Supplementary Appendix for the paper

Machine learning for set-identified linear models

by Vira Semenova

Abstract

The Supplementary Appendix contains proofs of some results stated in the paper "Machine Learning for Set-Identified Linear Models" by Vira Semenova. Section 8 contains the demonstration of the bias calculation for the partially linear predictor. Section 9 contains a general recipe to obtain an orthogonal moment function starting from a non-orthogonal one, extending previous work of Newey (1994), Newey and Stoker (1993), etc. from point to set-identified case. Section 10 contains the proofs of Theorem 3,4,5 from the paper and the proofs of supplementary lemmas.

8 Partially linear predictor in one-dimensional case

Consider the setting from Example 3 when the endogenous variable D is one-dimensional. Then the identified set $\mathcal{B}$ is a closed interval

$$\mathcal{B} = [\beta_L, \beta_U].$$

Given an i.i.d sample $(W_i)_{i=1}^N = (D_i, X_i, Y_{L,i}, Y_{U,i})_{i=1}^N$, I derive a root-$N$ consistent asymptotically normal estimator, $[\hat{\beta}_L, \hat{\beta}_U]$, of the identified set and construct a confidence region for the identified set $\mathcal{B}$.

I characterize the upper bound β_U as a solution to a semiparametric moment equation. Inspecting (2.8), one can see that the identified set (2.8) consists of the ordinary least squares coefficients where the first-stage residual $(D - \eta_0(X))$ is the regressor and $Y \in [Y_L, Y_U]$ is an outcome. To achieve the upper bound β_U, or, equivalently, the largest possible least squares coefficient, I construct a random variable Y^{UBG} as

$$Y^{\text{UBG}}(\eta) = \begin{cases} Y_L, & D - \eta(X) \leq 0, \\ Y_U, & D - \eta(X) > 0. \end{cases} \tag{8.1}$$

Intuitively, $Y^{\text{UBG}}(\eta)$, referred to as an upper bound generator, takes the largest possible value Y_U when $D - \eta(X)$ is positive and the smallest possible value Y_L otherwise[5]. As a result, the upper bound is characterized by the semiparametric moment equation

$$\mathbb{E}(Y^{\text{UBG}}(\eta_0) - (D - \eta_0(X))\beta_U)(D - \eta_0(X)) = 0 \tag{8.2}$$

(see, e.g. Beresteanu and Molinari (2008) or Bontemps et al. (2012)). The major difficulty when estimating β_U comes from the nuisance function $\eta_0(X) = \mathbb{E}[D|X]$, which is a function of high-dimensional covariates vector and must be estimated by regularized machine learning methods in order to achieve consistency.

I describe the naive approach to estimate β_U and explain why it does not work. To abstract away from other estimation issues, I use different samples for the first and second stages. Given the sample $(W_i)_{i=1}^N$, I split it into a main sample J_1 and an auxiliary sample J_2 of equal size $n = [N/2]$ such that $J_1 \cup J_2 = \{1, 2, \ldots, N\}$. I use the auxiliary sample J_2 to construct an estimator $\hat{\eta}(X)$. Then, I construct an estimate of the upper bound generator $\hat{Y}_i^{\text{UBG}}$ and regress it on the estimated first-stage residual $D_i - \hat{\eta}(X_i)$

$$\hat{\beta}_U^{\text{NAIVE}} = (\sum_{i \in J_1} (D_i - \hat{\eta}(X_i))^2)^{-1} \sum_{i \in J_1} (D_i - \hat{\eta}(X_i))\hat{Y}_i^{\text{UBG}}.$$

Unfortunately, the naive estimator converges at a rate slower than $\sqrt{N}$

$$\sqrt{N}|\hat{\beta}_U^{\text{NAIVE}} - \beta_U| \to \infty \tag{8.3}$$

and cannot be used to conduct inference about β_U using standard Gaussian approximation. The behavior of the naive estimator is shown in Figure 1(a).

The slow convergence of the naive estimator $\hat{\beta}_U^{\text{NAIVE}}$ is due to the slower-than-root-N convergence of the first-stage estimator of $\eta_0(X)$. In order to estimate $\eta_0(X)$ consistently in a high-dimensional framework, I must employ modern regularized methods, such as boosting, random forest, and lasso, that rely on regularization constraints to achieve convergence. This regulariza-

[5] In what follows, I assume that the residual V has a continuous distribution and is equal to zero with probability zero.

