Online Appendix to Locally Robust Semiparametric Estimation

Victor Chernozhukov
MIT

Juan Carlos Escanciano
Universidad Carlos III de Madrid

Hidehiko Ichimura
University of Arizona and University of Tokyo

Whitney K. Newey
MIT and NBER

James M. Robins
Harvard University

November 2021

Abstract

1 Appendix B: Partial Robustness of Plug-In GMM

Partial robustness refers to identifying moments where $E[g(W,\theta_0,\bar{\gamma})] = 0$ for some $\bar{\gamma} \neq \gamma_0$. Partial robustness of identifying moments that are affine in γ with $E[\phi(W,\gamma,\alpha_0,\theta_0)]$ affine in γ can be characterized by the FSIF, since double robustness implies

$$E[g(W,\theta_0,\gamma)] = -E[\phi(W,\gamma,\alpha_0,\theta_0)].$$

We give two examples of partial robustness results that follow from double robustness.

EXAMPLE B1: For a linear functional $\theta_0 = E[m(W,\gamma_0)]$ of a regression function $\gamma_0(X) = E[Y|X]$, let $b(X)$ be a $p\times 1$ vector of functions of X and $\bar{\gamma}(X) = b(X)'\delta$, $\delta = (E[b(X)b(X)'])^{-1}E[b(X)Y]$, be the best linear predictor of $\gamma_0(X)$ by $b(X)$.

THEOREM B1: *If $E[b(X)b(X)']$ is nonsingular and $\alpha_0(X) = \rho_0'b(X)$ for some ρ_0 then $\theta_0 = E[m(W,\bar{\gamma})]$.*

Proof: By orthogonality of the least square projection and by $\alpha_0(X)$ being a linear combination of $b(X)$ it follows that $E[\alpha_0(X)\{Y - \bar{\gamma}(X)\}] = 0$. Then by the proof of Theorem 5 in Appendix A,

$$E[m(W, \bar{\gamma})] - \theta_0 = -E[\alpha_0(X)\{Y - \bar{\gamma}(X)\}] = 0. \ Q.E.D.$$

This result generalizes Stoker's (1986) result that linear regression coefficients equal average derivatives when the regressors are multivariate Gaussian to any linear functional $m(w, \gamma)$ and nonlinear $b(X)$. Stoker's (1986) result can also be extended to instrumental variables.

EXAMPLE B2: Consider the average derivative $\theta_0 = E[\partial\gamma_0(Z)/\partial z_1]$ where $g(w, \gamma, \theta) = \partial\gamma(z)/\partial z_1 - \theta$. Let $\delta = (E[b(X)p(Z)'])^{-1}E[b(X)Y]$ be the limit of the linear instrumental variables estimator with right hand side variables $p(Z)$ and the same number of instruments $b(X)$, and $\bar{\gamma}(Z) = p(Z)'\delta$ the linear instrumental variables estimand.

THEOREM B2: *If* $-\partial \ln f_0(Z)/\partial z_1 = c'p(Z)$ *for a constant vector* c, $E[p(Z)p(Z)']$ *is nonsingular, and* $E[b(X)|Z] = \Pi p(Z)$ *for a square nonsingular* Π *then* $\theta_0 = E[\partial\bar{\gamma}(Z)/\partial z_1]$.

Proof: For $\alpha_0(X) = -c'\Pi^{-1}b(X)$ note that $E[\alpha_0(X)|Z] = -c'\Pi^{-1}\Pi p(Z) = -c'p(Z)$. Then integration by parts gives

$$\begin{aligned} E[g(W, \theta_0, \bar{\gamma})] &= E[c'p(Z)\{\bar{\gamma}(Z) - \gamma_0(Z)\}] = -E[E[\alpha_0(X)|Z]\{\bar{\gamma}(Z) - \gamma_0(Z)\}] \\ &= E[\alpha_0(X)\{Y - \bar{\gamma}(Z)\}] = -c'\Pi^{-1}E[b(X)\{Y - \bar{\gamma}(Z)\}] = 0. \ Q.E.D. \end{aligned}$$

2 Appendix C: Comparing Debiased and Plug-in GMM

To highlight the importance of orthogonal moment functions we compare the properties of debiased GMM with a corresponding cross-fit plug-in GMM estimator

$$\tilde{\theta} = \arg\min_{\theta\in\Theta} \hat{g}(\theta)'\hat{\Upsilon}\hat{g}(\theta).$$

We show that confidence intervals based on the plug-in GMM estimator are invalid with first step model selection and plug-in GMM is not root-n consistent with a Lasso first step. We also show that equation (6.2) of Section 6 must be included among regularity conditions for plug-in GMM.

We show these model selection and regularization problems of plug-in GMM when the parameter of interest is $\theta_0 = E[m(\tilde{W}, \gamma_0)]$, γ_0 is the linear, mean-square projection of an observed variable Y on a set Γ of functions of X that is linear and closed in mean-square, and $\tilde{W}$ is a subvector of W that does not include Y. As noted in Section 5 this object includes many interesting parameters as special cases. We also will assume that γ_0 is a linear combination

$\gamma_0(X) = \breve{b}(X)'\breve{\beta}$ of a vector $\breve{b}(X)$ of functions of X with each $b_j(X) \in \Gamma$. Plug-in GMM will have poor properties more generally but its poor properties here are particularly compelling because it would be good for any estimator to work well when γ_0 is a finite dimensional regression.

There is a general explanation of why standard confidence intervals are asymptotically incorrect for plug-in GMM if the first step $\hat{\gamma}$ incorporates model selection. Models that are selected with probability approaching one (w.p.a.1) for a fixed data generating process must also be selected with w.p.a.1 for any regular, root-n local alternatives, including those where the model is incorrect. This property follows by the contiguity of regular local alternatives, where all events that occur w.p.a.1 for the fixed model occur w.p.a.1 for the local alternative. The selected model being incorrect under a local alternative typically leads to asymptotic bias of a plug-in estimator, giving a limiting distribution with nonzero mean. Consequently the usual asymptotic confidence intervals, that are based on a zero mean limiting distribution, are invalid. The results of Leeb and Potscher (2005) can be explained in this way as is the following result for plug-in GMM estimation of $\theta_0 = E[m(\tilde{W}, \gamma_0)]$.

THEOREM C1: *If with probability approaching one for* $(\ell = 1, ..., L)$, $\hat{\gamma}_\ell(x)$ *is equal to ordinary least squares from regressing* Y_i *on* $\breve{b}(X_i)$ *over* $i \notin I_\ell$; *i)* $\breve{b}_j \in \Gamma$, $(j = 1, ..., J)$, *and* $\gamma_0(x) = \breve{b}(x)'\breve{\beta}$; *ii)* $\breve{Q} =: E[\breve{b}(X)\breve{b}(X)']$ *is nonsingular,* $\breve{b}(X)$ *is bounded, and* $E[Y^2] < \infty$; *iii)* $m(\tilde{W}, \gamma)$ *is linear in* γ, $E[m(\tilde{W}, \gamma)^2] \leq C \|\gamma\|^2$, *and there is bounded* $\alpha_0(X)$ *with* $E[m(\tilde{W}, \gamma)] = E[\alpha_0(X)\gamma(X)]$ *for all* $\gamma \in \Gamma$*; then for* $\breve{\alpha}(x) = \breve{b}(x)'\breve{G}^{-1}E[\breve{b}(X)\alpha_0(X)]$,

$$\sqrt{n}(\tilde{\theta} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \zeta(W_i) + o_p(n^{-1/2}),\ \zeta(W) = m(\tilde{W}, \gamma_0) - \theta_0 + \breve{\alpha}(X)[Y - \gamma_0(X)].$$

Proof: Let $\tilde{\theta}_\ell = \sum_{i \in I_\ell} m(\tilde{W}_i, \hat{\gamma}_\ell)/n_\ell$, $\tilde{M}_\ell = \sum_{i \in I_\ell} m(\tilde{W}_i, \breve{b})/n_\ell$, and $\breve{M} = E[\alpha_0(X)\breve{b}(X)]$ and Note that by i) and ii), $\breve{M}'\breve{\beta} = \theta_0$ and with probability approaching one (w.p.a.1), $\tilde{\theta}_\ell = \tilde{M}_\ell'\hat{\beta}_\ell$ for each ℓ. By ii) and iii) $\tilde{M}_\ell = \breve{M} + O_p(n^{-1/2})$. It also follows in a standard way that $\hat{\beta}_\ell = \breve{\beta} + \breve{Q}^{-1}\sum_{i \notin I_\ell} \breve{b}(X_i)\varepsilon_i/(n - n_\ell) + o_p(n^{-1/2}) = \breve{\beta} + O_p(n^{-1/2})$ for $\varepsilon_i = Y_i - \gamma_0(X_i)$. Then w.p.a.1, for $\bar{m}_\ell = \sum_{i \in I_\ell} m(\tilde{W}_i, \gamma_0)/n_\ell$

$$\begin{aligned}\tilde{\theta}_\ell &= \tilde{M}_\ell'\hat{\beta}_\ell = \tilde{M}_\ell'\breve{\beta} + \breve{M}'(\hat{\beta}_\ell - \breve{\beta}) + (\tilde{M}_\ell - \breve{M})'(\hat{\beta}_\ell - \breve{\beta}) \\ &= \bar{m}_\ell + \breve{M}'\breve{Q}^{-1}\frac{1}{n - n_\ell}\sum_{i \notin I_\ell} \breve{b}(X_i)\varepsilon_i + o_p(n^{-1/2}) = \bar{m}_\ell + \frac{1}{n - n_\ell}\sum_{i \notin I_\ell} \breve{\alpha}(X_i)\varepsilon_i + o_p(n^{-1/2}),\end{aligned}$$

where the last equality follows by $\breve{\alpha}(X) = \breve{M}'\breve{Q}^{-1}\breve{b}(X)$. The conclusion then follows by $\tilde{\theta} = \sum_{\ell=1}^{L}(n_\ell/n)\tilde{\theta}_\ell$ and Lemma 1 of Newey and Robins (2017). *Q.E.D.*

By $\breve{\alpha}(x) = \breve{b}(x)'\breve{G}^{-1}E[\breve{b}(X)\alpha_0(X)]$ it follows that the influence function of $\tilde{\theta}$ is different than the unique influence function of $E_F[m(\tilde{W}, \gamma(F))]$ when $\alpha_0(X)$ is not linear in $\breve{b}(X)$. Then there

will be directions of departure of $\gamma(F)$ from γ_0 that make $\sqrt{n}(\tilde{\theta}-\theta_0)$ be asymptotically biased; see Van der Vaart (1991). The following result displays such directions:

