}^{1/2}. \end{aligned}$$

By (O.2),

$$\sup_{u \in \mathcal{U}^\epsilon} \left\| \hat{\Psi}_{u0}^{-1} \mathbb{E}_n\left[f(X)\zeta_u\right] \right\|_\infty \leq C' \sqrt{\log(p \vee N_n \vee n)/n}$$

with probability $1-o(1)$, so that with the same probability,

$$\begin{aligned} &\sup_{u \in \mathcal{U}^\epsilon, u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leq \epsilon} \left\| \left(\hat{\Psi}_{u0}^{-1} - \hat{\Psi}_{u'0}^{-1}\right)\hat{\Psi}_{u0} \right\|_\infty \left\| \hat{\Psi}_{u0}^{-1} \mathbb{E}_n\left[f(X)\zeta_u\right] \right\|_\infty \\ &\quad \leq \sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leq \epsilon} \max_{j \leq p} \left\{\mathbb{E}_n\left[f_j(X)^2 (\zeta_u - \zeta_{u'})^2\right]\right\}^{1/2} C' \sqrt{\frac{\log(p \vee N_n \vee n)}{n}} \\ &\quad \leq \frac{o(1)}{\sqrt{n}}, \end{aligned}$$

where the last inequality follows by Condition WL(iii).

The last term in (O.3) is of the order $o(n^{-1/2})$ with probability $1-o(1)$ since by Condition WL,

$$\sup_{u,u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leq \epsilon} \left\| \mathbb{E}_n\left[f(X)(\zeta_u - \zeta_{u'})\right] \right\|_\infty \leq \delta_n n^{-1/2}$$

with probability $1-\Delta_n$, and noting that by Condition WL, $\sup_{u \in \mathcal{U}} \|\hat{\Psi}_{u0}^{-1}\|_\infty$ is uniformly bounded with probability at least $1-o(1)-\Delta_n$.

The results above imply that (O.3) is bounded by $o(1)/\sqrt{n}$ with probability $1-o(1)$. Since $\frac{1}{2}\sqrt{\log(2pN_n/\gamma)} \leq \Phi^{-1}(1-\gamma/\{2pN_n\})$ for $n$ large enough (since $\gamma/\{2pN_n\} \to 0$ and standard tail bounds), we have that with probability $1-o(1)$,

$$\begin{aligned} &\frac{(c'-c)}{\sqrt{n}} \Phi^{-1}\left(1-\gamma/\{2pN_n\}\right) \\ &\quad \geq \sup_{u \in \mathcal{U}^\epsilon, u' \in \mathcal{U}, d_{\mathcal{U}}(u,u') \leq \epsilon} \left\| \hat{\Psi}_{u0}^{-1} \mathbb{E}_n\left[f(X)\zeta_u\right] - \hat{\Psi}_{u'0}^{-1} \mathbb{E}_n\left[f(X)\zeta_{u'}\right] \right\|_\infty, \end{aligned}$$

and the result follows. *Q.E.D.*

PROOF OF LEMMA J.2: We start with the last statement of the lemma since it is more difficult (others will use similar calculations). Consider the class of functions $\mathcal{F} = \{Y_u : u \in \mathcal{U}\}$, $\mathcal{F}' = \{\mathrm{E}_P[Y_u|X] : u \in \mathcal{U}\}$, and $\mathcal{G} = \{\zeta_u^2 = (Y_u - \mathrm{E}_P[Y_u|X])^2 : u \in \mathcal{U}\}$. Let $F$ be a measurable envelope for $\mathcal{F}$ which satisfies $F \leq B_n$.

Because $\mathcal{F}$ is a VC-class of functions with VC index $C'd_u$, by Lemma L.1(1) we have

$$\log N(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \lesssim 1 + [d_u \log(e/\epsilon) \vee 0]. \tag{O.6}$$

To bound the covering number for $\mathcal{F}'$, we apply Lemma L.2, and since $\mathrm{E}[F|X] \leq F$, we have

$$\log \sup_Q N(\epsilon\|F\|_{Q,2}, \mathcal{F}', \|\cdot\|_{Q,2}) \leq \log \sup_Q N\left(\frac{\epsilon}{2}\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\right). \tag{O.7}$$

Since $\mathcal{G} \subset (\mathcal{F} - \mathcal{F}')^2$, $G = 4F^2$ is an envelope for $\mathcal{G}$ and the covering number for $\mathcal{G}$ satisfies

$$\begin{aligned}
&\log N(\epsilon\|4F^2\|_{Q,2}, \mathcal{G}, \|\cdot\|_{Q,2}) \\
&\quad\overset{\text{(i)}}{\leq} 2\log N\left(\frac{\epsilon}{2}\|2F\|_{Q,2}, \mathcal{F} - \mathcal{F}', \|\cdot\|_{Q,2}\right) \\
&\quad\overset{\text{(ii)}}{\leq} 2\log N\left(\frac{\epsilon}{4}\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\right) + 2\log N\left(\frac{\epsilon}{4}\|F\|_{Q,2}, \mathcal{F}', \|\cdot\|_{Q,2}\right) \\
&\quad\overset{\text{(iii)}}{\leq} 4\log \sup_Q N\left(\frac{\epsilon}{8}\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\right),
\end{aligned} \tag{O.8}$$

where (i) and (ii) follow by Lemma L.1(2), and (iii) follows from (O.7).

Hence, the entropy bound for the class $\mathcal{M} = \bigcup_{j\in[p]} \mathcal{M}_j$, where $\mathcal{M}_j = \{f_j^2(X)\mathcal{G}\}$, $j \in [p]$ and envelope $M = 4K_n^2F^2$, satisfies

$$\begin{aligned}
&\log N(\epsilon\|M\|_{Q,2}, \mathcal{M}, \|\cdot\|_{Q,2}) \\
&\quad\overset{\text{(a)}}{\leq} \log p + \max_{j\in[p]} \log N(\epsilon\|4K_n^2F^2\|_{Q,2}, \mathcal{M}_j, \|\cdot\|_{Q,2}) \\
&\quad\overset{\text{(b)}}{\leq} \log p + \log N(\epsilon\|4F^2\|_{Q,2}, \mathcal{G}, \|\cdot\|_{Q,2}) \\
&\quad\overset{\text{(c)}}{\leq} \log p + 4\log \sup_Q N\left(\frac{\epsilon}{8}\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\right) \\
&\quad\overset{\text{(d)}}{\lesssim} \log p + [(1 + d_u)\log(e/\epsilon) \vee 0],
\end{aligned}$$

where (a) follows by Lemma L.1(2) for union of classes, (b) holds by Lemma L.1(2) when one class has only a single function, (c) by (O.8), and (d) follows from (O.6) and $\epsilon \leq 1$. Therefore, since $\sup_{u\in\mathcal{U}} \max_{j\leq p} \mathrm{E}_P[f_j^2(X)\zeta_u^2]$ is bounded away from zero and from above, by Lemma C.1 we have with probability $1 - O(1/\log n)$ that

$$\begin{aligned}
&\sup_{u\in\mathcal{U}} \max_{j\leq p} \left|(\mathbb{E}_n - \mathrm{E}_P)[f_j^2(X)\zeta_u^2]\right| \\
&\quad\lesssim \sqrt{\frac{(1 + d_u)\log(npK_n^2B_n^2)}{n}} + \frac{(1 + d_u)K_n^2B_n^2}{n}\log(npB_n^2K_n^2),
\end{aligned}$$

using the envelope $M = 4K_n^2 B_n^2$, $v = C'$, $a = pn$, and a constant $\sigma$.

Consider the first term. By Lemma C.1 we have with probability $1 - O(1/\log n)$ that

$$\begin{aligned}
&\sup_{d_{\mathcal{U}}(u,u')\le 1/n} \left\| \mathbb{E}_n\left[ f(X)(\zeta_u - \zeta_{u'}) \right] \right\|_\infty \\
&= \sup_{d_{\mathcal{U}}(u,u')\le 1/n} \frac{1}{\sqrt{n}} \max_{j\le p} \left| \mathbb{G}_n\left( f_j(X)(\zeta_u - \zeta_{u'}) \right) \right| \\
&\lesssim \frac{1}{\sqrt{n}} \sqrt{ \frac{(1+d_u) L_n \log\left( pnK_nB_n \frac{n^\nu}{L_n} \right)}{n^\nu} } \\
&\quad + \frac{(1+d_u) K_n B_n \log\left( pnK_nB_n \frac{n^\nu}{L_n} \right)}{n}
\end{aligned}$$

using the envelope $F = 2K_nB_n$, $v = C'$, $a = pn$, the entropy bound in Lemma L.2, and $\sigma^2 \propto L_n n^{-\nu} \le F^2$ for all $n$ sufficiently large, because $L_n n^{-\nu} \searrow 0$ and

$$\begin{aligned}
&\sup_{d_{\mathcal{U}}(u,u')\le 1/n} \max_{j\le p} \mathrm{E}_P\left[ f_j(X)^2 (\zeta_u - \zeta_{u'})^2 \right] \\
&\le \sup_{d_{\mathcal{U}}(u,u')\le 1/n} \max_{j\le p} \mathrm{E}_P\left[ f_j(X)^2 (Y_u - Y_{u'})^2 \right] \\
&\le \sup_{d_{\mathcal{U}}(u,u')\le 1/n} L_n \left| u - u' \right|^\nu \max_{j\le p} \mathrm{E}_P\left[ f_j(X)^2 \right] \\
&\le C L_n n^{-\nu}.
\end{aligned}$$

To bound the second term in the statement of the lemma, it follows that

$$\begin{aligned}
&\sup_{d_{\mathcal{U}}(u,u')\le 1/n} \max_{j\le p} \mathbb{E}_n\left[ f_j(X)^2 (\zeta_u - \zeta_{u'})^2 \right] \\
&= \sup_{d_{\mathcal{U}}(u,u')\le 1/n} \max_{j\le p} \mathbb{E}_n\left[ f_j(X)^2 \left( \mathrm{E}_P[Y_u - Y_{u'} | X] \right)^2 \right] \\
&\le \sup_{d_{\mathcal{U}}(u,u')\le 1/n} \max_{j\le p} \mathbb{E}_n\left[ f_j(X)^2 \mathrm{E}_P\left[ |Y_u - Y_{u'}|^2 | X \right] \right] \\
&\le \max_{j\le p} \mathbb{E}_n\left[ f_j(X)^2 \right] \sup_{d_{\mathcal{U}}(u,u')\le 1/n} L_n \left| u - u' \right|^\nu,
\end{aligned} \tag{O.9}$$

where the first inequality holds by Jensen's inequality, and the second inequality holds by assumption. Since $c \le \max_{j\le p} \{\mathrm{E}_P[f_j(X)^2]\}^{1/2} \le C$, the result follows by Lemma C.1 which yields with probability $1 - O(1/\log n)$

$$\max_{j\le p} \left| (\mathbb{E}_n - \mathrm{E}_P)\left[ f_j(X)^2 \right] \right| \lesssim \sqrt{ \frac{\log\left( pnK_n^2 \right)}{n} } + \frac{K_n^2}{n} \log\left( pnK_n^2 \right), \tag{O.10}$$

where we used the choice $C \le \sigma = C' \le F = K_n^2$, $v = C$, $a = pn$. *Q.E.D.*

O.2. *Proofs for Lasso With Functional Response: Linear Case*

PROOF OF LEMMA J.3: Let $\hat{\delta}_u = \hat{\theta}_u - \theta_u$. Throughout the proof we assume that the events $c_r^2 \geq \sup_{u\in\mathcal{U}} \mathbb{E}_n[r_u^2]$, $\lambda/n \geq c \sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)]\|_\infty$, and $\ell\hat{\Psi}_{u0} \leq \hat{\Psi}_u \leq L\hat{\Psi}_{u0}$ occur.