Figure 1: Finite-sample distribution of non-orthogonal (naive) and orthogonal estimates of the bounds

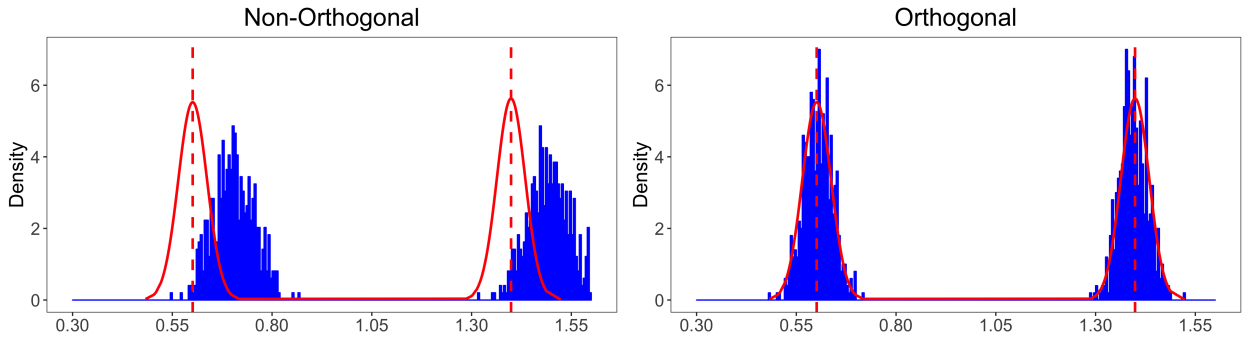

Figure 1 shows the finite-sample distribution (blue histogram) of naive (left panel) and orthogonal (right panel) estimates of the lower (β_L) and the upper (β_U) bounds of the identified set. The red curve shows the normal (infeasible) approximation when the first-stage parameter $\eta_0(X) = \mathbb{E}[D|X]$ is known. The dashed line should the true value of the bound. In the left panel, the distribution of the naive estimator is centered substantially far from the true value. The naive estimator is biased because the first-stage bias transmits into the bias of the bounds. In the right panel, the distributions are close. This estimator is approximately unbiased because the first-stage bias of $\hat{\eta}$ does not transmit into the bias of the bounds. The function $\mathbb{E}[D|X]$ is a linear sparse function of a high-dimensional vector X, so the gamma-lasso first-stage estimator of $\mathbb{E}[D|X]$ from Taddy (2011) has good prediction properties. I use the cross-fitting procedure with the number of folds $K = 2$.

tion creates bias in the first-stage estimates. The bias converges slower than root-N and carries over into the naive estimator $\hat{\beta}_U^{\text{NAIVE}}$.

I show that the major obstacle to optimal convergence and valid inference is the sensitivity of the moment function (8.2) with respect to the biased estimation of the first stage parameter η_0. Assume that I can somehow generate the true value of the upper bound generator $\mathcal{Y}_0 = Y^{\text{UBG}}(\eta_0)$. Consider a smooth moment function

$$m_0(W, \beta_U, \eta_0) = (\mathcal{Y}_0 - (D - \eta_0(X))\beta_U) \cdot (D - \eta_0(X)) \tag{8.4}$$

Then the difference between the infeasible moment equation $m_0(W, \beta_U, \eta_0)$, based on the true value of the nuisance parameter η_0, and the feasible yet slightly incorrect moment equation $m_0(W, \beta_U, \hat{\eta})$, based on the first-stage estimate $\hat{\eta}$ is proportional to the expected derivative of (8.4)

$$\mathbb{E}[m_0(W, \beta_U, \hat{\eta}) - m_0(W, \beta_U, \eta_0)] \approx \partial_{\eta_0} \mathbb{E}[m_0(W, \beta_U, \eta_0)(\hat{\eta}(X) - \eta_0(X))].$$

The derivative of (8.4) is non-zero

$$\partial_{\eta_0}\mathbb{E}m_0(W,\beta_U,\eta_0)(\hat{\eta}-\eta_0) = -\mathbb{E}[\mathcal{Y}_0(\hat{\eta}(X)-\eta_0(X))],$$

which is why the first-stage bias carries over into the second stage.