COROLLARY C2: *If the conditions of Theorem C1 are satisfied and the distribution of* Y *conditional on* $(\tilde{W},X)$ *has a pdf* $f_0(y|\tilde{w},x)$ *such that there is* $C>0$ *with* $E[\int\{\sup_{|a|\leq C}\{[df_0(y+a|\tilde{W},X)/da]^2/f_0(y+a|\tilde{W},X)\}\}dy]<\infty$, *then for* $\bar{\sigma}^2=E[\{\alpha_0(X)-\bar{\alpha}(X)\}^2]$, *any* μ, *and* $W_1,...,W_n$ *i.i.d. with CDF* F_n *having conditional pdf* $f_0(\tilde{y}-n^{-1/2}\mu\{\alpha_0(x)-\bar{\alpha}(x)\}|\tilde{w},x)$ *for* Y_i given $(\tilde{W}_i,X_i)=(\tilde{w},z)$ *and CDF* $F_0(\tilde{w},z)$ *for* $(\tilde{W}_i,X_i)$ *we have*

$$\sqrt{n}(\tilde{\theta}-\theta_n)\xrightarrow{d}N(\mu\bar{\sigma}^2,V),$$

where $\theta_n=E_{F_n}[m(\tilde{W},proj_{F_n}(Y|\Gamma))]$ *is the parameter of interest for* F_n.

Proof: For F_n the CDF defined in Corollary C2,

$$E_{F_n}[Y|X]=E[Y|X]+\left(\frac{\mu}{\sqrt{n}}\right)[\bar{\alpha}(x)-\alpha_0(x)].$$

Since the distribution under F_n of $(\tilde{W},X)$ is the same as under F_0 and $\breve{\alpha}(X)$ and $\alpha_0(X)$ are elements of Γ it follows by iterated projections that

$$\begin{aligned}\gamma_n(X)&:=proj_{F_n}(Y|\Gamma)(X)=proj_{F_n}(E_{F_n}[Y|X]|\Gamma)(X)=proj_{F_0}(E_{F_n}[Y|X]|\Gamma)(X)\\&=proj_{F_0}(E[Y|X]|\Gamma)(X)+\left(\frac{\mu}{\sqrt{n}}\right)[\bar{\alpha}(x)-\alpha_0(x)]=\gamma_0(X)+\left(\frac{\mu}{\sqrt{n}}\right)[\bar{\alpha}(x)-\alpha_0(x)].\end{aligned}$$

Note also that by $\breve{\alpha}(X)\in\Gamma$ and iterated projections,

$$\begin{aligned}\theta_n&=E_{F_n}[m(\tilde{W},\gamma_n)]=E[m(\tilde{W},\gamma_n)]=E[\alpha_0(X)\gamma_n(X)]\\&=\theta_0+\left(\frac{\mu}{\sqrt{n}}\right)E[\alpha_0(X)\{\breve{\alpha}(X)-\alpha_0(X)\}]=\theta_0-\left(\frac{\mu}{\sqrt{n}}\right)\bar{\sigma}^2,\\E_{F_n}[\zeta(W)]&=E[\breve{\alpha}(X)\{proj_{F_n}(Y|\Gamma)(X)-\gamma_0(X)\}]=-\left(\frac{\mu}{\sqrt{n}}\right)E[\breve{\alpha}(X)\{\alpha_0(X)-\breve{\alpha}(X)\}]=0.\end{aligned}$$

By hypothesis iv) the conditions of Proposition 1 of Bickel et al. (1993) are satisfied for the parametric model $f_0(y-\delta[\bar{\alpha}(x)-\alpha_0(x)]|x,z)$ with parameter δ. Then by Proposition 3 of Bickel et al. (1993) the sequence of distributions where $W_1,...,W_n$ are i.i.d. with CDF F_n are contiguous to the sequence where $W_1,...,W_n$ are i.i.d. with CDF F_0. Therefore the conclusion of Theorem C1 holds under F_n and

$$\begin{aligned}\sqrt{n}(\tilde{\theta}-\theta_n)&=\sqrt{n}(\hat{\theta}-\theta_0)+\sqrt{n}(\theta_0-\theta_n)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\zeta(W_i)+o_p(1)+\mu\bar{\sigma}^2\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^n\{\zeta(W_i)-E_{F_n}[\zeta(W)]\}+\mu\bar{\sigma}^2\xrightarrow{d}N(\mu\bar{\sigma}^2,V).\ Q.E.D.\end{aligned}$$

Intuitively, model selection for $\hat{\gamma}(x)$ gives a good estimator of $\gamma_0(X)$ but the variables selected for estimating $\gamma_0(X)$ may not include all the variables on which $\alpha_0(X)$ depends, leading to $\breve{\alpha}(X) \neq \alpha_0(X)$ and hence to asymptotic bias One way to avoid the model selection problem for plug-in GMM is to limit selection to models that can approximate any unknown function. For example Newey (1994) did this for series estimation by requiring that selection is made only among models that can approximate any function in large samples. Forcing a flexible approximation in this way is not very feasible in high dimensional settings and is not needed for debiased GMM, where model selection can be applied separately to estimation of α and of γ and standard confidence intervals are correct.

Many machine learners employ regularization to obtain estimators of functions that approximately balance bias and standard deviation. For nonparametric estimation or machine learning with large sets of predictors the standard deviation of the predictor will shrink slower than $1/\sqrt{n}$, and hence so will the bias. This bias may pass through to plug in GMM and result in $\tilde{\theta}$ not being root-n consistent. This bias problem is clearly present for Lasso where penalization leads to bias for $\tilde{\theta}$ of size $\sqrt{\ln(p)/n}$. With bias of that size $\sqrt{n}$ times the bias will be of order $\sqrt{\ln(p)}$ which goes to infinity, so that that plug-in GMM is not root-n consistent, as we show in the next result. Debiased GMM has the small bias property discussed in Section 2 and so will be root-n consistent under sufficient regularity conditions, with bias being second order (size $\ln(p)/n$ for Lasso) permitting debiased GMM to be root-n consistent (by $\sqrt{n}\ln(p)/n \longrightarrow 0$ for Lasso).

THEOREM C3: *If conditions i) - iii) of Theorem C1* are satisfied, $c = E[\alpha_0(X)\breve{b}(X)']\breve{G}^{-1}\breve{e} \neq 0$ *for* $\breve{e} = (sgn(\breve{\beta}_1), ..., sgn(\breve{\beta}_s))'$*; w.p.a.1* $1(\hat{\beta}_{\ell j} = 0) = 1(\beta_j = 0)$ *for all* j; $p \longrightarrow \infty$; $r \longrightarrow 0$ *then*

$$\sqrt{n}\left(\tilde{\theta}^r - \theta_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta(W_i) + \sqrt{n}rc + o_p(1), \quad \left|\sqrt{n}\left(\tilde{\theta}^r - \theta_0\right)\right| \xrightarrow{p} \infty.$$

Proof: Consider the Lasso least squares estimator

$$\hat{\beta}_\ell^r = \arg\min_\beta \frac{1}{n - n_\ell}\sum_{i \notin I_\ell}[Y_i - \breve{b}(X_i)'\beta]^2 + 2r\sum_{j=1}^{s}|\beta_j|, \ \hat{\gamma}_\ell^r(x) = \breve{b}(x)'\hat{\beta}_\ell^r.$$

Note that $\hat{\gamma}_\ell(x) = \hat{\gamma}_\ell^r(x)$ w.p.a.1. Define $\breve{\beta}^r = \breve{\beta} + r\breve{Q}^{-1}\breve{e}$. Then $\hat{\beta}_\ell^r = \hat{\beta}_\ell + r\hat{Q}_\ell^{-1}\breve{e}$ and by standard

arguments and $r \longrightarrow 0$,

$$\begin{aligned}\hat{\beta}_\ell^r &= \breve{\beta}^r + \breve{Q}^{-1}\frac{1}{n-n_\ell}\sum_{i\notin I_\ell}\{\breve{b}(X_i)[Y_i - \breve{b}(X_i)'\breve{\beta}^r + r\breve{e}\} + o_p(n^{-1/2}) \\ &= \breve{\beta}^r + \breve{Q}^{-1}\frac{1}{n-n_\ell}\sum_{i\notin I_\ell}\breve{b}(X_i)[Y_i - \gamma_0(X_i)] + r\breve{Q}^{-1}(\hat{Q}_\ell\breve{Q}^{-1} - I)\breve{e} + o_p(n^{-1/2}) \\ &= \breve{\beta}^r + \breve{Q}^{-1}\frac{1}{n-n_\ell}\sum_{i\notin I_\ell}\breve{b}(X_i)[Y_i - \gamma_0(X_i)] + o_p(n^{-1/2}).\end{aligned}$$

It then follows similarly to the proof of Theorem C1 that

$$\begin{aligned}\tilde{\theta}_\ell^r &= \tilde{M}_\ell'\hat{\beta}_\ell^r = \tilde{M}_\ell'\breve{\beta}^r + \breve{M}'(\hat{\beta}_\ell^r - \breve{\beta}^r) + (\tilde{M}_\ell - \breve{M})'(\hat{\beta}_\ell^r - \breve{\beta}^r) \\ &= \bar{m}_\ell + \frac{1}{n_\ell}\sum_{i\in I_\ell} m(\tilde{W}_i, rb'Q^{-1}\breve{e}) + \breve{M}'\breve{Q}^{-1}\frac{1}{n-n_\ell}\sum_{i\notin I_\ell}\breve{b}(X_i)\varepsilon_i + o_p(n^{-1/2}) \\ &= \bar{m}_\ell + r(\tilde{M} - \breve{M})Q^{-1}\breve{e} + r\breve{M}Q^{-1}\breve{e} + \frac{1}{n-n_\ell}\sum_{i\notin I_\ell}\breve{\alpha}(X_i)\varepsilon_i + o_p(n^{-1/2}).\end{aligned}$$

The conclusion then follows by $r(\tilde{M} - \breve{M})Q^{-1}\breve{e} = o_p(n^{-1/2})$ which follows by $r \longrightarrow 0$. *Q.E.D.*

The hypothesis of this result that $1(\hat{\beta}_{\ell j} = 0) = 1(\beta_j = 0)$ w.p.a.1 is known to hold, for $\gamma_0(X)$ specified here, when all coefficients from regressing each $b_j(X_i)$ on $\breve{b}(X_i)$ are small enough in absolute value; see Zhao and Yu (2006).