By definition of $\hat{\theta}_u$,

$$\hat{\theta}_u \in \arg\min_{\theta\in\mathbb{R}^p} \mathbb{E}_n\big[\big(Y_u - f(X)'\theta\big)^2\big] + \frac{2\lambda}{n}\|\hat{\Psi}_u\theta\|_1,$$

and $\ell\hat{\Psi}_{u0} \leq \hat{\Psi}_u \leq L\hat{\Psi}_{u0}$, we have

$$\begin{aligned}
&\mathbb{E}_n\big[\big(f(X)'\hat{\delta}_u\big)^2\big] - 2\mathbb{E}_n\big[\big(Y_u - f(X)'\theta_u\big)f(X)\big]'\hat{\delta}_u \\
&\quad = \mathbb{E}_n\big[\big(Y_u - f(X)'\hat{\theta}_u\big)^2\big] - \mathbb{E}_n\big[\big(Y_u - f(X)'\theta_u\big)^2\big] \\
&\quad \leq \frac{2\lambda}{n}\|\hat{\Psi}_u\theta_u\|_1 - \frac{2\lambda}{n}\|\hat{\Psi}_u\hat{\theta}_u\|_1 \\
&\quad \leq \frac{2\lambda}{n}\|\hat{\Psi}_u\hat{\delta}_{uT_u}\|_1 - \frac{2\lambda}{n}\|\hat{\Psi}_u\hat{\delta}_{uT_u^c}\|_1 \\
&\quad \leq \frac{2\lambda}{n}L\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u}\|_1 - \frac{2\lambda}{n}\ell\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u^c}\|_1.
\end{aligned} \tag{O.11}$$

Therefore, by $c_r^2 \geq \sup_{u\in\mathcal{U}} \mathbb{E}_n[r_u^2]$ and $\lambda/n \geq c\sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}^{-1}\mathbb{E}_n[\zeta_u f(X)]\|_\infty$, we have

$$\begin{aligned}
&\mathbb{E}_n\big[\big(f(X)'\hat{\delta}_u\big)^2\big] \\
&\quad \leq 2\mathbb{E}_n\big[r_u f(X)\big]'\hat{\delta}_u + 2\big(\hat{\Psi}_{u0}^{-1}\mathbb{E}_n\big[\zeta_u f(X)\big]\big)'(\hat{\Psi}_{u0}\hat{\delta}_u) \\
&\qquad + \frac{2\lambda}{n}L\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u}\|_1 - \frac{2\lambda}{n}\ell\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u^c}\|_1 \\
&\quad \leq 2c_r\big\{\mathbb{E}_n\big[\big(f(X)'\hat{\delta}_u\big)^2\big]\big\}^{1/2} + 2\big\|\hat{\Psi}_{u0}^{-1}\mathbb{E}_n\big[\zeta_u f(X)\big]\big\|_\infty\|\hat{\Psi}_{u0}\hat{\delta}_u\|_1 \\
&\qquad + \frac{2\lambda}{n}L\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u}\|_1 - \frac{2\lambda}{n}\ell\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u^c}\|_1 \\
&\quad \leq 2c_r\big\{\mathbb{E}_n\big[\big(f(X)'\hat{\delta}_u\big)^2\big]\big\}^{1/2} + \frac{2\lambda}{cn}\|\hat{\Psi}_{u0}\hat{\delta}_u\|_1 + \frac{2\lambda}{n}L\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u}\|_1 \\
&\qquad - \frac{2\lambda}{n}\ell\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u^c}\|_1 \\
&\quad \leq 2c_r\big\{\mathbb{E}_n\big[\big(f(X)'\hat{\delta}_u\big)^2\big]\big\}^{1/2} + \frac{2\lambda}{n}\left(L + \frac{1}{c}\right)\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u}\|_1 \\
&\qquad - \frac{2\lambda}{n}\left(\ell - \frac{1}{c}\right)\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u^c}\|_1.
\end{aligned} \tag{O.12}$$

Let

$$\tilde{\mathbf{c}} := \frac{cL+1}{c\ell - 1}\sup_{u\in\mathcal{U}}\|\hat{\Psi}_{u0}\|_\infty\big\|\hat{\Psi}_{u0}^{-1}\big\|_\infty.$$

Therefore, if $\hat{\delta}_u \notin \Delta_{\tilde{\mathbf{c}},u} = \{\delta \in \mathbb{R}^p : \|\delta_{T_u^c}\|_1 \leq \tilde{\mathbf{c}}\|\delta_{T_u}\|_1\}$, we have that $(L + \frac{1}{c})\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u}\|_1 \leq (\ell - \frac{1}{c})\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u^c}\|_1$ so that

$$\{\mathbb{E}_n[(f(X)'\hat{\delta}_u)^2]\}^{1/2} \leq 2c_r.$$

Otherwise assume $\hat{\delta}_u \in \Delta_{\tilde{\mathbf{c}},u}$. In this case (O.12), the definition of $\kappa_{\tilde{\mathbf{c}}}$, and $\|\hat{\delta}_{uT_u}\|_1 \leq \sqrt{s}\|\hat{\delta}_{uT_u}\|$, we have

$$\begin{aligned}
&\mathbb{E}_n[(f(X)'\hat{\delta}_u)^2] \\
&\quad \leq 2c_r\{\mathbb{E}_n[(f(X)'\hat{\delta}_u)^2]\}^{1/2} \\
&\qquad + \frac{2\lambda}{n}\left(L + \frac{1}{c}\right)\|\hat{\Psi}_{u0}\|_\infty\sqrt{s}\{\mathbb{E}_n[(f(X)'\hat{\delta}_u)^2]\}^{1/2}/\kappa_{\tilde{\mathbf{c}}},
\end{aligned}$$

which implies

$$\{\mathbb{E}_n[(f(X)'\hat{\delta}_u)^2]\}^{1/2} \leq 2c_r + \frac{2\lambda\sqrt{s}}{n\kappa_{\tilde{\mathbf{c}}}}\left(L + \frac{1}{c}\right)\|\hat{\Psi}_{u0}\|_\infty. \tag{O.13}$$

To establish the $\ell_1$-bound, first assume that $\hat{\delta}_u \in \Delta_{2\tilde{\mathbf{c}},u}$. In that case,

$$\begin{aligned}
\|\hat{\delta}_u\|_1 &\leq (1 + 2\tilde{\mathbf{c}})\|\hat{\delta}_{uT_u}\|_1 \leq (1 + 2\tilde{\mathbf{c}})\sqrt{s}\{\mathbb{E}_n[(f(X)'\hat{\delta}_u)^2]\}^{1/2}/\kappa_{2\tilde{\mathbf{c}}} \\
&\leq (1 + 2\tilde{\mathbf{c}})\left\{2\frac{\sqrt{s}c_r}{\kappa_{2\tilde{\mathbf{c}}}} + \frac{2\lambda s}{n\kappa_{\tilde{\mathbf{c}}}\kappa_{2\tilde{\mathbf{c}}}}\left(L + \frac{1}{c}\right)\|\hat{\Psi}_{u0}\|_\infty\right\},
\end{aligned}$$

where we used that $\|\hat{\delta}_{uT_u}\|_1 \leq \sqrt{s}\|\hat{\delta}_{uT_u}\|$, the definition of the restricted eigenvalue, and the prediction rate derived in (O.13).

Otherwise note that $\hat{\delta}_u \notin \Delta_{2\tilde{\mathbf{c}},u}$ implies that $(L + \frac{1}{c})\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u}\|_1 \leq \frac{1}{2}(\ell - \frac{1}{c})\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u^c}\|_1$ so that (O.12) yields

$$\begin{aligned}
&\frac{1}{2}\frac{2\lambda}{n}\left(\ell - \frac{1}{c}\right)\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u^c}\|_1 \\
&\quad \leq \{\mathbb{E}_n[(f(X)'\hat{\delta}_u)^2]\}^{1/2}(2c_r - \{\mathbb{E}_n[(f(X)'\hat{\delta}_u)^2]\}^{1/2}) \leq c_r^2,
\end{aligned}$$

where we used that $\max_t t(2c_r - t) \leq c_r^2$. Therefore,

$$\begin{aligned}
\|\hat{\delta}_u\|_1 &\leq \left(1 + \frac{1}{2\tilde{\mathbf{c}}}\right)\|\hat{\delta}_{uT_u^c}\|_1 \leq \left(1 + \frac{1}{2\tilde{\mathbf{c}}}\right)\|\hat{\Psi}_{u0}^{-1}\|_\infty\|\hat{\Psi}_{u0}\hat{\delta}_{uT_u^c}\|_1 \\
&\leq \left(1 + \frac{1}{2\tilde{\mathbf{c}}}\right)\frac{c\|\hat{\Psi}_{u0}^{-1}\|_\infty}{\ell c - 1}\frac{n}{\lambda}c_r^2.
\end{aligned}$$

*Q.E.D.*

PROOF OF LEMMA J.4: *Step 1.* Let $L_u = 4c_0\|\hat{\Psi}_{u0}^{-1}\|_\infty[\frac{nc_r}{\lambda} + \frac{\sqrt{s}}{\kappa_{\tilde{\mathbf{c}}}}\|\hat{\Psi}_{u0}\|_\infty]$. By Step 2 below and the definition of $L_u$, we have

$$\hat{s}_u \leq \phi_{\max}(\hat{s}_u)L_u^2. \tag{O.14}$$

Consider any $M \in \mathcal{M} = \{m \in \mathbb{N} : m > 2\phi_{\max}(m)\sup_{u\in\mathcal{U}} L_u^2\}$, and suppose $\hat{s}_u > M$.