To overcome the transmission of the bias, I replace the moment equation (8.2) by another moment equation that is less sensitive to the biased estimation of its first-stage parameters. Using the classic idea from Frisch-Waugh-Lowell, I replace $\mathcal{Y}_0$ by the second-stage residual $\mathcal{Y}_0 - \mathbb{E}[\mathcal{Y}_0|X]$. The derivative of the new moment equation takes the form

$$-\mathbb{E}[(\mathcal{Y}_0 - \mathbb{E}[\mathcal{Y}_0|X])(\hat{\eta}(X)-\eta_0(X))] = 0.$$

The new moment equation takes the form

$$\mathbb{E}(\mathcal{Y}_0 - \mathbb{E}[\mathcal{Y}_0|X]) - (D-\eta_0(X))\beta_U)\cdot(D-\eta_0(X)) = 0$$

and can be interpreted as the ordinary least squares regression of the second-stage residual $\mathcal{Y}_0 - \mathbb{E}[\mathcal{Y}_0|X]$ on the first-stage residual $D-\eta_0(X)$. This equation is known as a doubly-robust moment equation (Robins and Rotnitzky (1995), Robins et al. (1994), Chernozhukov et al. (2017a)) from point-identified case, where an observed outcome Y appeared in place of the constructed (and unobserved) upper bound generator $\mathcal{Y}_0$.

I argue that the estimation error of the upper bound generator $Y^{\text{UBG}}(\eta_0)$ can be ignored when the first-stage residual $V = D-\eta_0(X)$ is continuously distributed. Then this estimation error matters (i.e., $Y^{\text{UBG}}(\hat{\eta}) \neq Y^{\text{UBG}}(\eta_0)$) only if the first-stage residual is small enough

$$|Y^{\text{UBG}}(\hat{\eta}) - Y^{\text{UBG}}(\eta_0)| \leq \begin{cases} Y_U - Y_L, & 0 < |D-\eta_0(X)| < |\hat{\eta}(X)-\eta_0(X)| \\ 0, & \text{otherwise} \end{cases}.$$

When the residual $D-\eta_0(X)$ is sufficiently continuous, the probability of the event $Y^{\text{UBG}}(\hat{\eta}) \neq Y^{\text{UBG}}(\eta_0)$ is smaller than the estimation error $|\hat{\eta}(X)-\eta_0(X)|$. Assuming that the estimation error $|\hat{\eta}(X)-\eta_0(X)|$ itself converges at $o(N^{-1/4})$ rate, I show that this error can be ignored since its

contribution to bias is second-order.

The proposed estimator has two stages. In the first-stage, I estimate the conditional expectations

$$\{\eta_0(X), \mathbb{E}[\mathcal{Y}_0|X]\}$$

of the endogenous variable D and of the upper bound generator Y^{UBG}, respectively, using machine learning tools. In the second stage, I regress the estimated second-stage residual on the estimated first-stage residual. I use different samples in the first and the second stages (a more sophisticated form of sample splitting, called cross-fitting, is defined in Section 3). The behavior of the proposed estimator is shown in Figure 1(b).

Algorithm 3 Upper Bound on the Partially Linear Predictor

Let $\gamma_{U,0}(X) := \mathbb{E}[\mathcal{Y}_0|X]$.
Input: an i.i.d sample $(W_i)_{i=1}^N = (D_i, X_i, Y_{L,i}, Y_{U,i})_{i=1}^N$, estimated values $(\hat{\eta}(X_i), \hat{\gamma}_U(X_i))_{i\in J_2}$, where $\hat{\gamma}_U(\cdot)$ is estimated using the auxiliary sample J_2.

1: Estimate the upper bound generator for every $i \in J_1$

$$\hat{Y}_i^{\text{UBG}} := \begin{cases} Y_{L,i}, & D_i - \hat{\eta}(X_i) \leq 0, \\ Y_{U,i}, & D_i - \hat{\eta}(X_i) > 0. \end{cases}$$

2: Estimate $\hat{\beta}_U$ by Ordinary Least Squares using the second-stage residual of the upper bound generator as the dependent variable and the first-stage residual V as the regressor

$$\hat{\beta}_U = (\sum_{i\in J_1} (D_i - \hat{\eta}(X_i))^2)^{-1} \sum_{i\in J_1} (D_i - \hat{\eta}(X_i))[\hat{Y}_i^{\text{UBG}} - \hat{\gamma}_U(X_i)]. \tag{8.5}$$

Return: $\hat{\beta}_U$.