Here we see that the Lasso plug-in estimator is not root-n consistent when $\gamma_0(X)$ is a linear combination of a finite number of elements of Γ. In general plug-in GMM will not be root-n consistent with a Lasso first step, though it is beyond the scope of this paper to show this. More generally plug-in GMM will also have large bias for first step machine learners other than Lasso, e.g. as found for random forests in a Monte Carlo example in Chernozhukov et al. (2018).

It is helpful to compare the key asymptotic property of debiased GMM in equation (6.1) with a corresponding key property of plug-in GMM,

$$\hat{g}(\theta_0) = \frac{1}{n}\sum_{i=1}^n \psi(W_i, \gamma_0, \alpha_0, \theta_0) + o_p(n^{-1/2}). \tag{2.1}$$

When equation (2.1) is not satisfied plug-in GMM can have invalid confidence intervals or not be root-n consistent as discussed in connection with equation (4.5). Lemma 8 shows that the corresponding property for debiased GMM is satisfied under general and simple regularity conditions. In contrast, equation (2.1) requires that $\tilde{\phi} := \sum_{\ell=1}^L \sum_{i\in I_\ell}\phi(W_i, \hat{\gamma}_\ell, \alpha_0, \theta_0)/n = o_p(n^{-1/2})$, which is more complicated than Lemma 8 and specific to the first step.

THEOREM C4: *If equation (4.5) is satisfied for $\hat{\alpha}_\ell = \alpha_0$ and $\tilde{\theta}_\ell = \theta_0$ then equation (2.1) is satisfied if and only if $\tilde{\phi} = o_p(n^{-1/2})$.*

Proof: Let $\hat{\psi} = \sum_{i=1}^{n} \psi(W_i, \gamma_0, \alpha_0, \theta_0)/n$ and $\tilde{\phi} = \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \phi(W_i, \hat{\gamma}_\ell, \alpha_0, \theta_0)/n$. Note that

$$\hat{g}(\theta_0) - \hat{\psi} = \left\{ \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \psi(W_i, \hat{\gamma}_\ell, \alpha_0, \theta_0)/n - \hat{\psi} \right\} - \tilde{\phi} = o_p(n^{-1/2}) - \tilde{\phi},$$

by the hypothesis of Theorem C4. Therefore if $\tilde{\phi} = o_p(n^{-1/2})$ then $\hat{g}(\theta_0) - \hat{\psi} = o_p(n^{-1/2})$. Similarly $\tilde{\phi} = o_p(n^{-1/2}) - [\hat{g}(\theta_0) - \hat{\psi}]$, so if $\hat{g}(\theta_0) - \hat{\psi} = o_p(n^{-1/2})$ then $\tilde{\phi} = o_p(n^{-1/2})$. *Q.E.D.*

This result shows that equation (6.2) is an important regularity condition for plug-in GMM, as discussed in Section 6.

3 Appendix D: Convergence Rates for Lasso Minimum Distance

In this Appendix we summarize regularity conditions and convergence rates for a Lasso minimum distance estimator of an object $\alpha_0(X) = a(X)/w(X)$, using a dictionary $(b_1(X), b_2(X), ...)$, a subvector $b(X) = (b_1(X), ..., b_p(X))'$, an estimator $\hat{G}$ of $G = E[w(X)b(X)b(X)']$, and an estimator $\hat{M}$ of $M = E[b(X)a(X)]$. The conditions and results here make use of results given in Chernozhukov, Newey, and Singh (2020) and Bradic et al. (2021). For a matrix $A = [a_{ij}]$ let $|A|_\infty = \max_{i,j} |a_{ij}|$, for a vector b let $|b|_1 = \sum_{j=1}^{p} |b_j|$, and for a measurable function $a(X)$ of X let $\|a\| = \sqrt{E[a(X)^2]}$. We consider an estimator of $\alpha_0(x)$ given by

$$\hat{\alpha}(x) = b(x)'\hat{\rho}, \ \hat{\rho} = \arg\min_{\rho} \{\hat{\rho}'\hat{G}\hat{\rho} - 2\hat{M}'\rho + 2r \sum_{j=1}^{p} |\rho_j|\}.$$

Assumption D1: *There is $C > 0$ such that i) with probability one $\max_{1 \leq j \leq p} |b_j(X)| \leq C$ and $w(X) \geq 1/C$ ii) $\left|\hat{M} - M\right|_\infty = O_p(\varepsilon_n)$; iii) $\varepsilon_n = o(r)$.*

Assumption D2: *i) $\left|\hat{G} - G\right|_\infty = O_p(\varepsilon_n)$ and ii) for every n there is a $p \times 1$ vector ρ_n such that $|\rho_n|_1 \leq C$ and $\|\alpha_0 - b'\rho_n\|^2 = O(\varepsilon_n)$.*

Lemma D1: *If Assumptions D1 and D2 are satisfied then $\|\hat{\alpha} - \alpha_0\| = O_p(\sqrt{r})$.*

Proof: Assumptions D1 and D2 imply Assumptions 1-3 of Chernozhukov, Newey, and Singh (2020, CNS). The conclusion then follows by Theorem 1 of CNS. *Q.E.D.*

This result gives a convergence rate for $\hat{\alpha}_\ell$ of $\sqrt{r}$. We can speed up this convergence rate under a stronger approximate sparsity condition and a sparse eigenvalue condition.

ASSUMPTION D3: *There exists* $C > 1$, $\xi > 0$ *such that for all* $\bar{s}$ *with* $\bar{s} \leq C(\varepsilon_n^2)^{-1/(1+2\xi)}$ *there is* $\bar{\rho}$ *with* $\bar{s}$ *nonzero elements such that* $\|\alpha_0 - b'\bar{\rho}\| \leq C(\bar{s})^{-\xi}$

For any $\rho = (\rho_1, ..., \rho_p)$ let $\mathcal{J} = \{1, ..., p\}$, $\mathcal{J}_\rho$ be the subset of $\mathcal{J}$ with $\rho_j \neq 0$, and $\mathcal{J}_\rho^c$ be the complement of $\mathcal{J}_\rho$ in $\mathcal{J}$ and let $\rho_L = \arg\min_\rho\{\|\alpha_0 - b'\rho\|^2 + 2\varepsilon_n |\rho|_1\}$ be the population Lasso coefficients for penalty $2\varepsilon_n$. The next condition is a sparse eigenvalue condition for the population matrix G.

ASSUMPTION D4: G *is nonsingular and has largest eigenvalue uniformly bounded in* n. *Also there is* $k > 3$ *such such that for* $\rho = \rho_L$,

$$\inf_{\{\delta:\delta\neq 0, \sum_{j\in\mathcal{J}_{\rho_L}^c} |\delta_j| \leq k\sum_{j\in\mathcal{J}_{\rho_L}} |\delta_j|\}} \frac{\delta' G \delta}{\sum_{j\in\mathcal{J}_{\rho_L}} \delta_j^2} > 0.$$

LEMMA D2: *If Assumptions D1-D4 are satisfied then*

$$\|\hat{\alpha} - \alpha_0\| = O_p((\varepsilon_n)^{\frac{-1}{1+2\xi}} r).$$

Proof: Assumptions D1-D4 imply Assumptions 1-5 of CNS, so the conclusion follows by Theorem 3 of CNS. *Q.E.D.*

The convergence rate here is faster than that of Lemma D1 when r is not too much larger than ε_n. For Theorem 12 it is useful to have a uniform convergence rate for a Lasso $\hat{\gamma}$. Let J denote a subset of $\{1, ..., p\}$, ρ_J be the vector consisting of $\rho_{Jj} = \rho_j$ for $j \in J$ and $\rho_{Jj} = \rho_j = 0$ otherwise, and ρ_{J^c} be the corresponding vector for J^c.

ASSUMPTION D5: $\hat{G} = \sum_{i=1}^n b(X_i)b(X_i)'/n$, $G = E[b(X)b(X)']$ *has largest eigenvalue bounded uniformly in* n, *and there is* $C, c > 0$ *such that for all* $s \approx C\varepsilon_n^{-2}$ *with probability approaching one*

$$\min_{J\leq s} \min_{\|\rho_{J^c}\|_1 \leq 3\|\rho_J\|_1} \frac{\rho'\hat{G}\rho}{\rho_J'\rho_J} \geq c.$$

This is a sample sparse eigenvalue condition, e.g. see Bickel, Ritov, Tsybakov (2009).

LEMMA D3: *If Assumptions D1, D3, and D5 are satisfied,* $r = o(n^c\varepsilon_n)$ *for all* $c > 0$, *and there exists* $C > 0$ *such that* $p \leq Cn^C$, *then for all* $c > 0$,

$$\|\hat{\alpha} - \alpha_0\| = o_p(n^c\varepsilon_n^{2\xi/(2\xi+1)}).$$

If in addition $\xi > 1/2$, $\alpha_0(x) = \sum_{j=1}^\infty \rho_{j0}b_j(x)$, *and* $\sum_{j>p} |\rho_{j0}| \leq C\varepsilon_n^{(2\xi-1)/(2\xi+1)}$ *then for all* $c > 0$,

$$\sup_x |\hat{\alpha}(x) - \alpha_0(x)| = o_p(n^c\varepsilon_n^{(2\xi-1)/(2\xi+1)}).$$

Proof: The first conclusion follows by Lemmas A2 and A7 of Bradic et al. (2021) and boundedness of the largest eigenvalue of G. Also, by uniform boundedness of $b_j(x)$, Lemmas A4 and A7 of Bradic et al. (2021), and the triangle inequality,

$$|\hat{\alpha}(x) - \alpha_0(x)| \leq C \sum_{j=1}^{p} |\hat{\rho}_j - \rho_{j0}| + C \sum_{j>p} |\rho_{j0}| = O_p(\varepsilon_n^{(2\xi-1)/(2\xi+1)}) = o_p(n^c \varepsilon_n^{(2\xi-1)/(2\xi+1)}).Q.E.D.$$

4 Appendix E: Proofs of Theorems 9-12.

Theorem 9 will follow from Lemma 8 and two useful Lemmas. Let $\Psi := E[\psi(W, \gamma_0, \alpha_0, \theta_0)\psi(W, \gamma_0, \alpha_0, \theta_0)']$.