Next recall the sublinearity of the maximum sparse eigenvalue (for a proof, see Lemma 3 in Belloni and Chernozhukov (2013)), namely, for any integer $k \geq 0$ and constant $\ell \geq 1$, we have $\phi_{\max}(\ell k) \leq \lceil \ell \rceil \phi_{\max}(k)$, where $\lceil \ell \rceil$ denotes the ceiling of $\ell$. Therefore,

$$\hat{s}_u \leq \phi_{\max}(M\hat{s}_u/M)L_u^2 \leq \left\lceil \frac{\hat{s}_u}{M} \right\rceil \phi_{\max}(M)L_u^2.$$

Thus, since $\lceil k \rceil \leq 2k$ for any $k \geq 1$, we have $M \leq 2\phi_{\max}(M)L_u^2$ which violates the condition that $M \in \mathcal{M}$. Therefore, we have $\hat{s}_u \leq M$.

In turn, applying (O.14) once more with $\hat{s}_u \leq M$, we obtain $\hat{s}_u \leq \phi_{\max}(M)L_u^2$. The result follows by minimizing the bound over $M \in \mathcal{M}$.

*Step 2.* In this step we establish that, uniformly over $u \in \mathcal{U}$,

$$\sqrt{\hat{s}_u} \leq 4\sqrt{\phi_{\max}(\hat{s}_u)} \|\hat{\Psi}_{u0}^{-1}\|_\infty c_0 \left[ \frac{nc_r}{\lambda} + \frac{\sqrt{s}}{\kappa_{\tilde{\mathbf{c}}}} \|\hat{\Psi}_{u0}\|_\infty \right].$$

Let $R_u = (r_{u1}, \dots, r_{un})'$, $\mathbf{Y}_u = (Y_{u1}, \dots, Y_{un})'$, $\bar{\zeta}_u = (\zeta_{u1}, \dots, \zeta_{un})'$, and $F = [f(X_1); \dots; f(X_n)]'$. We have from the optimality conditions that the Lasso estimator $\hat{\theta}_u$ satisfies

$$\mathbb{E}_n\big[\hat{\Psi}_{ujj}^{-1} f_j(X)\big(Y_u - f(X)'\hat{\theta}_u\big)\big] = \operatorname{sign}(\hat{\theta}_{uj})\lambda/n \quad \text{for each } j \in \hat{T}_u.$$

Therefore, noting that $\|\hat{\Psi}_u^{-1}\hat{\Psi}_{u0}\|_\infty \leq 1/\ell$, we have

$$\begin{aligned}
\sqrt{\hat{s}_u}\lambda &= \big\| \big(\hat{\Psi}_u^{-1}F'(\mathbf{Y}_u - F\hat{\theta}_u)\big)_{\hat{T}_u} \big\| \\
&\leq \big\| \big(\hat{\Psi}_u^{-1}F'\bar{\zeta}_u\big)_{\hat{T}_u} \big\| + \big\| \big(\hat{\Psi}_u^{-1}F'R_u\big)_{\hat{T}_u} \big\| + \big\| \big(\hat{\Psi}_u^{-1}F'F(\theta_u - \hat{\theta}_u)\big)_{\hat{T}_u} \big\| \\
&\leq \sqrt{\hat{s}_u} \big\| \hat{\Psi}_u^{-1}\hat{\Psi}_{u0} \big\|_\infty \big\| \hat{\Psi}_{u0}^{-1}F'\bar{\zeta}_u \big\|_\infty + n\sqrt{\phi_{\max}(\hat{s}_u)} \big\| \hat{\Psi}_u^{-1} \big\|_\infty c_r \\
&\quad + n\sqrt{\phi_{\max}(\hat{s}_u)} \big\| \hat{\Psi}_u^{-1} \big\|_\infty \big\| F(\hat{\theta}_u - \theta_u) \big\|_{\mathbb{P}_n,2} \\
&\leq \sqrt{\hat{s}_u}(1/\ell) \big\| \hat{\Psi}_{u0}^{-1}F'\bar{\zeta}_u \big\|_\infty \\
&\quad + n\sqrt{\phi_{\max}(\hat{s}_u)} \frac{\big\| \hat{\Psi}_{u0}^{-1} \big\|_\infty}{\ell} \big\{ c_r + \big\| F(\hat{\theta}_u - \theta_u) \big\|_{\mathbb{P}_n,2} \big\},
\end{aligned}$$

where we used that $\|v\| \leq \|v\|_0^{1/2}\|v\|_\infty$ and

$$\begin{aligned}
&\big\| \big(F'F(\theta_u - \hat{\theta}_u)\big)_{\hat{T}_u} \big\| \\
&\quad \leq \sup_{\|\delta\|_0 \leq \hat{s}_u, \|\delta\| \leq 1} \big| \delta'F'F(\theta_u - \hat{\theta}_u) \big| \\
&\quad \leq \sup_{\|\delta\|_0 \leq \hat{s}_u, \|\delta\| \leq 1} \big\| \delta'F' \big\| \big\| F(\theta_u - \hat{\theta}_u) \big\| \\
&\quad \leq \sup_{\|\delta\|_0 \leq \hat{s}_u, \|\delta\| \leq 1} \big\{ \delta'F'F\delta \big\}^{1/2} \big\| F(\theta_u - \hat{\theta}_u) \big\| \\
&\quad \leq n\sqrt{\phi_{\max}(\hat{s}_u)} \big\| f(X)'(\theta_u - \hat{\theta}_u) \big\|_{\mathbb{P}_n,2}.
\end{aligned}$$

Since $\lambda/c \geq \sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}^{-1} F'\bar{\zeta}_u\|_\infty$, and by Lemma J.3, we have that the estimate $\hat{\theta}_u$ satisfies $\|f(X)'(\hat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leq 2c_r + 2(L + \frac{1}{c})\frac{\lambda\sqrt{s}}{n\kappa_{\tilde{c}}}\|\hat{\Psi}_{u0}\|_\infty$ so that

$$\sqrt{\hat{s}_u} \leq \frac{\sqrt{\phi_{\max}(\hat{s}_u)}\dfrac{\|\hat{\Psi}_{u0}^{-1}\|_\infty}{\ell}\left[\dfrac{3nc_r}{\lambda} + 3\left(L + \dfrac{1}{c}\right)\dfrac{\sqrt{s}}{\kappa_{\tilde{c}}}\|\hat{\Psi}_{u0}\|_\infty\right]}{\left(1 - \dfrac{1}{c\ell}\right)}$$

$$\leq 4\frac{\left(L + \dfrac{1}{c}\right)}{\left(1 - \dfrac{1}{c\ell}\right)}\frac{1}{\ell}\sqrt{\phi_{\max}(\hat{s}_u)}\|\hat{\Psi}_{u0}^{-1}\|_\infty\left[\frac{nc_r}{\lambda} + \frac{\sqrt{s}}{\kappa_{\tilde{c}}}\|\hat{\Psi}_{u0}\|_\infty\right].$$

The result follows by noting that $(L + [1/c])/(1 - 1/[\ell c]) = c_0\ell$ by definition of $c_0$. *Q.E.D.*

PROOF OF LEMMA J.5: Define $m_u := (\mathrm{E}[Y_{u1}|X_1], \ldots, \mathrm{E}[Y_{un}|X_n])'$, $\bar{\zeta}_u := (\zeta_{u1}, \ldots, \zeta_{un})'$, and the $n \times p$ matrix $F := [f(X_1); \ldots; f(X_n)]'$. For a set of indices $S \subset \{1, \ldots, p\}$, we define $\hat{P}_S = F[S](F[S]'F[S])^{-1}F[S]'$ to denote the projection matrix on the columns associated with the indices in $S$ where we interpret $\hat{P}_S$ as a null operator if $S$ is empty.

Since $Y_{ui} = m_{ui} + \zeta_{ui}$, we have

$$m_u - F\tilde{\theta}_u = (I - \hat{P}_{\tilde{T}_u})m_u - \hat{P}_{\tilde{T}_u}\bar{\zeta}_u,$$

where $I$ is the identity operator. Therefore,

$$\|m_u - F\tilde{\theta}_u\| \leq \|(I - \hat{P}_{\tilde{T}_u})m_u\| + \|\hat{P}_{\tilde{T}_u}\bar{\zeta}_u\|. \tag{O.15}$$

Since $\|F[\tilde{T}_u]/\sqrt{n}(F[\tilde{T}_u]'F[\tilde{T}_u]/n)^{-1}\| \leq \sqrt{1/\phi_{\min}(\tilde{s}_u)}$, the last term in (O.15) satisfies

$$\begin{aligned}\|\hat{P}_{\tilde{T}_u}\bar{\zeta}_u\| &\leq \sqrt{1/\phi_{\min}(\tilde{s}_u)}\|F[\tilde{T}_u]'\bar{\zeta}_u/\sqrt{n}\| \\ &\leq \sqrt{1/\phi_{\min}(\tilde{s}_u)}\sqrt{\tilde{s}_u}\|F'\bar{\zeta}_u/\sqrt{n}\|_\infty.\end{aligned}$$

By Lemma J.1 with $\gamma = 1/n$, we have that with probability $1 - o(1)$, uniformly in $u \in \mathcal{U}$,

$$\begin{aligned}\|F'\bar{\zeta}_u/\sqrt{n}\|_\infty &\leq C\sqrt{\log(p \vee n^{d_u+1})}\max_{1\leq j\leq p}\sqrt{\mathbb{E}_n[f_j(X)^2\zeta_u^2]} \\ &= C\sqrt{\log(p \vee n^{d_u+1})}\|\hat{\Psi}_{u0}\|_\infty.\end{aligned}$$

The result follows.