Sample Splitting. I can use machine learning methods in the first stage because of sample splitting. In the absence of sample splitting, the estimation error of the first-stage machine learning estimator may be correlated with the true values of the first and second-stage residuals. This correlation leads to bias, referred to as overfitting bias. The behavior of the overfit estimator is shown in Figure 2 (a).

While sample splitting helps overcome overfitting bias, it cuts the sample used for the estimation in half. This problem can lead to the loss of efficiency in small samples. To overcome this problem, I use the cross-fitting technique from Chernozhukov et al. (2017a) defined in Section 3.

Figure 2: Finite-sample distribution of the orthogonal estimator without and with sample splitting

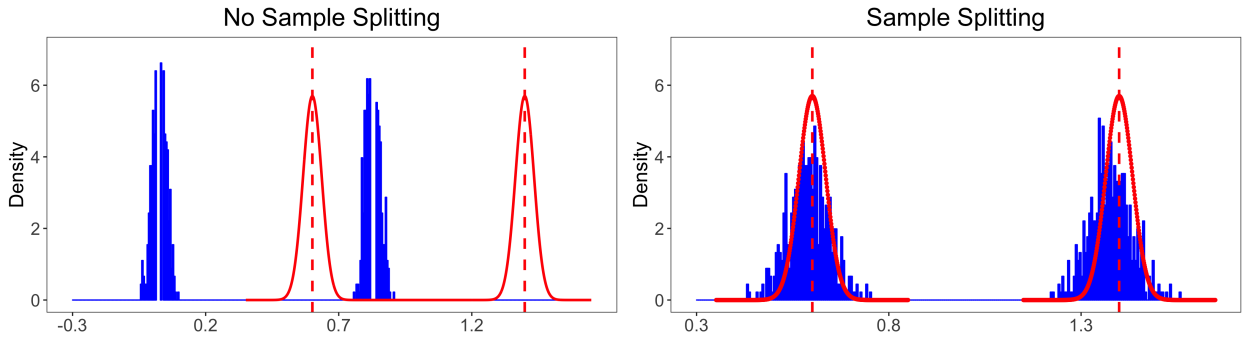

Figure 2 shows the finite-sample distribution (blue histogram) of the orthogonal estimator without (left panel) and with (right panel) sample splitting. The red curve shows the normal (infeasible) approximation when the first-stage parameter $\eta_0(X) = \mathbb{E}[D|X]$ is known. The dashed line should the true value of the bound. In the left panel, the distribution of the naive estimator is centered substantially far from the true value. The naive estimator is biased because of overfitting. In the right panel, the distributions are close. This estimator is approximately unbiased because different samples are used in the first and the second stages. The function $\mathbb{E}[D|X]$ is a linear sparse function of a high-dimensional vector X, so the gamma-lasso first-stage estimator of $\mathbb{E}[D|X]$ from Taddy (2011) has good prediction properties. I use the cross-fitting procedure with the number of folds $K = 2$.

Specifically, I partition the sample into two halves. To estimate the residuals for each half, I use the other half to estimate the first-stage nuisance parameter. Then, the upper bound is estimated using the whole sample. As a result, each observation is used both in the first and second stages, improving efficiency in small samples.

Sketch of the pointwise result. I end this section with a sketch of my pointwise result. Let $[\hat{\beta}_L, \hat{\beta}_U]^\top$ be a vector of the estimators of the lower and upper bounds defined in Algorithm 3. My estimator is root-N consistent and asymptotically Gaussian

$$\sqrt{N} \begin{pmatrix} \hat{\beta}_L - \beta_L \\ \hat{\beta}_U - \beta_U \end{pmatrix} \Rightarrow N(0, \Omega), \tag{8.6}$$

where the sample size N converges to infinity, $\Rightarrow$ denotes convergence in distribution, and Ω is a covariance matrix. The confidence region of level $\alpha \in (0,1)$ for the identified set $[\beta_L, \beta_U]$ takes the

form

$$[\hat{\beta}_L - N^{-1/2}\hat{C}_{\alpha/2}, \quad \hat{\beta}_U + N^{-1/2}\hat{C}_{1-\alpha/2}],$$