Lemma E1: If Assumptions 1 and 4 are satisfied then $\hat{\Psi} \xrightarrow{p} \Psi$.

Proof: Define the remainders $\hat{R}_{1\ell i}$, $\hat{R}_{2\ell i}$, $\hat{R}_{3\ell i}$, and $\hat{\Delta}_\ell(W_i)$ as in the proof of Lemma 8. Also, define $\hat{R}_{4\ell i} = g(W_i, \hat{\gamma}_\ell, \tilde{\theta}_\ell) - g(W_i, \hat{\gamma}_\ell, \theta_0)$. By Assumption 4

$$E[\left\|\hat{R}_{4\ell i}\right\|^2 |\mathcal{W}_\ell^c] \xrightarrow{p} 0.$$

It similarly follows from Assumption 1 and $\int \left\|\hat{\Delta}_\ell(w)\right\|^2 F_0(dw) \xrightarrow{p} 0$ that for $i \in I_\ell$,

$$E[\left\|\hat{R}_{k\ell i}\right\|^2 |\mathcal{W}_\ell^c] \xrightarrow{p} 0,\ k = 1, 2, 3,\ E[\left\|\hat{\Delta}_\ell(W_i)\right\|^2 |\mathcal{W}_\ell^c] \xrightarrow{p} 0.$$

Then it follows that for $\psi_i = \psi(W_i, \gamma_0, \alpha_0, \theta_0)$,

$$E[\frac{1}{n}\sum_{i\in I_\ell} \left\|\hat{\psi}_{i\ell} - \psi_i\right\|^2 |\mathcal{W}_\ell^c] \leq \frac{Cn_\ell}{n}\left(\sum_{k=1}^{4} E[\left\|\hat{R}_{k\ell i}\right\|^2 |\mathcal{W}_\ell^c] + E[\left\|\hat{\Delta}_\ell(W_i)\right\|^2 |\mathcal{W}_\ell^c]\right) \xrightarrow{p} 0.$$

Therefore $\sum_{i\in I_\ell} \left\|\hat{\psi}_{i\ell} - \psi_i\right\|^2 / n \xrightarrow{p} 0$ by the conditional Markov inequality. It follows by the triangle inequality that for $\tilde{\Psi} := \sum_{i=1}^{n} \psi_i \psi_i'/n$,

$$\begin{aligned}\left\|\hat{\Psi} - \tilde{\Psi}\right\| &\leq \sum_{\ell=1}^{L} \frac{1}{n} \sum_{i\in I_\ell} (\left\|\hat{\psi}_{i\ell} - \psi_i\right\|^2 + 2\left\|\psi_i\right\| \left\|\hat{\psi}_{i\ell} - \psi_i\right\|) \\ &\leq o_p(1) + 2\sum_{\ell=1}^{L} \sqrt{\frac{1}{n}\sum_{i\in I_\ell} \left\|\hat{\psi}_{i\ell} - \psi_i\right\|^2} \sqrt{\frac{1}{n}\sum_{i\in I_\ell} \left\|\psi_i\right\|^2} = o_p(1)(1 + O_p(1)) \xrightarrow{p} 0.\end{aligned}$$

We also have $\tilde{\Psi} \xrightarrow{p} \Psi$ by Khintchine's law of large numbers, so the conclusion follows by the triangle inequality. $Q.E.D.$

LEMMA E2: *If Assumption 5 is satisfied and* $\bar{\theta} \xrightarrow{p} \theta_0$ *then* $\partial \hat{g}(\bar{\theta})/\partial\theta \xrightarrow{p} G$.

Proof: Let $\hat{G}_\ell = n_\ell^{-1} \sum_{i \in I_\ell} \partial g(W_i, \hat{\gamma}_\ell, \bar{\theta})/\partial\theta$ and $\tilde{G}_\ell = n_\ell^{-1} \sum_{i \in I_\ell} \partial g(W_i, \hat{\gamma}_\ell, \theta_0)/\partial\theta$. By ii) and

$$E[\frac{1}{n_\ell} \sum_{i \in I_\ell} d(W_i, \hat{\gamma}_\ell) | \mathcal{W}_\ell^c] = E[d(W_i, \hat{\gamma}_\ell) | \mathcal{W}_\ell^c] \leq C,$$

with probability approaching one. Then by the conditional Markov inequality, $\sum_{i \in I_\ell} d(W_i, \hat{\gamma}_\ell) = O_p(1)$. Then by conditions i) and ii) and the triangle inequality, with probability approaching one

$$\left\| \hat{G}_\ell - \tilde{G}_\ell \right\| \leq n_\ell^{-1} \sum_{i \in I_\ell} d(W_i, \hat{\gamma}_\ell) \left\| \bar{\theta} - \theta_0 \right\|^{1/C} = O_p(1) o_p(1) \xrightarrow{p} 0.$$

Then $\hat{G}_\ell - \tilde{G}_\ell \xrightarrow{p} 0$ follows by the conditional Markov inequality. For $\bar{G}_\ell = n_\ell^{-1} \sum_{i \in I_\ell} \partial g(W_i, \gamma_0, \theta_0)/\partial\theta$ it follows similarly from condition iii) that $\tilde{G}_\ell - \bar{G}_\ell \xrightarrow{p} 0$. By Khintchine's law of large numbers $\bar{G}_\ell \xrightarrow{p} G$, so the conclusion follows by the triangle inequality. $Q.E.D.$

Proof of Theorem 9: Follows in a standard way from Lemmas 8, E1, and E2. $Q.E.D.$

Proof of Theorem 10: Let $g(w, \gamma, \theta) = m(w, \gamma) - \theta$ and $\phi(w, \gamma, \alpha, \theta) = \alpha(x)\lambda(w, \gamma)$. Assumption 1 is satisfied by conditions ii) and iii). Also, by i) $\int \hat{\alpha}_\ell(x)\lambda(w, \gamma_0) F_0(dw) = 0$ with probability approaching one so Assumption 3 is satisfied by conditions i) and v). Also, note that $\hat{\Delta}_\ell(w) = [\hat{\alpha}_\ell(x) - \alpha_0(x)][\lambda(w, \hat{\gamma}_\ell) - \lambda(w, \gamma_0)]$. If condition iv) a) is satisfied then

$$\int \hat{\Delta}_\ell(w)^2 F_0(dw) = \int [\hat{\alpha}_\ell(x) - \alpha_0(x)]^2 [\lambda(w, \hat{\gamma}_\ell) - \lambda(w, \gamma_0)]^2 F_0(dw) \xrightarrow{p} 0$$

and by iterated expectations and the Cauchy-Schwartz inequality

$$\begin{aligned}
\left| \sqrt{n} \int \hat{\Delta}_\ell(w) F_0(dw) \right| &= \sqrt{n} \left| \int [\hat{\alpha}_\ell(x) - \alpha_0(x)][\lambda(w, \hat{\gamma}_\ell) - \lambda(w, \gamma_0)] F_0(dw) \right| \\
&= \sqrt{n} \left| \int [\hat{\alpha}_\ell(x) - \alpha_0(x)][\bar{\lambda}(x, \hat{\gamma}_\ell) - \bar{\lambda}(x, \gamma_0)] F_0(dx) \right| \\
&\leq \sqrt{n} \left\| \hat{\alpha}_\ell - \alpha_0 \right\| \left\| \bar{\lambda}(\hat{\gamma}_\ell) - \bar{\lambda}(\gamma_0) \right\| \xrightarrow{p} 0,
\end{aligned}$$

so that Assumptions 2 i) and 4 are satisfied, giving the conclusion. If condition iv) b) is satisfied then by the Cauchy-Schwartz and conditional Markov inequalities,

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{i \in I_\ell} \left\| \hat{\Delta}_\ell(w) \right\| &\leq \sqrt{n} \left(\frac{1}{n} \sum_{i \in I_\ell} [\hat{\alpha}_\ell(X_i) - \alpha_0(X_i)]^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i \in I_\ell} [\lambda(W_i, \hat{\gamma}_\ell) - \lambda(W_i, \gamma_0)]^2 \right)^{1/2} \\
&= \sqrt{n} O_p(\|\hat{\alpha}_\ell - \alpha_0\|) O_p(\|\lambda(\hat{\gamma}_\ell) - \lambda(\gamma_0)\|) = o_p(1),
\end{aligned}$$

so Assumption 2 ii) is satisfied, giving all the conditions of Lemma 8. Also it is straightforward to show that $\|\bar{\alpha}_\ell - \alpha_0\| \xrightarrow{p} 0$ and by the conditional Markov inequality,

$$\int \hat{\Delta}_\ell(w)^2 F_0(dw) \leq (M^2 + C) \left\| \lambda(\hat{\gamma}_\ell) - \lambda(\gamma_0) \right\|^2 \xrightarrow{p} 0,$$

so Assumption 4 is also satisfied. The conclusion then follows from Lemmas 8 and E1. *Q.E.D.*

The following result is useful for showing that Assumption 3 is satisfied for Example 2.

LEMMA E3: *If Assumption 6 is satisfied then for Example 2* $\left|\bar{\psi}(\gamma,\alpha_0,\theta_0)\right| \leq C\left\|\gamma-\gamma_0\right\|^2$.

Proof: Let $\lambda(W,\gamma(X)) = 1(Y<\gamma(X))-\zeta$, $U = Y-\gamma_0(X)$, and $\Delta(X) = \gamma(X)-\gamma_0(X)$. By eq. (2.9) $E[\alpha_0(X)\lambda(W,\gamma_0(X))] = 0$. It follows by the definition of $\alpha_0(X)$ and the orthogonality for a projection that for any $b \in \Gamma$,

$$E[v_m(X)b(X)] + E[f(0|X)\alpha_0(X)b(X)] = -E[f(0|X)\{-f(0|X)^{-1}v_m(X) - \alpha_0(X)\}b(X)] = 0.$$

Assumption 6 and Taylor expansion with LaGrange remainder give

$$\begin{aligned} E[\lambda(Y,\gamma(X)) - \lambda(Y,\gamma_0(X))|X] &= E[1(U<\Delta(X)) - 1(U<0)|X] = \int_0^{\Delta(X)} f(u|X)du \\ &= f(0|X)\Delta(X) + [\partial f(\delta(X)|X)/\partial u]\Delta(X)^2, \end{aligned}$$

where $\delta(X)$ is between $\Delta(X)$ and zero. Therefore, by $\theta_0 = E[v_m(X)\gamma_0(X)]$,

$$\begin{aligned} \left|\bar{\psi}(\gamma,\alpha_0,\theta_0)\right| &= |E[v_m(X)\Delta(X) + \alpha_0(X)\{\lambda(W,\gamma(X)) - \lambda(W,\gamma_0(X))\}]| \\ &= |E[v_m(X)\Delta(X) + E[f(0|X)\alpha_0(X)\Delta(X)] + E[\{\partial f(\delta(X)|X)/\partial u\}\Delta(X)^2]| \\ &= \left|E[\{\partial f(\delta(X)|X)/\partial u\}\Delta(X)^2]\right| \leq C\left\|\Delta\right\|^2 = C\left\|\gamma-\gamma_0\right\|^2. \; Q.E.D. \end{aligned}$$

The following result gives a convergence rate for the $\hat{Q}_\ell$ of Example 2. Let $\varepsilon_n = \sqrt{\ln(p)/(hn)} + h^2 + n^{-d_\gamma}$.