The last statement follows from noting that the mean square approximation error provides an upper bound to the best mean square approximation error based on the model $\tilde{T}_u$ provided that the model include the Lasso's mode, that is, $\hat{T}_u \subseteq \tilde{T}_u$. Indeed, we have

$$\begin{aligned}&\sup_{u\in\mathcal{U}}\min_{\mathrm{supp}(\theta)\subseteq\tilde{T}_u}\|\mathrm{E}_P[Y_u|X] - f(X)'\theta\|_{\mathbb{P}_n,2} \\ &\quad\leq \sup_{u\in\mathcal{U}}\min_{\mathrm{supp}(\theta)\subseteq\hat{T}_u}\|\mathrm{E}_P[Y_u|X] - f(X)'\theta\|_{\mathbb{P}_n,2}\end{aligned}$$

$$
\begin{aligned}
&\leq \sup_{u\in\mathcal{U}} \left\| \mathrm{E}_P[Y_u|X] - f(X)'\hat{\theta}_u \right\|_{\mathbb{P}_n,2} \\
&\leq c_r + \sup_{u\in\mathcal{U}} \left\| f(X)'\theta_u - f(X)'\hat{\theta}_u \right\|_{\mathbb{P}_n,2} \\
&\leq 3c_r + \left(L + \frac{1}{c}\right) \frac{2\lambda\sqrt{s}}{n\kappa_{\tilde{\mathbf{c}}}} \sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}\|_\infty,
\end{aligned}
$$

where we invoked Lemma J.3 to bound $\|f(X)'(\hat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2}$. *Q.E.D.*

### O.3. *Proofs for Lasso With Functional Response: Logistic Case*

PROOF OF LEMMA J.6: Let $\delta_u = \hat{\theta}_u - \theta_u$ and $S_u = -\mathbb{E}_n[f(X)\zeta_u]$. By definition of $\hat{\theta}_u$ we have $M_u(\hat{\theta}_u) + \frac{\lambda}{n}\|\hat{\Psi}_u\hat{\theta}_u\|_1 \leq M_u(\theta_u) + \frac{\lambda}{n}\|\hat{\Psi}_u\theta_u\|_1$. Thus,

$$
\begin{aligned}
M_u(\hat{\theta}_u) - M_u(\theta_u) &\leq \frac{\lambda}{n}\|\hat{\Psi}_u\theta_u\|_1 - \frac{\lambda}{n}\|\hat{\Psi}_u\hat{\theta}_u\|_1 \\
&\leq \frac{\lambda}{n}\|\hat{\Psi}_u\delta_{u,T_u}\|_1 - \frac{\lambda}{n}\|\hat{\Psi}_u\delta_{u,T_u^c}\|_1 \\
&\leq \frac{\lambda L}{n}\|\hat{\Psi}_{u0}\delta_{u,T_u}\|_1 - \frac{\lambda\ell}{n}\|\hat{\Psi}_{u0}\delta_{u,T_u^c}\|_1.
\end{aligned}
\tag{O.16}
$$

Moreover, by convexity of $M_u(\cdot)$ and Hölder's inequality, we have

$$
\begin{aligned}
&M_u(\hat{\theta}_u) - M_u(\theta_u) \\
&\quad \geq \partial_\theta M_u(\theta_u) \geq -\frac{\lambda}{n}\frac{1}{c}\|\hat{\Psi}_{u0}\delta_u\|_1 - \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\left\|\sqrt{w_u}f(X)'\delta_u\right\|_{\mathbb{P}_n,2}
\end{aligned}
\tag{O.17}
$$

because

$$
\begin{aligned}
\left|\partial_\theta M_u(\theta_u)'\delta_u\right| &= \left|S_u'\delta_u + \left\{\partial_\theta M_u(\theta_u) - S_u\right\}'\delta_u\right| \\
&\leq \left|S_u'\delta_u\right| + \left|\left\{\partial_\theta M_u(\theta_u) - S_u\right\}'\delta_u\right| \\
&\leq \left\|\hat{\Psi}_{u0}^{-1}S_u\right\|_\infty \|\hat{\Psi}_{u0}\delta_u\|_1 + \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\left\|\sqrt{w_u}f(X)'\delta_u\right\|_{\mathbb{P}_n,2} \\
&\leq \frac{\lambda}{n}\frac{1}{c}\|\hat{\Psi}_{u0}\delta_u\|_1 + \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\left\|\sqrt{w_u}f(X)'\delta_u\right\|_{\mathbb{P}_n,2},
\end{aligned}
\tag{O.18}
$$

where we used that $\lambda/n \geq c\sup_{u\in\mathcal{U}}\|\hat{\Psi}_{u0}^{-1}S_u\|_\infty$ and that $\partial_\theta M_u(\theta_u) = \mathbb{E}_n[\{\zeta_u + r_u\}f(X)]$ so that

$$
\begin{aligned}
\left|\left\{\partial_\theta M_u(\theta_u) - S_u\right\}'\delta_u\right| &= \left|\mathbb{E}_n\left[r_u f(X)'\delta_u\right]\right| \\
&\leq \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\left\|\sqrt{w_u}f(X)'\delta_u\right\|_{\mathbb{P}_n,2}.
\end{aligned}
\tag{O.19}
$$

Combining (O.16) and (O.17), we have

$$
\begin{aligned}
&\frac{\lambda}{n}\frac{c\ell - 1}{c}\|\hat{\Psi}_{u0}\delta_{u,T_u^c}\|_1 \\
&\quad \leq \frac{\lambda}{n}\frac{Lc + 1}{c}\|\hat{\Psi}_{u0}\delta_{u,T_u}\|_1 + \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\left\|\sqrt{w_u}f(X)'\delta_u\right\|_{\mathbb{P}_n,2},
\end{aligned}
\tag{O.20}
$$

and for $\tilde{\mathbf{c}} = \frac{Lc+1}{\ell c-1} \sup_{u\in\mathcal{U}} \|\hat{\Psi}_{u0}\|_\infty \|\hat{\Psi}_{u0}^{-1}\|_\infty \geq 1$, we have

$$\|\delta_{u,T_u^c}\|_1 \leq \tilde{\mathbf{c}}\|\delta_{u,T_u}\|_1 + \frac{n}{\lambda}\frac{c\|\hat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}.$$

Suppose $\delta_u \notin \Delta_{2\tilde{\mathbf{c}},u}$, namely $\|\delta_{u,T_u^c}\|_1 \geq 2\tilde{\mathbf{c}}\|\delta_{u,T_u}\|_1$. Thus,

$$\begin{aligned}
\|\delta_u\|_1 &\leq \big(1+\{2\tilde{\mathbf{c}}\}^{-1}\big)\|\delta_{u,T_u^c}\|_1 \\
&\leq \big(1+\{2\tilde{\mathbf{c}}\}^{-1}\big)\tilde{\mathbf{c}}\|\delta_{u,T_u}\|_1 \\
&\quad + \big(1+\{2\tilde{\mathbf{c}}\}^{-1}\big)\frac{n}{\lambda}\frac{c\|\hat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\
&\leq \big(1+\{2\tilde{\mathbf{c}}\}^{-1}\big)\frac{1}{2}\|\delta_{u,T_u^c}\|_1 \\
&\quad + \big(1+\{2\tilde{\mathbf{c}}\}^{-1}\big)\frac{n}{\lambda}\frac{c\|\hat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}.
\end{aligned}$$

The relation above implies that if $\delta_u \notin \Delta_{2\tilde{\mathbf{c}},u}$,

$$\begin{aligned}
\text{(O.21)}\quad \|\delta_u\|_1 &\leq \frac{4\tilde{\mathbf{c}}}{2\tilde{\mathbf{c}}-1}\big(1+\{2\tilde{\mathbf{c}}\}^{-1}\big)\frac{n}{\lambda}\frac{c\|\hat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2} \\
&\quad \times \|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} \\
&\leq \frac{6c\|\hat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\frac{n}{\lambda}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} =: \mathrm{I}_u,
\end{aligned}$$

where we used that $\frac{4\tilde{\mathbf{c}}}{2\tilde{\mathbf{c}}-1}(1+\{2\tilde{\mathbf{c}}\}^{-1}) \leq 6$ since $\tilde{\mathbf{c}} \geq 1$. Combining the bound with the bound

$$\|\delta_{u,T_u}\|_1 \leq \frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2} =: \mathrm{II}_u, \quad \text{if} \quad \delta_u \in \Delta_{2\tilde{\mathbf{c}},u},$$

we have that $\delta_u$ satisfies

$$\text{(O.22)}\quad \|\delta_{u,T_u}\|_1 \leq \mathrm{I}_u + \mathrm{II}_u.$$

For every $u \in \mathcal{U}$, since

$$\begin{aligned}
A_u &= \Delta_{2\tilde{\mathbf{c}},u} \\
&\quad \cup \left\{\delta : \|\delta\|_1 \leq \frac{6c\|\hat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}\frac{n}{\lambda}\|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}\right\},
\end{aligned}$$

it follows that $\delta_u \in A_u$, and we have

$$\begin{aligned}
&\frac{1}{3}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}^2 \wedge \left\{\frac{\bar{q}_{A_u}}{3}\|\sqrt{w_u}f(X)'\delta_u\|_{\mathbb{P}_n,2}\right\} \\
&\quad \leq_{(1)} M_u(\hat{\theta}_u) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\delta_u
\end{aligned}$$

$$
\begin{aligned}
&+2\|\tilde r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\big\|\sqrt{w_u}f(X)'\delta_u\big\|_{\mathbb{P}_n,2}\\
&\leq_{(2)}\left(L+\frac{1}{c}\right)\frac{\lambda}{n}\|\hat\Psi_{u0}\delta_{u,T_u}\|_1+3\|\tilde r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\big\|\sqrt{w_u}f(X)'\delta_u\big\|_{\mathbb{P}_n,2}\\
&\leq_{(3)}\left(L+\frac{1}{c}\right)\|\hat\Psi_{u0}\|_\infty\frac{\lambda}{n}\{\mathrm{I}_u+\mathrm{II}_u\}\\
&\quad+3\|\tilde r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\big\|\sqrt{w_u}f(X)'\delta_u\big\|_{\mathbb{P}_n,2}\\
&\leq_{(4)}\left\{\left(L+\frac{1}{c}\right)\|\hat\Psi_{u0}\|_\infty\frac{\lambda\sqrt{s}}{n\bar\kappa_{2\tilde{\mathbf{c}}}}+9\tilde{\mathbf{c}}\|\tilde r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}\big\|\sqrt{w_u}f(X)'\delta_u\big\|_{\mathbb{P}_n,2},
\end{aligned}
$$

where (1) follows by Lemma O.2 with $A_u$, (2) follows from (O.18) and $|r_{ui}|\leq|\tilde r_{ui}|$, (3) follows by $\|\hat\Psi_{u0}\delta_{u,T_u}\|_1\leq\|\hat\Psi_{u0}\|_\infty\|\delta_{u,T_u}\|_1$ and (O.22), (4) follows from simplifications and $|r_{ui}|\leq|\tilde r_{ui}|$. Since the inequality $(x^2\wedge ax)\leq bx$ holding for $x>0$ and $b<a<0$ implies $x\leq b$, the above system of the inequalities, provided that for every $u\in\mathcal{U}$

$$
\bar q_{A_u}>3\left\{\left(L+\frac{1}{c}\right)\|\hat\Psi_{u0}\|_\infty\frac{\lambda\sqrt{s}}{n\bar\kappa_{2\tilde{\mathbf{c}}}}+9\tilde{\mathbf{c}}\|\tilde r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\},
$$

implies that

$$
\begin{aligned}
\big\|\sqrt{w_u}f(X)'\delta_u\big\|_{\mathbb{P}_n,2}&\leq 3\left\{\left(L+\frac{1}{c}\right)\|\hat\Psi_{u0}\|_\infty\frac{\lambda\sqrt{s}}{n\bar\kappa_{2\tilde{\mathbf{c}}}}+9\tilde{\mathbf{c}}\|\tilde r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}\\
&=:\mathrm{III}_u\quad\text{for every } u\in\mathcal{U}.
\end{aligned}
$$