where the critical values $\hat{C}_{\alpha/2}, \hat{C}_{1-\alpha/2}$ are

$$\begin{pmatrix} \hat{C}_{\alpha/2} \\ \hat{C}_{1-\alpha/2} \end{pmatrix} = \hat{\Omega}^{1/2} \begin{pmatrix} \Phi^{-1}(\sqrt{1-\alpha}) \\ \Phi^{-1}(\sqrt{1-\alpha}) \end{pmatrix}$$

and $\Phi^{-1}(t)$ is the inverse of the standard normal distribution. I estimate the covariance matrix Ω using a version of weighted bootstrap given in Definition 6.

9 General Recipe for the Construction of an Orthogonal Moment Condition

In this section, I provide a general recipe to construct a near-orthogonal moment condition for the support function starting from a non-orthogonal moment condition (3.4), extending the previous work of (Hardle and Stoker (1989), Newey (1994), Chernozhukov et al. (2017b), Ichimura and Newey (2017)) from a point- to a set-identified case. Adding generality helps to understand the derivation. Suppose I am interested in a function $M(p)$ defined by the moment condition

$$M(p) = \mathbb{E}m(W, p, \eta_0),$$

where $\eta_0(X)$ is a functional parameter. To make the moment condition above insensitive to the biased estimation of η_0, I add a bias correction term $\alpha(W, p, \xi(p))$ that enjoys the following two properties. First, the bias correction term has zero mean

$$\mathbb{E}[\alpha(W, p, \xi_0(p))] = 0,$$

so that the new moment condition is still valid. Second, I require that the function

$$g(W,p,\xi(p)) = m(W,p,\eta) + \alpha(W,p,\xi(p)) \tag{9.1}$$

obeys the Neyman-orthogonality condition (Assumption 3).

Lemma[6] 10 derives a general form of a bias correction term for the case $\eta_0(X)$ is defined via the conditional exogeneity restriction (9.2). In our applications, we consider two important cases of this Lemma: a conditional expectation function (Lemma 11) and a conditional quantile function (Lemma 12). Lemma 10 is the extension of Ichimura and Newey (2017)'s result to the set-identified case.

Lemma 10 (Bias Correction Term for a Nuisance Function Determined by a Conditional Exogeneity Restriction)**.** *Suppose the true value $\eta_0 = \eta_0(X)$ of a functional nuisance parameter η satisfies the generalized conditional exogeneity restriction*

$$\mathbb{E}R[(W,\eta_0(X))|X] = 0, \tag{9.2}$$

where $R(W,\eta) : \mathcal{W} \bigtimes \mathcal{T} \to \mathcal{R}^L$ is a known measurable map that maps a data vector W and a square-integrable vector-function η into a subset of $\mathcal{R}^L$. Define the bias correction term $\alpha(W,p,\xi(p))$ for the moment $m(W,p,\eta)$ as

$$\alpha(W,p,\xi(p)) := -\gamma(p,X)I(X)^{-1}R(W,\eta(X)), \tag{9.3}$$

where the nuisance parameter $\xi(p) = \xi(p,x)$ is a P-square integrable vector-valued function of x $\xi(p,x) = \{\gamma(p,x), I(x), \eta(x)\}$. The true value $\xi_0(p,x)$ of $\xi(p,x)$ is

$$\xi_0(p,x) = \{\eta_0(x), \gamma_0(p,x), I_0(x)\},$$

where $\eta_0(x)$ is the original functional parameter defined by (9.2), $\gamma_0(p,x) = \partial_{\eta_0(x)}\mathbb{E}[m(W,p,\eta_0)|X = x]$, and $I_0(x) := \partial_{\eta_0}\mathbb{E}[R(W,\eta)|X = x]$ is the Gateaux derivative of the expected generalized residual $\mathbb{E}[R(W,\eta)|X]$ with respect to η conditionally on X. Furthermore, the function $g(W,p,\xi(p))$ in

[6]Lemma 10 was co-developed in the co-authored project "Plug-in Regularized Estimation of High-Dimensional Parameters in Nonlinear Semiparametric Models" with Vasilis Syrgkanis, Denis Nekipelov, and Victor Chernozhukov.