LEMMA E4: *If Assumption 7 is satisfied there is C such that $|b_j(X)| \leq C$ for all j then for* $Q = -E[f(0|X)b(X)b(X)']$,

$$\left|\hat{Q}_\ell - Q\right|_\infty = O_p(\varepsilon_n), \;\; \left|\hat{M}_\ell - M\right|_\infty = O_p(\varepsilon_n).$$

Proof: Consider

$$\hat{Q}_{\ell\ell'} := \frac{-1}{n_{\ell'}}\sum_{i\in I_{\ell'}} \frac{1}{h}K\left(\frac{Y_i - \hat{\gamma}_{\ell\ell'}(X_i)}{h}\right)b(X_i)b(X_i)'.$$

For notational convenience we drop the ℓ and ℓ' subscripts and replace $\sum_{i\in I_{\ell'}}$ with $\sum_{i=1}^n$ while retaining independence of $\hat{\gamma}$ from the data being averaged over. Let $K_h(u) = h^{-1}K(u/h)$, $f(u|x)$ denote the conditional pdf of $U = Y-\gamma_0(X)$ given $X = x$, and $\hat{\Delta}(X) = \gamma_0(X) - \hat{\gamma}(X)$. Note by two change of variables $v = (u+\hat{\Delta})/h$, and $\tilde{v} = u/h$

$$\begin{aligned} |E[K_h(Y-\hat{\gamma}(X))|X] - E[K_h(U)|X]| &= \left|\int K_h(u-\hat{\Delta}(X))f(u|X)du - \int K_h(u)f(u|X)du\right| \\ &\leq \int |K(v)|\left|f(hv+\hat{\Delta}(X)|X) - f(hv|X)\right|dv \leq C\hat{\Delta}(X). \end{aligned}$$

Also, note that by a mean value expansion for small enough h,

$$f(hv|X) = \sum_{k=0}^{1} v^k h^k \frac{d^k f(0|X)}{du^k} + h^2 R(v, X),$$
$$|E[K_h(U)|X] - f(0|X)| = \left| \int K(v)[f(hv|X) - f(0|X)]dv \right| \leq Ch^2$$

Note also that conditional on $\hat{\gamma}$, by the conditional Markov inequality and $\int \left|\hat{\Delta}(x)\right| F_0(x) \leq \|\hat{\gamma} - \gamma_0\| = O_p(n^{-d_\gamma})$ we have $\sum_{i=1}^n \left|\hat{\Delta}(X_i)\right| /n = O_p(n^{-d_\gamma})$. Therefore by $|b(X_i)|_\infty \leq C$ we have

$$\left| \frac{1}{n} \sum_{i=1}^n b(X_i) b(X_i)' \{E[K_h(Y_i - \hat{\gamma}(X_i))|X_i] - f(0|X_i)\} \right|_\infty$$
$$\leq C \frac{1}{n} \sum_{i=1}^n \left|\hat{\Delta}(X_i)\right| + Ch^2 = O_p(n^{-d_\gamma} + h^2),$$

Note also that by a change of variable $v = (U + \hat{\Delta}(X))/h$,

$$|b_j(X) b_{j'}(X) K_h(Y - \hat{\gamma}(X_i))| \leq Ch^{-1},$$
$$E[b_j(X)^2 b_{j'}(X)^2 K_h(Y - \hat{\gamma}(X))^2] \leq CE[K_h(U + \hat{\Delta}(X))^2] \leq Ch^{-1} E[\int K(v)^2 f(hv - \hat{\Delta}(X)|X)] \leq Ch^{-1}.$$

It then follows by Lemma 19.32 of Van der Vaart (1998) and a standard argument that

$$\left| \hat{Q} + \frac{1}{n} \sum_{i=1}^n b(X_i) b(X_i)' E[K_h(Y_i - \hat{\gamma}(X_i))|X_i] \right|_\infty = O_p(\sqrt{\frac{\ln(p)}{hn}}).$$

Then by the triangle inequality $\left|\hat{Q} + \sum_{i=1}^n b(X_i) b(X_i)' f(0|X_i)/n\right|_\infty = O_p(\varepsilon_n)$. In addition it follows by a standard application of Hoeffding's inequality that

$$\left| \frac{-1}{n} \sum_{i=1}^n b(X_i) b(X_i)' f(0|X_i) - Q \right|_\infty = O_p(\sqrt{\frac{\ln(p)}{n}}) = O_p(\varepsilon_n),$$

so $\left|\hat{Q} - Q\right|_\infty = O_p(\varepsilon_n)$ follows by the triangle inequality. The first conclusion follows by another application of the triangle inequality. The second conclusion follows by Assumption 7 vi) and another application of Hoeffding's inequality since $\sqrt{\ln(p)/n} \leq \varepsilon_n$ for n large enough. *Q.E.D.*

The next result gives a convergence rates for the $\hat{\alpha}_\ell$ of Example 2, using the conditions of Appendix D.

LEMMA E5: *If Assumptions 7, D1, and D2 are satisfied then* $\|\hat{\alpha}_\ell - \alpha_0\| = O_p(\sqrt{r})$. *If Assumption D3 and D4 are also satisfied then for the sparse approximation rate* $\xi \geq 1/2$ *from Assumption D3 we have* $\|\hat{\alpha}_\ell - \alpha_0\| = O_p(r^{2\xi/(1+2\xi)})$.

Proof: Note that $-M_j = -E[v_m(X)b_j(X)] = E[f(0|X)\alpha_0(X)b_j(X)]$ and $\|a\|^2 \leq CE[f(0|X)a(W)^2]$. Then the conclusion follows by Lemmas D1 and D2. $Q.E.D.$

Proof of Theorem 11: We proceed by showing that each of the conditions of Theorem 10 are satisfied for $\lambda(W,\gamma) = 1(Y < \gamma(X)) - \zeta$, $\gamma_0(X) = \arg\min_{\gamma\in\Gamma} E[v(Y-\gamma(X))]$ from Section 2.1, and α_0 given before Theorem 11. It follows similar to the proof of Lemma E3 that for any $b \in \Gamma$ the first order condition

$$0 = \frac{d}{dt}E[v(Y - \gamma_0(X) - tb(X))] = E[\lambda(W,\gamma_0)b(X)]$$

are satisfied. Also, $\hat{\alpha}_\ell(X)$ is a linear combination of elements of Γ, so condition i) of Theorem 10 is satisfied. Condition ii) of Theorem 10 holds by hypothesis ii), vi), and $\lambda(W,\gamma)$ bounded. Condition iii) of Theorem 10 is satisfied by hypothesis ii),

$$\begin{aligned}\int [\lambda(w,\hat{\gamma}_\ell) - \lambda(w,\gamma_0)]^2 F_0(dw) &\leq \int 1(|u| \leq |\hat{\gamma}_\ell(x) - \gamma_0(x)|) f(u|x) F_0(dx) \\ &\leq C \int |\hat{\gamma}_\ell(x) - \gamma_0(x)|\, F_0(dx) \leq C\,\|\hat{\gamma}_\ell - \gamma_0\| \xrightarrow{p} 0,\end{aligned}$$

and Lemma E5. To see that condition iv) a) of Theorem 10 is satisfied, note that by $\lambda(w,\gamma)$ bounded,

$$\int [\hat{\alpha}_\ell(x) - \alpha_0(x)]^2 [\lambda(w,\hat{\gamma}_\ell) - \lambda(w,\gamma_0)]^2 F_0(dw) \leq C \int [\hat{\alpha}_\ell(x) - \alpha_0(x)]^2 F_0(dw) \xrightarrow{p} 0.$$

Also, similarly to the proof of Lemma E3,

$$\begin{aligned}\left\|\bar{\lambda}(\hat{\gamma}_\ell) - \bar{\lambda}(\gamma_0)\right\|^2 &= \int [\int_{-\infty}^{\hat{\gamma}(x)-\gamma_0(x)} f(u|x)du - \int_{-\infty}^{0} f(u|x)du]^2 F_0(dx) \\ &\leq \int C|\hat{\gamma}(x) - \gamma_0(x)|^2 F_0(dx) = C\,\|\hat{\gamma} - \gamma_0\|^2.\end{aligned}$$

Then condition iv a) of Theorem 10 follows by Lemma E5 and either hypothesis iv) or v). Finally, condition v) of Theorem 10 is satisfied by Lemma E3 and hypothesis iii). $Q.E.D.$

Next we give results useful in the proof of Theorem 12. We first give a result on uniform convergence of the Lasso estimator of the value function difference γ_{20} described in Section 3. This result may be useful more generally for dynamic structural models to provide a machine learner of expected value function differences.