The second result follows from the definition of $\bar\kappa_{2\tilde{\mathbf{c}}}$, (O.21), and the bound on $\|\sqrt{w_u}\times f(X)'\delta_u\|_{\mathbb{P}_n,2}$ just derived, namely, for every $u\in\mathcal{U}$, we have

$$
\begin{aligned}
\|\delta_u\|_1&\leq 1\{\delta_u\in\Delta_{2\tilde{\mathbf{c}},u}\}\|\delta_u\|_1+1\{\delta_u\notin\Delta_{2\tilde{\mathbf{c}},u}\}\|\delta_u\|_1\\
&\leq(1+2\tilde{\mathbf{c}})\mathrm{II}_u+\mathrm{I}_u\\
&\leq 3\left\{\frac{(1+2\tilde{\mathbf{c}})\sqrt{s}}{\bar\kappa_{2\tilde{\mathbf{c}}}}+\frac{6c\|\hat\Psi_{u0}^{-1}\|_\infty}{\ell c-1}\frac{n}{\lambda}\left\|\frac{r_u}{\sqrt{w_u}}\right\|_{\mathbb{P}_n,2}\right\}\mathrm{III}_u.
\end{aligned}
$$

*Q.E.D.*

PROOF OF LEMMA J.7: The proofs of both bounds are similar to the proof of sparsity for the linear case (Lemma J.4) differing only on the definition of $L_u$ which is a consequence of pre-sparsity bounds established in Step 2 and Step 3.

*Step 1.* To establish the first bound by Step 2 below, triangle inequality, and the definition of $\psi(A_u)$, we have

$$
\begin{aligned}
\sqrt{\hat s_u}&\leq\frac{c(n/\lambda)}{(c\ell-1)}\sqrt{\phi_{\max}(\hat s_u)}\big\|f(X)'(\hat\theta_u-\theta_u)-r_u\big\|_{\mathbb{P}_n,2}\\
&\leq\frac{c(n/\lambda)}{(c\ell-1)}\sqrt{\phi_{\max}(\hat s_u)}\left\{\frac{\big\|\sqrt{w_u}f(X)'(\hat\theta_u-\theta_u)\big\|_{\mathbb{P}_n,2}}{\psi(A_u)}+\|r_u\|_{\mathbb{P}_n,2}\right\}
\end{aligned}
$$

uniformly in $u \in \mathcal{U}$. By Lemma J.6, $\psi(A_u) \leq 1$, and $\|r_u\|_{\mathbb{P}_n,2} \leq \|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}$, we have

$$\begin{aligned}\sqrt{\hat{s}_u} &\leq \sqrt{\phi_{\max}(\hat{s}_u)}\frac{c(n/\lambda)}{(c\ell-1)\psi(A_u)}\\ &\quad\times\left\{3\left(L+\frac{1}{c}\right)\|\hat{\Psi}_{u0}\|_\infty\frac{(\lambda/n)\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}+28\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}\\ &\leq \sqrt{\phi_{\max}(\hat{s}_u)}\frac{c_0}{\psi(A_u)}\left\{3\|\hat{\Psi}_{u0}\|_\infty\frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}+28\tilde{\mathbf{c}}\frac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\right\}.\end{aligned}$$

Let $L_u = \frac{c_0}{\psi(A_u)}\{3\|\hat{\Psi}_{u0}\|_\infty\frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}+28\tilde{\mathbf{c}}\frac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\}$. Thus we have

$$\hat{s}_u \leq \phi_{\max}(\hat{s}_u)L_u^2, \tag{O.23}$$

which has the same structure as (O.14) in Step 1 of the proof of Lemma J.4.

Consider any $M \in \mathcal{M} = \{m \in \mathbb{N} : m > 2\phi_{\max}(m)\sup_{u\in\mathcal{U}} L_u^2\}$, and suppose $\hat{s}_u > M$. By the sublinearity of the maximum sparse eigenvalue (Lemma 3 in Belloni and Chernozhukov (2013)), for any integer $k \geq 0$ and constant $\ell \geq 1$, we have $\phi_{\max}(\ell k) \leq \lceil \ell \rceil \phi_{\max}(k)$, where $\lceil \ell \rceil$ denotes the ceiling of $\ell$. Therefore

$$\hat{s}_u \leq \phi_{\max}(M\hat{s}_u/M)L_u^2 \leq \left\lceil \frac{\hat{s}_u}{M}\right\rceil \phi_{\max}(M)L_u^2.$$

Thus, since $\lceil k \rceil \leq 2k$ for any $k \geq 1$, we have $M \leq 2\phi_{\max}(M)L_u^2$ which violates the condition that $M \in \mathcal{M}$. Therefore, we have $\hat{s}_u \leq M$. In turn, applying (O.23) once more with $\hat{s}_u \leq M$, we obtain $\hat{s}_u \leq \phi_{\max}(M)L_u^2$. The result follows by minimizing the bound over $M \in \mathcal{M}$.

Next we establish the second bound. By Step 3 below, we have

$$\sqrt{\hat{s}_u} \leq \frac{2c(n/\lambda)}{(c\ell-1)}\sqrt{\phi_{\max}(\hat{s}_u)}\big\|\sqrt{w_u}\{f(X)'(\hat{\theta}_u-\theta_u)-\tilde{r}_u\}\big\|_{\mathbb{P}_n,2}.$$

By Lemma J.6 and that $\|\sqrt{w_u}\tilde{r}_u\|_{\mathbb{P}_n,2} \leq \|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}$, we have

$$\begin{aligned}\sqrt{\hat{s}_u} &\leq \sqrt{\phi_{\max}(\hat{s}_u)}\frac{2c(n/\lambda)}{(c\ell-1)}\\ &\quad\times\left\{3\left(L+\frac{1}{c}\right)\|\hat{\Psi}_{u0}\|_\infty\frac{(\lambda/n)\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}+28\tilde{\mathbf{c}}\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\}\\ &\leq \sqrt{\phi_{\max}(\hat{s}_u)}2c_0\left\{3\|\hat{\Psi}_{u0}\|_\infty\frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}+28\tilde{\mathbf{c}}\frac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\right\}.\end{aligned}$$

Let $L_u = 2c_0\{3\|\hat{\Psi}_{u0}\|_\infty\frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{\mathbf{c}}}}+28\tilde{\mathbf{c}}\frac{n\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}}{\lambda}\}$. Thus again we obtained the relation (O.14) and the proof follows similarly to Step 1 in the proof of Lemma J.4.

*Step 2.* In this step we show that, uniformly over $u \in \mathcal{U}$,

$$\sqrt{\hat{s}_u} \leq \frac{c(n/\lambda)}{(c\ell-1)}\sqrt{\phi_{\max}(\hat{s}_u)}\big\|f(X)'(\hat{\theta}_u-\theta_u)-r_u\big\|_{\mathbb{P}_n,2}. \tag{O.24}$$

Let $\Lambda_{ui} := \mathrm{E}_P[Y_{ui}|X_i]$ and $S_u = -\mathbb{E}_n[f(X)\zeta_u] = -\mathbb{E}_n[(Y_u - \Lambda_u)f(X)]$. Let $\hat{T}_u = \mathrm{supp}(\hat\theta_u)$, $\hat{s}_u = \|\hat\theta_u\|_0$, $\delta_u = \hat\theta_u - \theta_u$, and $\hat\Lambda_{ui} = \exp(f(X_i)'\hat\theta_u)/\{1 + \exp(f(X_i)'\hat\theta_u)\}$. For any $j \in \hat{T}_u$, we have $|\mathbb{E}_n[(Y_u - \hat\Lambda_u)f_j(X)]| = \hat\Psi_{ujj}\lambda/n$.

Since $\ell\hat\Psi_{u0} \leq \hat\Psi_u$ implies $\|\hat\Psi_u^{-1}\hat\Psi_{u0}\|_\infty \leq 1/\ell$, the first relation follows from

$$\begin{aligned}
\frac{\lambda}{n}\sqrt{\hat{s}_u} &= \big\|\big(\hat\Psi_u^{-1}\mathbb{E}_n\big[(Y_u - \hat\Lambda_u)f(X)\big]\big)_{\hat{T}_u}\big\| \\
&\leq \big\|\hat\Psi_u^{-1}\hat\Psi_{u0}\big\|_\infty \big\|\hat\Psi_{u0}^{-1}\mathbb{E}_n\big[(Y_u - \Lambda_u)f_{\hat{T}_u}(X)\big]\big\| \\
&\quad + \big\|\hat\Psi_u^{-1}\hat\Psi_{u0}\big\|_\infty \big\|\hat\Psi_{u0}^{-1}\big\|_\infty \big\|\mathbb{E}_n\big[(\hat\Lambda_u - \Lambda_u)f_{\hat{T}_u}(X)\big]\big\| \\
&\leq \sqrt{\hat{s}_u}(1/\ell)\big\|\hat\Psi_{u0}^{-1}\mathbb{E}_n\big[\zeta_u f(X)\big]\big\|_\infty \\
&\quad + (1/\ell)\big\|\hat\Psi_{u0}^{-1}\big\|_\infty \sup_{\|\theta\|_0 \leq \hat{s}_u, \|\theta\|=1} \mathbb{E}_n\big[|\hat\Lambda_u - \Lambda_u|\big|f(X)'\theta\big|\big] \\
&\leq \frac{\lambda}{\ell c n}\sqrt{\hat{s}_u} + \sqrt{\phi_{\max}(\hat{s}_u)}(1/\ell)\big\|\hat\Psi_{u0}^{-1}\big\|_\infty \big\|f(X)'\delta_u - r_u\big\|_{\mathbb{P}_n,2}
\end{aligned}$$

uniformly in $u \in \mathcal{U}$, where we used that $\Lambda$ is 1-Lipschitz. This relation implies (O.24).