(10.10) has zero Gateaux derivative with respect to $\xi(p)$ at $\xi_0(p)$ uniformly on $\mathcal{P}$

$$\partial_{\xi_0(p)}\mathbb{E}g(W,p,\xi_0(p))[\xi(p)-\xi_0(p)]=0 \quad \forall p\in\mathcal{P}.$$

Lemma 11 is a special case of Lemma 10 when $R(W,\eta(X))=U-\eta(X)$. This result is an extension of Newey (1994)'s result to the set-identified case.

Lemma 11 (Bias Correction Term for Conditional Expectation Function)**.** *Suppose the true value $\eta_0(X)$ of a functional parameter $\eta=\eta(X)$ is the conditional expectation of an observed random variable U given X*

$$\eta_0(x)=\mathbb{E}[U|X=x].$$

Define the bias correction term $\alpha(W,p,\xi(p))$ for the moment $m(W,p,\eta)$

$$\alpha(W,p,\xi(p)):=\gamma(p,X)[U-\eta(X)],$$

where $\xi(p)=\xi(p,x)$ is a P-square integrable vector-valued function of x $\xi(p,x)=\{\eta(x),\gamma(p,x)\}$. The true value $\xi_0(p,x)$ is equal to

$$\xi_0(p,x)=\{\eta_0(x),\gamma_0(p,x)\},$$

where $\gamma_0(p,x)$ is the expectation function conditional on X of the moment derivative

$$\gamma_0(p,x):=\partial_\eta\mathbb{E}[m(W,p,\eta_0)|X=x].$$

Then, the function $g(W,p,\xi(p))$ in (9.1) has zero Gateaux derivative with respect to ξ at ξ_0 for each $p\in\mathcal{P}$

$$\partial_\xi\mathbb{E}g(W,p,\xi_0(p))[\xi-\xi_0]=0 \quad \forall p\in\mathcal{P}.$$

Lemma 11 is a special case of Lemma 10 when $R(W,\eta(X))=1_{U\leq\eta(X)}-u_0$. This result is an extension of Ichimura and Newey (2017)'s result (Proposition 7) to the set-identified case.

Lemma 12 (Bias Correction Term for Conditional Quantile Function)**.** *Suppose the true value* $\eta_0(X)$ *of the functional parameter* $\eta(X)$ *is the conditional quantile of an observed random variable* U *given* X *at a given quantile level* $u_0 \in (0,1)$

$$\eta_0(X) = Q_{U|X=x}(u_0, x).$$

Define the bias correction term $\alpha(W, p, \xi(p))$ *for the moment* $m(W, p, \eta)$

$$\alpha(W, p, \xi(p)) = -\gamma(p, X)\frac{1_{U \leq \eta(X)} - u_0}{l(X)},$$

where $\xi(p,x)$ *is a P-square integrable vector-valued function of* p *and* x $\xi(p,x) = \{\eta(x), \gamma(p,x), l(x)\}$. *The true value* $\xi_0(p,x)$ *is equal to*

$$\xi_0(p,x) = \{\eta_0(x), \gamma_0(p,x), f_{U|X}(\eta_0(X))\},$$

where $\gamma_0(p,x)$ *is the expectation function conditional on* X *of the moment derivative*

$$\gamma_0(p,x) = \partial_\eta \mathbb{E}[m(W, p, \eta_0)|X = x]$$

and $f_{U|X}(\eta_0(X))$ *is the conditional density of* U *given* X *evaluated at* $\eta_0(X)$. *Then, the function* $g(W, p, \xi(p))$ *in (9.1) has zero Gateaux derivative with respect to* ξ *at* ξ_0 *for each* $p \in \mathcal{P}$

$$\partial_\xi \mathbb{E} g(W, p, \xi_0)[\xi - \xi_0] = 0 \quad \forall p \in \mathcal{P}.$$

Lemma 13 discusses the empirically relevant case where there are multiple components appearing in an initial moment condition (10.9).