LEMMA E6: *If Assumption 8 is satisfied then for* $\gamma_{20}(X_t) = E[H(\gamma_{10}(X_{t+1}))|X_t]$ and each ℓ

$$\begin{aligned}\sup_x |\hat{\gamma}_{2\ell}(x) - \gamma_{20}(x)| &= O_p(n^{-d_1(2\xi_1-1)/(2\xi_1+1)}\ln(n)),\ \ |\hat{\gamma}_{3\ell} - \gamma_{30}| = O_p(n^{-d_1}), \\ \|\hat{\gamma}_{2\ell} - \gamma_{20}\| &= O_p(n^{-d_1 2\xi_1/(2\xi_1+1)}\ln(n)),\ \ (\ell = 1,...,L).\end{aligned}$$

Proof: Let $Q_2 = E[b(X)b(X)'\gamma_{10}(X)]$. Let

$$\tilde{M}_{2\ell} = \frac{1}{(n-n_\ell)T}\sum_{\ell'\neq\ell}\sum_{i\in I_{\ell'}}\sum_{t=1}^{T} Y_{2it}b(X_{it})H\left(\gamma_{10}(X_{i,t+1})\right).$$

and $\mathcal{W}_\ell^c$ denote all observations not in I_ℓ. It follows by Assumption 8 and the Cauchy-Schwartz inequalities that

$$\frac{1}{n_{\ell'}T}E[\sum_{i\in I_{\ell'}}\sum_{t=1}^{T}|\hat{\gamma}_{1\ell\ell'}(X_{i,t+1})-\gamma_{10}(X_{i,t+1})|\,|\mathcal{W}_\ell^c] = \int|\hat{\gamma}_{1\ell\ell'}(x)-\gamma_{10}(x)|\,F_0(dx) \leq \|\hat{\gamma}_{1\ell\ell'}-\gamma_{10}\| = O_p(n^{-d_1}).$$

Then by $H(p)$ having bounded derivative on $[\varepsilon, 1-\varepsilon]$, Assumption 8, and the conditional Markov inequality

$$\left|\hat{M}_{2\ell}-\tilde{M}_{2\ell}\right|_\infty \leq \frac{C}{(n-n_\ell)}\sum_{\ell'\neq\ell}\sum_{i\in I_{\ell'}}\sum_{t=1}^{T}|\hat{\gamma}_{1\ell\ell'}(X_{i,t+1})-\gamma_{10}(X_{i,t+1})| = O_p(n^{-d_1}). \tag{4.1}$$

For $M_2 = E[Y_{2t}b(X_t)H(\gamma_{10}(X_{t+1}))]$ it follows by a standard maximal inequality that $\left|\tilde{M}_{2\ell}-M_2\right|_\infty = O_p(\sqrt{\ln(p)/n})$. Then by the triangle inequality we have $\left|\hat{M}_{2\ell}-M_2\right|_\infty = O_p(n^{-d_1})$. The first conclusion then follows by Lemma D3.

For the second conclusion let

$$\tilde{\gamma}_{3\ell} := \frac{1}{\hat{P}_{1\ell}(n-n_\ell)T}\sum_{\ell'\neq\ell}\sum_{i\in I_{\ell'}}\sum_{t=1}^{T} Y_{1it}H\left(\gamma_{10}(X_{i,t+1})\right).$$

It follows similarly to equation (4.1) that $|\hat{\gamma}_{3\ell}-\tilde{\gamma}_{3\ell}| = O_p(n^{-d_1})$. Also by standard arguments $|\tilde{\gamma}_{3\ell}-\gamma_{30}| = O_p(1/\sqrt{n}) = O_p(n^{-d_1})$, so the second conclusion follows by the triangle inequality.

The third conclusion follows from Lemma D3. *Q.E.D.*

The following Lemma gives a convergence rate for the preliminary plug-in $\tilde{\theta}_\ell$.

LEMMA E7: *If Assumption 8 and the hypotheses of Theorem 12 are satisfied then*

$$\tilde{\theta}_\ell = \theta_0 + O_p(n^{-d_1[(2\xi_1-1)/(2\xi_1+1)]}).$$

Proof: Follows from Lemma E6 by standard for quasi maximum likelihood with $\hat{\gamma}_2(x)$ and $\hat{\gamma}_3$ plugged-in. *Q.E.D.*

LEMMA E8: *If Assumption 8 and the hypotheses of Theorem 12 are satisfied then*

$$\|\hat{\alpha}_{1\ell}-\alpha_{10}\| = O_p(n^{-d_1[(2\xi_1-1)/(2\xi_1+1)]2\xi_2/[2\xi_2+1]}[\ln(n)]^2 + n^{-\xi_3/(2\xi_3+1)}\ln(n) + n^{-d_1}),$$
$$\|\hat{\alpha}_{2\ell}-\alpha_{20}\| = O_p(n^{-d_1[(2\xi_1-1)/(2\xi_1+1)]}\ln(n)),\ \ \|\hat{\alpha}_{3\ell}-\alpha_{30}\| = O_p(n^{-d_1(2\xi_1-1)/(2\xi_1+1}),$$
$$\|H(\hat{\gamma}_{1\ell})-H(\gamma_{10})\| = O_p(n^{-d_1}).$$

Proof: First, note that $a(x) = a(x, \theta_0, \gamma_{20}, \gamma_{30})$ is bounded by $D(x)$ bounded and $H(\gamma_{10}(x))$ is bounded by $\gamma_{10}(x) \in (\varepsilon, 1-\varepsilon)$, so that by Assumption 8 and the fixed trimming, with $\Lambda(a) > 0$ and twice continuous differentiability of $\Lambda(a)$,

$$\sup_x |\hat{\alpha}_{2\ell k}(x, y_2) - \alpha_{02k}(x, y_2)| \leq C \sup_x |\hat{a}(x) - a(x)| \leq C(\left\|\tilde{\theta}_\ell - \theta_0\right\| + \sup_x |\hat{\gamma}_{2\ell}(x) - \gamma_{20}(x)|$$
$$+ |\hat{\gamma}_{3\ell} - \gamma_{30}|) = O_p(n^{-d_1(2\xi_1-1)/(2\xi_1+1)}), \quad a(x, \tilde{\theta}_\ell, \hat{\gamma}_{2\ell}, \hat{\gamma}_{3\ell}) = \hat{a}(x),$$

giving the second conclusion. The third conclusion follows in a standard way.

Let $\hat{\zeta}_{1\ell k}(x)$ denote Lasso with regressors $b(X_{i,t+1})$, dependent variable equal to the k_{th} element $\hat{\alpha}_{2\ell itk}$ of $\hat{\alpha}_{2\ell it} = \hat{\alpha}_{2\ell}(X_{it}, Y_{2it})$, and penalization r_2 for $i \notin I_\ell$. This estimator has

$$\hat{M}_\ell = \frac{1}{(n-n_\ell)T} \sum_{i \notin I_\ell} \sum_{t=1}^{T} \hat{\alpha}_{2\ell itk} b(X_{i,t+1}), \ \hat{Q}_\ell = \frac{1}{(n-n_\ell)T} \sum_{i \notin I_\ell} \sum_{t=1}^{T} b(X_{i,t+1}) b(X_{i,t+1})',$$

and r_2 replacing r_1. By Lemma E6, Lemma E7, uniform boundedness of the elements of $b(x)$, and Bernstein's inequality, for $M = E[\alpha_{20k}(X_{it}, Y_{2it}) b(X_{i,t+1})]$,

$$\left|\hat{M}_\ell - \tilde{M}_\ell\right|_\infty \leq C \sup_x |\hat{\alpha}_{2\ell k}(x, y_2) - \alpha_{02k}(x, y_2)| = O_p(n^{-d_1(2\xi_1-1)/(2\xi_1+1)} \ln(n)),$$

$$\left|\tilde{M}_\ell - M\right|_\infty = O_p(\sqrt{\ln(p)/n}), \ \tilde{M}_\ell := \frac{1}{(n-n_\ell)T} \sum_{i \notin I_\ell} \sum_{t=1}^{T} \alpha_{20k}(X_{it}, Y_{2it}) b(X_{i,t+1}).$$

Then by the triangle inequality and another application of Bernstein's inequality,

$$\left|\hat{M}_\ell - M\right|_\infty = O_p(n^{-d_1(2\xi_1-1)/(2\xi_1+1)} \ln(n)).$$

Then by Lemma D3,

$$\left\|\hat{\zeta}_{1\ell} - \zeta_{10}\right\| = O_p(n^{-d_1[(2\xi_1-1)/(2\xi_1+1)]2\xi_2/[2\xi_2+1]} [\ln(n)]^2).$$

It follows similarly that for $\hat{\zeta}_{2\ell}(x)$ denoting Lasso with regressors $b(X_{i,t+1})$ and dependent variable Y_{1it},

$$\left\|\hat{\zeta}_{2\ell} - \zeta_{20}\right\| = O_p(n^{-\xi_3/(2\xi_1+1)} \ln(n)).$$

Also note that $H_p(\hat{\gamma}_{1\ell}(x))$ and $H_p(\gamma_{10}(x))$ are bounded by the fixed trimming, which together with $\|\hat{\gamma}_{1\ell} - \gamma_{10}\| = O_p(n^{-d_1})$ also gives $\|H_p(\hat{\gamma}_{1\ell}) - H_p(\gamma_{10})\| = O_p(n^{-d_1})$.Then by the triangle inequality and boundedness of $\zeta_{10}(x)$ and $\zeta_{20}(x)$,

$$\begin{aligned}
\|\hat{\alpha}_{1\ell} - \alpha_{10}\| &= \left\|(\hat{\zeta}_{1\ell} + \hat{\alpha}_{3\ell}\hat{\zeta}_{2\ell}) H_p(\hat{\gamma}_{1\ell}) - (\zeta_{10} + \alpha_{30}\zeta_{20}) H_p(\gamma_{10})\right\| \\
&\leq \left\| \left[(\hat{\zeta}_{1\ell} - \zeta_{10}) + \hat{\alpha}_3(\hat{\zeta}_{2\ell} - \zeta_{20})\right] H_p(\hat{\gamma}_{1\ell})\right\| \\
&+ \|(\zeta_{10} + \hat{\alpha}_{3\ell}\zeta_{20}) [H_p(\hat{\gamma}_{1\ell}) - H_p(\gamma_{10})]\| + \|(\hat{\alpha}_{3\ell} - \alpha_{30})\zeta_{20} H_p(\gamma_{10})\| \\
&= O_p(n^{-d_1[(2\xi_1-1)/(2\xi_1+1)]2\xi_2/[2\xi_2+1]} [\ln(n)]^2 + n^{-\xi_3/(2\xi_1+1)} \ln(n)) + O_p(n^{-d_1}).
\end{aligned}$$

The first conclusion follows by the triangle inequality. The last conclusion follows similarly to $\|H_p(\hat{\gamma}_{1\ell}) - H_p(\gamma_{10})\| = O_p(n^{-d_1})$. *Q.E.D.*

Proof of Theorem 12: We proceed by verifying Assumptions 1-4 in Section 6 and the conditions of Theorem 9 for $\gamma = (\gamma_1, \gamma_2, \gamma_3)$. Assumption 1 follows by Lemmas E6 and E7 and by $a(x)$, Y_{2t}, $H(\gamma_{10}(X_t))$, $\gamma_{20}(X_t)$, $\alpha_{10}(X_t)$, $\alpha_{20}(X_t, Y_{2t})$, $\alpha_{30}(X_t)$ all bounded, similarly to the proof of Lemma E8.