*Step 3.* In this step we show that if $\max_{i\leq n}|f(X_i)'(\hat\theta_u - \theta_u) - \tilde{r}_{ui}| \leq 1$, we have

$$\sqrt{\hat{s}_u} \leq \frac{2c(n/\lambda)}{(c\ell - 1)}\sqrt{\phi_{\max}(\hat{s}_u)}\big\|\sqrt{w_u}\big\{f(X)'(\hat\theta_u - \theta_u) - \tilde{r}_u\big\}\big\|_{\mathbb{P}_n,2}. \tag{O.25}$$

Note that uniformly in $u \in \mathcal{U}$, Lemma O.5 establishes that $|\hat\Lambda_{ui} - \Lambda_{ui}| \leq w_{ui}2|f(X)'\delta_u - \tilde{r}_{ui}|$ since $\max_{i\leq n}|f(X_i)'\delta_u - \tilde{r}_{ui}| \leq 1$ is assumed. Thus, combining this bound with the calculations performed in Step 2, we obtain

$$\begin{aligned}
\frac{\lambda}{n}\sqrt{\hat{s}_u} &\leq \frac{\lambda}{\ell c n}\sqrt{\hat{s}_u} \\
&\quad + (2/\ell)\big\|\hat\Psi_{u0}^{-1}\big\|_\infty\sqrt{\phi_{\max}(\hat{s}_u)}\big\|\sqrt{w_u}\big\{f(X)'\delta_u - \tilde{r}_u\big\}\big\|_{\mathbb{P}_n,2},
\end{aligned}$$

which implies (O.25). *Q.E.D.*

PROOF OF LEMMA J.8: Let $\tilde\delta_u = \tilde\theta_u - \theta_u$ and $\tilde{t}_u = \|\sqrt{w_u}f(X)'\tilde\delta_u\|_{\mathbb{P}_n,2}$ and $S_u = -\mathbb{E}_n[f(X)\zeta_u]$.

By Lemma O.2 with $A_u = \{\delta \in \mathbb{R}^p : \|\delta\|_0 \leq \tilde{s}_u + s_u\}$, we have

$$\begin{aligned}
\frac{1}{3}\tilde{t}_u^2 \wedge \left\{\frac{\bar{q}_{A_u}}{3}\tilde{t}_u\right\} &\leq M_u(\tilde\theta_u) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\tilde\delta_u + 2\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\tilde{t}_u \\
&\leq M_u(\tilde\theta_u) - M_u(\theta_u) + \|S_u\|_\infty\|\tilde\delta_u\|_1 + 3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\tilde{t}_u \\
&\leq M_u(\tilde\theta_u) - M_u(\theta_u) \\
&\quad + \tilde{t}_u\left\{\frac{\sqrt{\tilde{s}_u + s_u}\|S_u\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\tilde{s}_u + s_u)}} + 3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\},
\end{aligned}$$

where the second inequality holds by calculations as in (O.18) and Hölder's inequality, and the last inequality follows from

$$\|\tilde{\delta}_u\|_1 \leq \sqrt{\tilde{s}_u + s_u}\|\tilde{\delta}_u\|_1 \leq \frac{\sqrt{\tilde{s}_u + s_u}}{\sqrt{\phi_{\min}(\tilde{s}_u + s_u)}}\|f(X)'\tilde{\delta}_u\|_{\mathbb{P}_n,2}$$
$$\leq \frac{\sqrt{\tilde{s}_u + s_u}}{\sqrt{\phi_{\min}(\tilde{s}_u + s_u)}} \frac{\|\sqrt{w_u}f(X)'\tilde{\delta}_u\|_{\mathbb{P}_n,2}}{\psi_u(A_u)}$$

by the definition $\psi_u(A) := \min_{\delta\in A} \frac{\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}}{\|f(X)'\delta\|_{\mathbb{P}_n,2}}$.

Recall the assumed conditions $\bar{q}_{A_u}/6 > \{\frac{\sqrt{\tilde{s}_u+s_u}\|S_u\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\tilde{s}_u+s_u)}} + 3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\}$ and $\bar{q}_{A_u}/6 > \sqrt{M_u(\tilde{\theta}_u) - M_u(\theta_u)}$. If $\frac{1}{3}\tilde{t}_u^2 > \{\frac{\bar{q}_{A_u}}{3}\tilde{t}_u\}$, then

$$\frac{\bar{q}_{A_u}}{3}\tilde{t}_u \leq \frac{\bar{q}_{A_u}}{6}\sqrt{M_u(\tilde{\theta}_u) - M_u(\theta_u)} + \frac{\bar{q}_{A_u}}{6}\tilde{t}_u,$$

so that $\tilde{t}_u \leq \sqrt{0 \vee \{M_u(\tilde{\theta}_u) - M_u(\theta_u)\}}$ which implies the result. Otherwise, we have

$$\frac{1}{3}\tilde{t}_u^2 \leq \{M_u(\tilde{\theta}_u) - M_u(\theta_u)\}$$
$$+ \tilde{t}_u\left\{\frac{\sqrt{\tilde{s}_u + s_u}\|S_u\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\tilde{s}_u + s_u)}} + 3\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\right\},$$

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

$$\tilde{t}_u \leq \sqrt{3}\sqrt{0 \vee \{M_u(\tilde{\theta}_u) - M_u(\theta_u)\}}$$
$$+ 3\left\{\frac{\sqrt{\tilde{s}_u + s_u}\|S_u\|_\infty}{\psi_u(A_u)\sqrt{\phi_{\min}(\tilde{s}_u + s_u)}} + 3\|\tilde{r}_{ui}/\sqrt{w_{ui}}\|_{\mathbb{P}_n,2}\right\}.$$

*Q.E.D.*

### O.4. *Technical Lemmas: Logistic Case*

The proof of the following lower bound builds upon ideas developed in Belloni and Chernozhukov (2011) for high-dimensional quantile regressions.

LEMMA O.2—Minoration Lemma: *For any $u \in \mathcal{U}$ and $\delta \in A_u \subset \mathbb{R}^p$, we have*

$$M_u(\theta_u + \delta) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\delta$$
$$+ 2\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}$$
$$\geq \left\{\frac{1}{3}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}^2\right\} \wedge \left\{\frac{\bar{q}_{A_u}}{3}\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2}\right\},$$

*where*

$$\bar{q}_{A_u} = \inf_{\delta\in A_u} \frac{\mathbb{E}_n[w_u|f(X)'\delta|^2]^{3/2}}{\mathbb{E}_n[w_u|f(X)'\delta|^3]}.$$

PROOF: *Step 1* (Minoration). Consider the following nonnegative convex function:

$$F_u(\delta) = M_u(\theta_u + \delta) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\delta + 2\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\big\|\sqrt{w_u}f(X)'\delta\big\|_{\mathbb{P}_n,2}.$$

Note that if $\bar{q}_{A_u} = 0$, the statement is trivial since $F_u(\delta) \geq 0$. Thus we can assume $\bar{q}_{A_u} > 0$.

Step 2 below shows that for any $\delta = t\tilde{\delta} \in \mathbb{R}^p$ where $t \in \mathbb{R}$ and $\tilde{\delta} \in A_u$ such that $\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2} \leq \bar{q}_{A_u}$, we have

$$F_u(\delta) \geq \frac{1}{3}\big\|\sqrt{w_u}f(X)'\delta\big\|^2_{\mathbb{P}_n,2}. \tag{O.26}$$

Thus (O.26) covers the case that $\delta \in A_u$ and $\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2} \leq \bar{q}_{A_u}$.

In the case that $\delta \in A_u$ and $\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2} > \bar{q}_{A_u}$, by convexity[4] of $F_u$ and $F_u(0) = 0$ we have

$$\begin{aligned} F_u(\delta) &\geq \frac{\big\|\sqrt{w_u}f(X)'\delta\big\|_{\mathbb{P}_n,2}}{\bar{q}_{A_u}} F_u\left(\delta \frac{\bar{q}_{A_u}}{\big\|\sqrt{w_u}f(X)'\delta\big\|_{\mathbb{P}_n,2}}\right) \\ &\geq \frac{\bar{q}_{A_u}\big\|\sqrt{w_u}f(X)'\delta\big\|_{\mathbb{P}_n,2}}{3}, \end{aligned} \tag{O.27}$$

where the last step follows by (O.26) since

$$\big\|\sqrt{w_u}f(X)'\bar{\delta}\big\|_{\mathbb{P}_n,2} = \bar{q}_{A_u} \quad \text{for} \quad \bar{\delta} = \delta \frac{\bar{q}_{A_u}}{\big\|\sqrt{w_u}f(X)'\delta\big\|_{\mathbb{P}_n,2}}.$$

Combining (O.26) and (O.27), we have

$$F_u(\delta) \geq \left\{\frac{1}{3}\big\|\sqrt{w_u}f(X)'\delta\big\|^2_{\mathbb{P}_n,2}\right\} \wedge \left\{\frac{\bar{q}_{A_u}}{3}\big\|\sqrt{w_u}f(X)'\delta\big\|_{\mathbb{P}_n,2}\right\}.$$

*Step 2* (Proof of (O.26)). Let $\tilde{r}_{ui}$ be such that $\Lambda(f(X_i)'\theta_u + \tilde{r}_{ui}) = \Lambda(f(X_i)'\theta_u) + r_{ui} = \mathrm{E}_P[Y_{ui}|X_i]$. Defining $g_{ui}(t) = \log\{1 + \exp(f(X_i)'\theta_u + \tilde{r}_{ui} + tf(X_i)'\delta)\}$, $\tilde{g}_{ui}(t) = \log\{1 + \exp(f(X_i)'\theta_u + tf(X_i)'\delta)\}$, $\Lambda_{ui} := \mathrm{E}_P[Y_{ui}|X_i]$, $\tilde{\Lambda}_{ui} := \exp(f(X_i)'\theta_u)/\{1 + \exp(f(X_i)'\theta_u)\}$, we have

$$\begin{aligned} &M_u(\theta_u + \delta) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\delta \\ &\quad = \mathbb{E}_n\big[\log\big\{1 + \exp\big(f(X)'\{\theta_u + \delta\}\big)\big\} - Y_u f(X)'(\theta_u + \delta)\big] \\ &\qquad - \mathbb{E}_n\big[\log\big\{1 + \exp\big(f(X)'\theta_u\big)\big\} - Y_u f(X)'\theta_u\big] \\ &\qquad - \mathbb{E}_n\big[(\tilde{\Lambda}_u - Y_u)f(X)'\delta\big] \\ &\quad = \mathbb{E}_n\big[\log\big\{1 + \exp\big(f(X)'\{\theta_u + \delta\}\big)\big\} \\ &\qquad - \log\big\{1 + \exp\big(f(X)'\theta_u\big)\big\} - \tilde{\Lambda}_u f(X)'\delta\big] \end{aligned} \tag{O.28}$$

[4]If $\phi$ is a convex function with $\phi(0) = 0$, for $\alpha \in (0,1)$ we have $\phi(t) \geq \phi(\alpha t)/\alpha$. Indeed, by convexity, $\phi(\alpha t + (1-\alpha)0) \leq (1-\alpha)\phi(0) + \alpha\phi(t) = \alpha\phi(t)$.

$$
\begin{aligned}
&= \mathbb{E}_n\big[\tilde{g}_u(1) - \tilde{g}_u(0) - \tilde{g}_u'(0)\big] \\
&= \mathbb{E}_n\big[g_u(1) - g_u(0) - g_u'(0)\big] \\
&\quad + \mathbb{E}_n\big[\big\{\tilde{g}_u(1) - g_u(1)\big\} - \big\{\tilde{g}_u(0) - g_u(0)\big\} - \big\{\tilde{g}_u'(0) - g_u'(0)\big\}\big].
\end{aligned}
$$

Note that the function $g_{ui}$ is three times differentiable and satisfies

$$
\begin{aligned}
&g_{ui}'(t) = \big(f(X_i)'\delta\big)\Lambda_{ui}(t), \\
&g_{ui}''(t) = \big(f(X_i)'\delta\big)^2 \Lambda_{ui}(t)\big[1 - \Lambda_{ui}(t)\big], \quad \text{and} \\
&g_{ui}'''(t) = \big(f(X_i)'\delta\big)^3 \Lambda_{ui}(t)\big[1 - \Lambda_{ui}(t)\big]\big[1 - 2\Lambda_{ui}(t)\big],
\end{aligned}
$$

where $\Lambda_{ui}(t) := \exp(f(X_i)'\theta_u + \tilde{r}_{ui} + tf(X_i)'\delta)/\{1 + \exp(f(X_i)'\theta_u + \tilde{r}_{ui} + tf(X)'\delta)\}$. Thus we have $|g_{ui}'''(t)| \leq |f(X)'\delta| g_{ui}''(t)$. Therefore, by Lemmas O.3 and O.4 given following the conclusion of this proof, we have

$$
\begin{aligned}
(\text{O.29}) \quad & g_{ui}(1) - g_{ui}(0) - g_{ui}'(0) \\
&\geq \frac{\big(f(X_i)'\delta\big)^2 w_{ui}}{\big(f(X_i)'\delta\big)^2}\big\{\exp\big(-\big|f(X_i)'\delta\big|\big) + \big|f(X_i)'\delta\big| - 1\big\} \\
&\geq w_{ui}\left\{\frac{\big|f(X_i)'\delta\big|^2}{2} - \frac{\big|f(X_i)'\delta\big|^3}{6}\right\}.
\end{aligned}
$$

Moreover, letting $Y_{ui}(t) = \tilde{g}_{ui}(t) - g_{ui}(t)$, we have

$$
\big|Y_{ui}'(t)\big| = \big|\big(f(X_i)'\delta\big)\big\{\Lambda_{ui}(t) - \tilde{\Lambda}_{ui}(t)\big\}\big| \leq \big|f(X_i)'\delta\big||\tilde{r}_{ui}|,
$$

where $\tilde{\Lambda}_{ui}(t) := \exp(f(X_i)'\theta_u + tf(X_i)'\delta)/\{1 + \exp(f(X_i)'\theta_u + tf(X_i)'\delta)\}$. Thus

$$
\begin{aligned}
(\text{O.30}) \quad & \big|\mathbb{E}_n\big[\big\{\tilde{g}_u(1) - g_u(1)\big\} - \big\{\tilde{g}_u(0) - g_u(0)\big\} - \big\{\tilde{g}_u'(0) - g_u'(0)\big\}\big]\big| \\
&= \big|\mathbb{E}_n\big[Y_u(1) - Y_u(0) - \{\tilde{\Lambda}_u - \Lambda_u\}f(X)'\delta\big]\big| \\
&\leq 2\mathbb{E}_n\big[|\tilde{r}_u|\big|f(X)'\delta\big|\big].
\end{aligned}
$$

Therefore, combining (O.28) with the bounds (O.29) and (O.30), we have

$$
\begin{aligned}
& M_u(\theta_u + \delta) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\delta \\
&\geq \frac{1}{2}\mathbb{E}_n\big[w_u\big|f(X)'\delta\big|^2\big] - \frac{1}{6}\mathbb{E}_n\big[w_u\big|f(X)'\delta\big|^3\big] \\
&\quad - 2\|\tilde{r}_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\big\|\sqrt{w_u}f(X)'\delta\big\|_{\mathbb{P}_n,2},
\end{aligned}
$$

which holds for any $\delta \in \mathbb{R}^p$.

Take any $\delta = t\tilde{\delta}$, $t \in \mathbb{R} \setminus \{0\}$, $\tilde{\delta} \in A_u$ such that $\|\sqrt{w_u}f(X)'\delta\|_{\mathbb{P}_n,2} \leq \bar{q}_{A_u}$. (Note that the case of $\delta = 0$ is trivial.) We have

$$
\mathbb{E}_n\big[w_u\big|f(X)'\delta\big|^2\big]^{1/2} = \big\|\sqrt{w_u}f(X)'\delta\big\|_{\mathbb{P}_n,2} \leq \bar{q}_{A_u}
$$

$$\begin{aligned}
&\leq \mathbb{E}_n\big[w_u\big|f(X)'\tilde{\delta}\big|^2\big]^{3/2}/\mathbb{E}_n\big[w_u\big|f(X)'\tilde{\delta}\big|^3\big]\\
&= \mathbb{E}_n\big[w_u\big|f(X)'\delta\big|^2\big]^{3/2}/\mathbb{E}_n\big[w_u\big|f(X)'\delta\big|^3\big],
\end{aligned}$$

since the scalar $t$ cancels out. Thus, $\mathbb{E}_n[w_u|f(X)'\delta|^3] \leq \mathbb{E}_n[w_u|f(X)'\delta|^2]$. Therefore we have

$$\frac{1}{2}\mathbb{E}_n\big[w_u\big|f(X)'\delta\big|^2\big] - \frac{1}{6}\mathbb{E}_n\big[w_u\big|f(X)'\delta\big|^3\big] \geq \frac{1}{3}\mathbb{E}_n\big[w_u\big|f(X)'\delta\big|^2\big]$$

and

$$\begin{aligned}
&M_u(\theta_u+\delta) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\delta\\
&\quad \geq \frac{1}{3}\mathbb{E}_n\big[w_u\big|f(X)'\delta\big|^2\big] - 2\left\|\frac{\tilde{r}_u}{\sqrt{w_u}}\right\|_{\mathbb{P}_n,2}\big\|\sqrt{w_u}f(X)'\delta\big\|_{\mathbb{P}_n,2},
\end{aligned}$$

which establishes that $F_u(\delta) := M_u(\theta_u+\delta) - M_u(\theta_u) - \partial_\theta M_u(\theta_u)'\delta + 2\|\frac{\tilde{r}_u}{\sqrt{w_u}}\|_{\mathbb{P}_n,2}\|\sqrt{w_u} \times f(X)'\delta\|_{\mathbb{P}_n,2}$ is larger than $\frac{1}{3}\mathbb{E}_n[w_u|f(X)'\delta|^2]$ for any $\delta = t\tilde{\delta}$, $t \in \mathbb{R}$, $\tilde{\delta} \in A_u$, and $\|\sqrt{w_u} \times f(X)'\delta\|_{\mathbb{P}_n,2} \leq \bar{q}_{A_u}$. *Q.E.D.*

LEMMA O.3—Lemma 1 From Bach (2010): *Let $g : \mathbb{R} \to \mathbb{R}$ be a three times differentiable convex function such that, for all $t \in \mathbb{R}$, $|g'''(t)| \leq Mg''(t)$ for some $M \geq 0$. Then, for all $t \geq 0$, we have*

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

LEMMA O.4: *For $t \geq 0$, we have* $\exp(-t) + t - 1 \geq \frac{1}{2}t^2 - \frac{1}{6}t^3$.

PROOF: 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, $f(t) \geq f(0) + tf'(0) = 0$. *Q.E.D.*

LEMMA O.5: *The logistic link function satisfies* $|\Lambda(t+t_0) - \Lambda(t_0)| \leq \Lambda'(t_0)\{\exp(|t|) - 1\}$. *If* $|t| \leq 1$, *we have* $\exp(|t|) - 1 \leq 2|t|$.

PROOF: Note that $|\Lambda''(s)| \leq \Lambda'(s)$ for all $s \in \mathbb{R}$, so that $-1 \leq \frac{d}{ds}\log(\Lambda'(s)) = \frac{\Lambda''(s)}{\Lambda'(s)} \leq 1$. Suppose $s \geq 0$. Therefore,

$$-s \leq \log\big(\Lambda'(s+t_0)\big) - \log\big(\Lambda'(t_0)\big) \leq s.$$

In turn, this implies $\Lambda'(t_0)\exp(-s) \leq \Lambda'(s+t_0) \leq \Lambda'(t_0)\exp(s)$. For $t > 0$, integrating one more time from 0 to $t$,

$$\Lambda'(t_0)\big\{1 - \exp(-t)\big\} \leq \Lambda(t+t_0) - \Lambda(t_0) \leq \Lambda'(t_0)\big\{\exp(t) - 1\big\}.$$

Similarly, for $t < 0$, integrating from $t$ to 0, we have

$$\Lambda'(t_0)\big\{1 - \exp(t)\big\} \leq \Lambda(t+t_0) - \Lambda(t_0) \leq \Lambda'(t_0)\big\{\exp(-t) - 1\big\}.$$

The first result follows by noting that $1 - \exp(-|t|) \leq \exp(|t|) - 1$. The second follows by verification. *Q.E.D.*

## APPENDIX P: SIMULATION EXPERIMENT

In this section, we present results from a brief simulation experiment. The results illustrate the performance of our proposed treatment effect estimator that makes use of estimating equations satisfying the key orthogonality condition given in Equation (1.2) in the main text and variable selection relative to an estimator that uses variable selection but is based on a "naive" estimating equation that does not satisfy the orthogonality condition. We find that inference based on the naive estimator can suffer from substantial size distortions and that the performance of this estimator is strongly dependent on features of the data-generating process (DGP). We also find that tests based on the estimator constructed using our procedure have size close to the nominal level uniformly across all DGPs we consider consistent with the theory developed in the paper.