To show Assumption 2 ii), note that

$$\begin{aligned}
\hat{\Delta}_\ell(w) &= \hat{\Delta}_{\ell 1}(w) + \hat{\Delta}_{\ell 2}(w) + \hat{\Delta}_{\ell 3}(w), \\
\hat{\Delta}_{\ell 1}(w) &= -\frac{1}{T}\sum_{t=1}^{T}[\hat{\alpha}_{1\ell}(x_t) - \alpha_{10}(x_t)][\hat{\gamma}_{1\ell}(x_t) - \gamma_{10}(x_t)], \\
\hat{\Delta}_{\ell 2}(w) &= \frac{1}{T}\sum_{t=1}^{T}[\hat{\alpha}_{2\ell}(x_t, y_{2t}) - \alpha_{20}(x_t, y_{2t})][H(\hat{\gamma}_{1\ell}(x_t)) - \hat{\gamma}_{2\ell}(x_t) - H(\gamma_{10}(x_t)) + \gamma_{20}(x_t)], \\
\hat{\Delta}_{\ell 3}(w) &= \frac{1}{T}\sum_{t=1}^{T}(\hat{\alpha}_{3\ell} - \alpha_{30})y_{2t}\{H(\hat{\gamma}_1(x_{t+1})) - \hat{\gamma}_3 - H(\hat{\gamma}_1(x_{t+1})) + \gamma_{30}\}.
\end{aligned}$$

Then by the first conclusion of Lemma E8 and conditions iv), v), and vi),

$$\begin{aligned}
\sqrt{n}\int \left\|\hat{\Delta}_{\ell 1}(w)\right\| F_0(dw) &\leq \sqrt{n}\,\|\hat{\alpha}_{1\ell} - \alpha_{10}\|\,\|\hat{\gamma}_{1\ell} - \gamma_{10}\| \\
&= O_p(\sqrt{n}\{n^{-d_1[(2\xi_1-1)/(2\xi_1+1)]2\xi_2/[2\xi_2+1]}[\ln(n)]^2 + n^{-\xi_3/(2\xi_3+1)}\ln(n) + n^{-d_1}\}n^{-d_1}) \\
&= o_p(1).
\end{aligned}$$

Next, by the second conclusion of Lemma E8 and condition vii)

$$\begin{aligned}
\sqrt{n}\int \left\|\hat{\Delta}_{\ell 2}(w)\right\| F_0(dw) &\leq \sqrt{n}\,\|\hat{\alpha}_{2\ell} - \alpha_{20}\|\,\|H(\hat{\gamma}_{1\ell}) - \hat{\gamma}_{2\ell} - H(\gamma_{10}) + \gamma_{20}\| \\
&= O_p(\sqrt{n}[n^{-d_1(2\xi_1-1)/(2\xi_1+1)}\ln(n)]n^{-d_1 2\xi_1/(2\xi_1+1)}\ln(n)) = o_p(1).
\end{aligned}$$

Next, by the third conclusion of Lemma E8, condition v), and $2\xi_1/(2\xi_1+1) < 1$ we have

$$\begin{aligned}
\sqrt{n}\int \left\|\hat{\Delta}_{\ell 3}(w)\right\| F_0(dw) &\leq \sqrt{n}\,\|\hat{\alpha}_{3\ell} - \alpha_{30}\|\,\|H(\hat{\gamma}_{1\ell}) - \hat{\gamma}_{3\ell} - H(\gamma_{10}) + \gamma_{30}\| \\
&= O_p(\sqrt{n}[n^{-d_1(2\xi_1-1)/(2\xi_1+1)}\ln(n)]n^{-d_1}) = o_p(1).
\end{aligned}$$

Assumption 2 ii) then follows by the triangle and conditional Markov inequality.

Next, Assumption 3 i) follows by the form of ϕ_1, ϕ_2, and ϕ_3 given in Section 3 and

$$\begin{aligned}
E[Y_{2t} - \gamma_{10}(X_t)|X_t] = 0,\ E[Y_{2t}\{H(\gamma_{10}(X_{t+1})) - \gamma_{20}(X_t)\}|X_t] = 0, \\
E[Y_{1t}\{H(\gamma_{10}(X_{t+1})) - \gamma_{30}\}] = 0.
\end{aligned}$$

We now proceed to verify that Assumption 3 iv) is satisfied. For ease of exposition we suppress the ℓ subscript. Let $\hat{a}(x) = a(x, \theta_0, \hat{\gamma}_2, \hat{\gamma}_3)$ and $\hat{\pi}(x) = \pi(a(x, \theta_0, \hat{\gamma}_2, \hat{\gamma}_3))$. Then

$$\begin{aligned}
\bar{\psi}(\hat{\gamma}, \alpha_0, \theta_0) &= T + T_1 + T_2 + T_3,\ T = \int D(x_t)\hat{\pi}(x_t)\{y_{2t} - \Lambda(\hat{a}(x_t))\}F_0(dw), \\
T_1 &= \int \alpha_{10}(x_t)[y_{2t} - \hat{\gamma}_1(x_t)]F_0(dw),\ T_3 = \alpha_{03}\int y_{1t}[H(\hat{\gamma}_{1\ell}(x_{t+1})) - \hat{\gamma}_3]F_0(dw), \\
T_2 &= \int \alpha_{20}(x_t, y_{2t})[H(\hat{\gamma}_1(x_{t+1})) - \hat{\gamma}_2(x_t)]F_0(dw).
\end{aligned}$$

Note that

$$\begin{aligned}
T &= \bar{T}_1 + \bar{T}_2 + R_1 + R_2, \\
\bar{T}_1 &= -\delta \int D(x_t)\pi(x_t)\Lambda_a(a(x_t))\left[\hat{\gamma}_2(x_t) - \gamma_{20}(x_t)\right]F_0(dw), \\
\bar{T}_2 &= A\left(\hat{\gamma}_3 - \gamma_{30}\right),\ A = \delta \int D(x_t)\pi(x_t)\Lambda_a(a(x_t))F_0(dw), \\
R_1 &= -\int D(x_t)\hat{\pi}(x_t)\Lambda_{aa}(\bar{a}(x_t))\left|\hat{a}(x_t) - a(x_t)\right|^2 F_0(dw), \\
R_2 &= -\int D(x_t)[\hat{\pi}(x_t) - \pi(x_t)]\Lambda_a(a(x_t))\left[\hat{a}(x_t) - a(x_t)\right]F_0(dw).
\end{aligned}$$

Also,

$$\begin{aligned}
\|R_1\| &\leq C\left(\|\hat{\gamma}_{2\ell} - \gamma_{20}\|^2 + |\hat{\gamma}_{3\ell} - \gamma_{30}|^2\right) = O_p(n^{-4d_1\xi_1/(2\xi_1+1)}) = o_p(n^{-1/2}), \\
\|R_2\| &\leq C\left(\|\hat{\gamma}_{2\ell} - \gamma_{20}\|^2 + |\hat{\gamma}_{3\ell} - \gamma_{30}|^2\right) = o_p(n^{-1/2}),
\end{aligned}$$

so that

$$T = \bar{T}_1 + \bar{T}_2 + o_p(n^{-1/2}).$$

Next, note that by the definition of $\alpha_{20}(x, y_2)$,

$$\bar{T}_1 = \int \alpha_{20}(x_t, y_{2t})\{\hat{\gamma}_2(x_t) - \gamma_{20}(x_t)\}F_0(dw).$$

Therefore

$$\bar{T}_1 + T_2 = \tilde{T}_2,\ \tilde{T}_2 = \int \alpha_{10}(x_t, y_{2t})[H(\hat{\gamma}_1(x_{t+1})) - \gamma_{20}(x_t)]F_0(dw).$$

Note that by $\gamma_{20}(x) = E[H(\gamma_{10}(X_{t+1}))|X_t = x, Y_{2t} = 1]$ and subtracting and adding the expression $\int \alpha_{20}(x_t, y_{2t})H(\gamma_{10}(x_{t+1}))F_0(dw)$ we obtain

$$\begin{aligned}
\tilde{T}_2 &= \int \alpha_{20}(x_t, y_{2t})[H(\hat{\gamma}_1(x_{t+1})) - H(\gamma_{10}(x_{t+1}))]F_0(dw) + \int \alpha_{20}(x_t, y_{2t})[H(\gamma_{10}(x_{t+1})) - \gamma_{20}(x_t)]F_0(dw) \\
&= \int \alpha_{20}(x_t, y_{2t})[H(\hat{\gamma}_1(x_{t+1})) - H(\gamma_{10}(x_{t+1}))]F_0(dw).
\end{aligned}$$

Expanding then gives

$$\tilde{T}_2 = \breve{T}_2 + R_3,\ \breve{T}_2 = \int \alpha_{20}(x_t, y_{2t}) H_p(\gamma_{10}(x_{t+1}))[\hat{\gamma}_1(x_{t+1}) - \gamma_{10}(x_{t+1})] F_0(dw),$$
$$R_3 = \int \alpha_{20}(x_t, y_{2t}) H_{pp}(\bar{\gamma}_1(x_{t+1}))[\hat{\gamma}_1(x_{t+1}) - \gamma_{10}(x_{t+1})]^2 F_0(dw),$$

where $\bar{\gamma}_1(x)$ is between $\hat{\gamma}_1(x_t)$ and $\gamma_{10}(x_t)$. It follows similarly to previous arguments that $\|R_3\| \leq C\|\hat{\gamma}_1 - \gamma_{10}\|^2 = O_p(n^{-2d_1}) = o_p(n^{-1/2})$, so that $\tilde{T}_2 = \breve{T}_2 + o_p(n^{-1/2})$. Also,

$$\breve{T}_2 = \int \zeta_{10}(x_t) H_p(\gamma_{10}(x_t))[\hat{\gamma}_1(x_t) - \gamma_{10}(x_t)] F_0(dw),\ \zeta_{10}(x) = E[\alpha_{20}(X_t, Y_{2t}) | X_{t+1} = x].$$

Note that

$$\alpha_{10}(x) = [\zeta_{10}(x) + \alpha_{30}\zeta_{20}(x)] H_p(\gamma_{10}(x)),\ \zeta_{20}(x) = E[Y_{1t} | X_{t+1} = x].$$

Then by $\int \zeta_{10}(x_t) H_p(\gamma_{10}(x_t))[y_{2t} - \gamma_{10}(x_t)] F_0(dw) = 0$ we have

$$\begin{aligned}\breve{T}_2 + T_1 &= \alpha_{30} \int \zeta_{20}(x_t) H_p(\gamma_{10}(x_t))[y_{2t} - \hat{\gamma}_1(x_t)] F_0(dw) \\ &= \alpha_{30} \int \zeta_{20}(x_t) H_p(\gamma_{10}(x_t))[\gamma_{10}(x_t) - \hat{\gamma}_1(x_t)] F_0(dw).\end{aligned}$$

Next, note that by iterated expectations and $\alpha_{30} = A/P_1$,

$$T_3 = \alpha_{03} \int y_{1t}[H(\hat{\gamma}_{1\ell}(x_{t+1})) - \hat{\gamma}_3] F_0(dw) = \alpha_{30} \int \zeta_{20}(x_t) H(\hat{\gamma}_1(x_t)) F_0(dw) - A\hat{\gamma}_3.$$

Note also that

$$A\gamma_3 = AE[y_{1t} H(\gamma_{10}(x_{t+1}))]/P_1 = \alpha_{30} \int \zeta_{20}(x_t) H(\gamma_{10}(x_t)) F_0(dw)$$

Then by an expansion

$$\begin{aligned}\bar{T}_2 + T_3 &= \alpha_{03} \int \zeta_{20}(x_t) H(\hat{\gamma}_1(x_t)) F_0(dw) - A\gamma_3 = \alpha_{03} \int \zeta_{20}(x_t)[H(\hat{\gamma}_1(x_t)) - H(\gamma_{10}(x_t))] F_0(dw) \\ &= \alpha_{03} \int \zeta_{20}(x_t)[H(\hat{\gamma}_1(x_t)) - H(\gamma_{10}(x_t))] F_0(dw) \\ &= \alpha_{30} \int \zeta_{20}(x_t) H_p(\gamma_{10}(x_t))[\hat{\gamma}_1(x_t) - \gamma_{10}(x_t)] F_0(dw) + R_4 = -(\breve{T}_2 + T_1) + R_4, \\ R_4 &= \alpha_{30} \int \zeta_{20}(x_t) H_{pp}(\bar{\gamma}_1(x_t))[\hat{\gamma}_1(x_t) - \gamma_{10}(x_t)]^2 F_0(dw),\end{aligned}$$

where $\bar{\gamma}_1(x)$ is between $\hat{\gamma}_1(x_t)$ and $\gamma_{10}(x_t)$. It follows similarly to previous arguments that $\|R_3\| \leq C\|\hat{\gamma}_1 - \gamma_{10}\|^2 = O_p(n^{-2d_1}) = o_p(n^{-1/2})$. Therefore

$$\bar{T}_2 + T_3 = -(\breve{T}_2 + T_1) + o_p(n^{-1/2}).$$

Summarizing, it follows from what has been shown that

$$\begin{aligned}\bar{\psi}(\hat{\gamma},\alpha_0,\theta_0) &= T+T_1+T_2+T_3 = \bar{T}_1+\bar{T}_2+T_1+T_2+T_3+o_p(n^{-1/2}) \\ &= \tilde{T}_2+\bar{T}_2+T_1+T_3+o_p(n^{-1/2}) = \breve{T}_2+\bar{T}_2+T_1+T_3+o_p(n^{-1/2}) \\ &= (\breve{T}_2+T_1)+(\bar{T}_2+T_3)+o_p(n^{-1/2}) = (\breve{T}_2+T_1)-(\breve{T}_2+T_1)+o_p(n^{-1/2}) \\ &= o_p(n^{-1/2}),\end{aligned}$$

giving Assumption 3 iv).

Next, note that by the fixed trimming $H\left(\hat{\gamma}_1(x)\right)$ and $\hat{\gamma}_1(x)$ are uniformly bounded. Also, by Lemma E6, $\hat{\gamma}_2(x)$ and $\hat{\gamma}_3$ are uniformly bounded with probability approaching one, so

$$\left\|\hat{\Delta}_\ell(w)\right\| \leq C(\|\hat{\alpha}_{1\ell}(x)-\alpha_{10}(x)\|+\|\hat{\alpha}_{2\ell}(x)-\alpha_{20}(x)\|+\|\hat{\alpha}_{3\ell}-\alpha_{30}\|).$$

The second condition of Assumption 4 then follows by Lemma E8. The first condition of Assumption 4 also follows in a straightforward manner from uniform boundedness of $\hat{\gamma}_2(x)$ and $\hat{\gamma}_3$ with probability approaching one.

Finally, Assumption 5 follows in a straightforward manner from the same boundedness properties, so the conclusion follows by Theorem 9. *Q.E.D.*

5 Appendix F: Consistency of Debiased GMM.

THEOREM A3: *If i)* $\hat{\Upsilon} \xrightarrow{p} \Upsilon$ *positive definite; ii)* $E[g(W,\gamma_0,\theta)]=0$ *if and only if* $\theta=\theta_0$; *iii)* Θ *is compact; iv)* $\int \|g(w,\hat{\gamma}_\ell,\theta)-g(w,\gamma_0,\theta)\|\,F_0(dw) \xrightarrow{p} 0$ *and* $E[\|g(W,\gamma_0,\theta)\|]<\infty$ *for all* $\theta\in\Theta$; *v) there is* $C>0$ *and* $d(W,\gamma)$ *such that for* $\|\gamma-\gamma_0\|$ *small enough and all* $\tilde{\theta}$, $\theta\in\Theta$ *and*

$$\left\|g(W,\gamma,\tilde{\theta})-g(W,\gamma,\theta)\right\| \leq d(W,\gamma)\left\|\tilde{\theta}-\theta\right\|^{1/C};\ E[d(W,\gamma)]<C.$$

vi) Assumption 1 ii) and iii) are satisfied, $\int \left\|\hat{\Delta}_\ell(w)\right\| F_0(dw) \xrightarrow{p} 0$, *and* $E[\|\phi(W,\gamma_0,\alpha_0,\theta_0)\|]<\infty$.

Proof: It follows from iv) that $\hat{g}(\theta) \xrightarrow{p} \bar{g}(\theta) := E[g(W,\gamma_0,\theta)]$ for all $\theta\in\Theta$. Let $\bar{\phi} := \sum_{i=1}^n \phi(W_i,\gamma_0,\alpha_0,\theta)/n$. In the notation in Lemma 8 it follows that $\phi(W_i,\hat{\gamma}_\ell,\hat{\alpha}_\ell,\tilde{\theta}_\ell)-\phi(W_i,\gamma_0,\alpha_0,\theta_0) = \hat{R}_{2\ell i}+\hat{R}_{3\ell i}+\hat{\Delta}_\ell(W_i)$. Then by $E[\phi(W,\gamma_0,\alpha_0,\theta_0)]=0$, condition vi), the conditional Markov inequality, Khintchine's law of large numbers, and the triangle inequality $\hat{\phi}=\hat{\phi}-\bar{\phi}+\bar{\phi} \xrightarrow{p} 0$. Therefore $\hat{\psi}(\theta)=\hat{g}(\theta)+\hat{\phi} \xrightarrow{p} \bar{g}(\theta)$ for all $\theta\in\Theta$. Next, by v) it follows that with probability approaching one, $\left\|\hat{g}(\tilde{\theta})-\hat{g}(\theta)\right\| \leq \hat{M}\left\|\tilde{\theta}-\theta\right\|^{1/C}$ for $\hat{M}=\sum_{\ell=1}^L\sum_{i\in I_\ell} d(W_i,\hat{\gamma}_\ell)/n$. Also $\hat{M}=O_p(1)$ by the conditional Markov inequality. Then by Corollary 2.2 of Newey (1991) we have $\sup_{\theta\in\Theta}\left\|\hat{\psi}(\theta)-\bar{g}(\theta)\right\| \xrightarrow{p} 0$. In addition condition v) implies that $\bar{g}(\theta)$ is continuos on Θ. The conclusion then follows similarly to the proof of Theorem 2.6 of Newey and McFadden (1994). *Q.E.D.*

REFERENCES

BICKEL, P.J., C.A.J. KLAASSEN, Y. RITOV, AND J.A. WELLNER (1993): *Efficient and Adaptive Estimation for Semiparametric Models*, Springer-Verlag, New York.

BICKEL, P.J., Y. RITOV, A.B. TSYBAKOV (2009): "Simultaneous Analysis of Lasso and Dantzig Selector," *Annals of Statistics* 37: 1705-1732.

BRADIC, J., V. CHERNOZHUKOV, W.K. NEWEY, Y. ZHU (2021): "Minimax Semiparametric Learning with Approximate Sparsity," https://arxiv.org/pdf/1912.12213.pdf.

CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, W. NEWEY, J. ROBINS (2018): "Debiased/Double Machine Learning for Treatment and Structural Parameters,*Econometrics Journal* 21, C1-C68.

CHERNOZHUKOV, V., W.K. NEWEY, AND R. SINGH (2020): "Learning L2-Continuous Regression Functionals via Regularized Riesz Representers," https://arxiv.org/pdf/1809.05224v3.pdf.

LEEB, H. AND B.M. POTSCHER (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory* 21, 21-59.

NEWEY, W.K. (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica* 59, 1161-1167.

NEWEY, W.K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* 62, 1349-1382.

NEWEY, W.K., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. 4, ed. by R. Engle, and D. McFadden, pp. 2113-2241. North Holland.

NEWEY, W.K., AND J. ROBINS (2017): "Cross Fitting and Fast Remainder Rates for Semiparametric Estimation," CEMMAP Working paper WP41/17.

STOKER, T. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica* 54, 1461-1482.

VAN DER VAART, A.W. (1991): "On Differentiable Functionals," *The Annals of Statistics,* 19, 178-204.

VAN DER VAART, A.W. (1998): *Asymptotic Statistics,* Cambridge University Press, Cambridge, England.

ZHAO, P., AND B. YU (2006): "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research* 7, 2541-2563.