For simplicity, we consider the case where the treatment, $d_i$, is exogenous conditional on control variables $x_i$. In this case, we can apply the results of the paper substituting $d_i$ for $z_i$ in each instance where instruments $z_i$ are used since $d_i$ is conditionally exogenous and thus a valid instrument for itself. All of the simulation results are based on data generated as

$$d_i = \mathbf{1}\left\{\frac{\exp\{x_i'(c_d\theta_0)\}}{1+\exp\{x_i'(c_d\theta_0)\}} > v_i\right\},$$
$$y_i = d_i\big[x_i'(c_y\theta_0)\big] + \zeta_i,$$

where $v_i \sim U(0,1)$, $\zeta_i \sim N(0,1)$, $v_i$ and $\zeta_i$ are independent, $p = \dim(x_i) = 250$, the covariates $x_i \sim N(0, \Sigma)$ with $\Sigma_{kj} = (0.5)^{|j-k|}$, and the sample size $n = 200$. $\theta_0$ is a $p \times 1$ vector with elements set as $\theta_{0,j} = (1/j)^2$ for $j = 1, \dots, p$. $c_d$ and $c_y$ are scalars that control the strength of the relationship between the controls, the outcome, and the treatment variable. We use several different combinations of $c_d$ and $c_y$, setting $c_d = \sqrt{\frac{(\pi^2/3)R_d^2}{(1-R_d^2)\theta_0'\Sigma\theta_0}}$ and $c_y = \sqrt{\frac{R_d^2}{(1-R_d^2)\theta_0'\Sigma\theta_0}}$ for all combinations of $R_d^2 \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and $R_y^2 \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

We report results for two different inference procedures in Figure S.9. The right panel of the figure shows size of 5% level $t$-tests for the average treatment effect where the point estimate is formed using our proposed estimator based on model selection and orthogonal estimating equations and the standard error is estimated using a plug-in estimator of the asymptotic variance. The left panel shows size of 5% level $t$-tests for the average treatment effect estimated as

$$\hat{\theta}_{\text{naive}} = \frac{1}{n}\sum_{i=1}^{n}\big(\hat{g}_y(1, x_i) - \hat{g}_y(0, x_i)\big),$$

where $\hat{g}_y(d, x_i)$ is a post-model-selection estimator of $\mathrm{E}[Y|D = d, X = x_i]$ and the standard error is estimated using a plug-in estimator of the asymptotic variance of $\hat{\theta}_{\text{naive}}$.

Both procedures rely on post-model-selection estimates of the conditional expectations $\mathrm{E}[Y|D = d, X = x_i]$, and we use exactly the same estimator of this quantity in both cases.

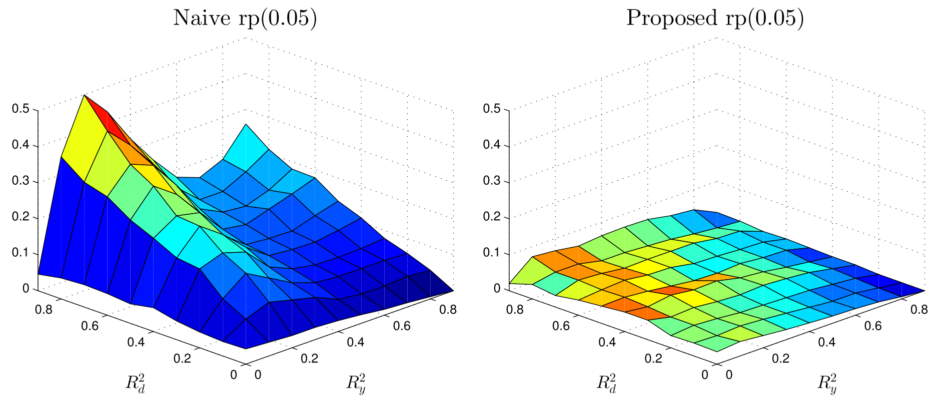


FIGURE S.9.—Rejection frequencies of 5% level tests for average treatment effect estimators following model selection. The left panel shows size of a test based on a "naive" estimator (Naive rp(0.05)), and the right panel shows size of a test based on our proposed procedure (Proposed rp(0.05)).

Specifically, we apply the Square-Root Lasso of Belloni, Chernozhukov, and Wang (2011) with outcome $Y$ and covariates $(D, D * X_1, \ldots, D * X_p, (1-D), (1-D) * X_1, \ldots, (1-D) * X_p)$ to select variables. We set the penalty level in the Square-Root Lasso using the "exact" option of Belloni, Chernozhukov, and Wang (2011) under the assumption of homoscedastic, Gaussian errors $\zeta_i$ with the tuning confidence level required in Belloni, Chernozhukov, and Wang (2011) set equal to 95%. After running the Square-Root Lasso, we then estimate regression coefficients by regressing $Y$ onto only those variables that were estimated to have nonzero coefficients by the Square-Root Lasso. We then form estimates of $\mathrm{E}[Y|D = 1, X = x_i]$ by plugging $(1, x_i')'$ into the estimated model for $i = 1, \ldots, n$ and form estimates of $\mathrm{E}[Y|D = 0, X = x_i]$ by plugging $(0, x_i')'$ into the estimated model for $i = 1, \ldots, n$.

For our proposed method, we also need an estimate of the propensity score. We obtain our estimates of the propensity score by using $\ell_1$-penalized logistic regression with $D$ as the outcome and $X$ as the covariates with penalty level set equal to $0.5\sqrt{n}\Phi^{-1}(1 - 1/2p)/n$, where $\Phi(\cdot)$ is the standard normal distribution function using the MATLAB function "glmlasso."[5] We standardize the variables in $X$ and set penalty loadings equal to 1. After running the $\ell_1$-penalized logistic regression, we estimate the propensity score by taking fitted values from the conventional logistic regression of $D$ onto only those variables that had nonzero estimated coefficients in the $\ell_1$-penalized logistic regression.

Looking at the results, we see that the behavior of the naive testing procedure depends heavily on the underlying coefficient sequence used to generate the data. There are substantial size distortions for many of the coefficient designs considered with good performance, size close to the nominal level, only occurring in a handful of cases. It is worth noting that, in practice, one does not know the underlying DGP and even estimation of the quantities necessary to know where one is in the figure may be infeasible even in this simple scenario. Our proposed procedure does a much better job at delivering accurate inference, producing tests with size close to the nominal level across all designs considered. That is, the simulation illustrates the uniformity derived in the theoretical development of our estimator, illustrating that its performance is relatively good uniformly

[5]This penalty level is equivalent to that discussed in the main paper since "glmlasso" scales the problem in a slightly different way.

across a variety of coefficient sequences. While simply illustrative, these simulation results reinforce the theoretical development of the main paper which proves that our proposed estimation and inference procedures have good properties uniformly across a variety of DGPs where approximate sparsity holds.

## REFERENCES


ABADIE, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263. [1,2]

ANDREWS, D. W. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, Vol. 4. Amsterdam: North-Holland, 2247–2294. [16]

BACH, F. (2010): "Self-Concordant Analysis for Logistic Regression," *Electronic Journal of Statistics*, 4, 384–414. [52]

BELLONI, A., AND V. CHERNOZHUKOV (2011): "$\ell_1$-Penalized Quantile Regression for High Dimensional Sparse Models," *The Annals of Statistics*, 39 (1), 82–130. [7,49]

——— (2013): "Least Squares After Model Selection in High-Dimensional Sparse Models," *Bernoulli*, 19, 521–547. [42,47]

BELLONI, A., V. CHERNOZHUKOV, AND L. WANG (2011): "Square-Root-LASSO: Pivotal Recovery of Sparse Signals via Conic Programming," *Biometrika*, 98, 791–806. [54]

BELLONI, A., V. CHERNOZHUKOV, AND Y. WEI (2013): "Honest Confidence Regions for Logistic Regression With a Large Number of Controls," Preprint. Available at arXiv:1304.3969. [7,8]

BENJAMIN, D. J. (2003): "Does 401(k) Eligibility Increase Saving? Evidence From Propensity Score Subclassification," *Journal of Public Economics*, 87, 1259–1290. [10]

BICKEL, P. J., Y. RITOV, AND A. B. TSYBAKOV (2009): "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37 (4), 1705–1732. [32,34]

CHERNOZHUKOV, V., AND C. HANSEN (2004): "The Impact of 401(k) Participation on the Wealth Distribution: An Instrumental Quantile Regression Analysis," *Review of Economics and Statistics*, 86 (3), 735–751. [10]

GHOSAL, S., A. SEN, AND A. W. VAN DER VAART (2000): "Testing Monotonicity of Regression," *The Annals of Statistics*, 28 (4), 1054–1082. [21,22]

HECKMAN, J., AND E. J. VYTLACIL (1999): "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences of the United States*, 96 (8), 4730–4734. [2]

HONG, H., AND D. NEKIPELOV (2010): "Semiparametric Efficiency in Nonlinear LATE Models," *Quantitative Economics*, 1, 279–304. [2]

IMBENS, G. W., AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. [1,2]

JING, B.-Y., Q.-M. SHAO, AND Q. WANG (2003): "Self-Normalized Cramér-Type Large Deviations for Independent Random Variables," *The Annals of Probability*, 31 (4), 2167–2215. [36]

MAMMEN, E. (1993): "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *The Annals of Statistics*, 21, 255–285. [11,12]

SHERMAN, R. (1994): "Maximal Inequalities for Degenerate $U$-Processes With Applications to Optimization Estimators," *The Annals of Statistics*, 22, 439–459. [22]

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer Series in Statistics. New York: Springer. [22,26,27,31]

VYTLACIL, E. J. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341. [2]



*Fuqua School of Business, Duke University, 100 Fuqua Drive, PO Box 90120, Office W312, Durham, NC 27708, U.S.A.; abn5@duke.edu,*

*Dept. of Economics, MIT, 50 Memorial Drive, E52-361B, Cambridge, MA 02142, U.S.A.; vchern@mit.edu,*

*Dept. of Economics, Boston University, 270 Bay State Road, Room 415A, Boston, MA 02215-1403, U.S.A.; ivanf@bu.edu,*

*and*

*Booth School of Business, University of Chicago, 5807 S. Woodlawn Ave., Chicago, IL 60637, U.S.A.; chansen1@chicagobooth.edu.*

*Co-editor Elie Tamer handled this manuscript.*

*Manuscript received 11 August, 2014; final version accepted 17 January, 2016; available online 24 March, 2016.*