Lemma 13 (Additive Structure of bias correction Term)**.** *Suppose* $\eta_0(X)$ *is an L-dimensional vector-function. Suppose each of its L distinct components* $l \in \{1, 2, \ldots, L\}$ *is defined by a separate exclusion restriction:* $\mathbb{E}[R_l(W, \eta_{l,0}(X))|X] = 0, l \in \{1, 2, \ldots, L\}$. *Then, the bias correction*

term $\alpha(W,p,\xi(p))$ is equal to the sum of L bias correction terms $\{1,2,\ldots,L\}$

$$\alpha(W,p,\xi(p)) = \sum_{l=1}^{L} \alpha_l(W,p,\xi_l(p)), \tag{9.4}$$

where each term $\alpha_l(W,p,\xi_l(p))$ corrects for the estimation of $\eta_l, l \in \{1,2,\ldots,L\}$ holding the other components η_{-l} fixed at their true value $\eta_{-l,0}$. The new nuisance function $\xi(p)$ is equal to the union $\cup_{l=1}^{L}\xi_l(p)$: $\xi = \cup_{l=1}^{L}\xi_l(p)$.

Lemma 13 is an extension of Newey (1994)'s result to the set-identified case.

Lemmas 11, 12, and 13 give a general recipe for the construction of the bias correction term $\alpha(W,p,\xi(p))$ starting from the moment condition (3.4), which is not orthogonal. Let η be an L-dimensional vector. First, for each $l \in \{1,2,\ldots,L\}$ I derive a bias correction term $\alpha_l(W,p,\xi_l(p))$ as if the nuisance parameter $\eta_{-l,0}$ were known. Then, the bias correction term $\alpha(W,p,\xi(p))$ is the sum of these L bias correction terms, and the new nuisance parameter $\xi(p)$ is the union $\cup_{l=1}^{L}\xi_l(p)$ of the nuisance parameters of each of the L terms.

In several applications, including the support function problem, the nuisance parameter η appears inside the weighting variable V defined in (2.2). As a result, the moment equation (3.4) depends on η in a non-smooth way. In particular, $V = V_\eta$ appears inside a function $x \to x1_{x>0}$ whose first derivative $1_{x>0}$ is not a differentiable function of x at $x = 0$.

I resolve this problem in two steps. First, I show that the difference between the expectations of the target function

$$m(W,p,\eta) = p^\top V_\eta (Y_L + (Y_U - Y_L)1_{\{p^\top V_\eta > 0\}})$$

and its smooth analog

$$m_0(W,p,\eta) = p^\top V_\eta (Y_L + (Y_U - Y_L)1_{\{p^\top V_{\eta_0} > 0\}})$$

is negligible under regularity conditions. Second, I derive the bias correction term for the smooth moment function $m_0(W,p,\eta)$. Lemma 9 provides the sufficient conditions for the first step. Lemmas 11, 12, and 13 give an orthogonalization recipe for the second step.

An argument similar to Lemma 9 was used to establish the consistency and asymptotic nor-

mality of Censored Least Absolute Deviation in Powell (1984).

10 Proofs

10.1 Proof of Section 2

Lemma 14 (Derivation of Equation (2.3))**.** *Let the Assumptions* 1 *and* 2*a of Lee (2008) hold. Let* $\Pr(D=0)=\Pr(D=1)=\frac{1}{2}$ *hold. Then, the bounds given in Equation (2.3) coincide with the bounds given in Proposition 1b of Lee (2008).*

Proof. The lower bound of Lee (2008) (Proposition 1b) is given by

$$\begin{aligned}
&\int_{x\in\mathcal{X}} f(x|D=0,S=1)\mathbb{E}[Y|D=1,S=1,Y\leq y_{\{p_0(x),x\}},X=x] \\
&=\int_{x\in\mathcal{X}} \frac{f(x|D=0,S=1)}{f(x|D=1,S=1)}\mathbb{E}[Y|D=1,S=1,Y\leq y_{\{p_0(x),x\}},X=x]f(x|D=1,S=1) \\
&=\int_{x\in\mathcal{X}} \frac{f(x|D=0,S=1)}{f(x|D=1,S=1)}\frac{1}{p_0(x)}\mathbb{E}[Y1_{\{Y\leq y_{\{p_0(x),x\}}\}}|D=1,S=1,X=x]f(x|D=1,S=1), \quad (10.1)
\end{aligned}$$

where $p_0(X)$ in my notation is $1-p(x)$ in Lee (2008)'s notation. Bayes' rule implies

$$\frac{f(x|D=0,S=1)}{f(x|D=1,S=1)}=\frac{\Pr(X=x,D=0,S=1)\Pr(D=1,S=1)}{\Pr(D=0,S=1)\Pr(X=x,D=1,S=1)} \quad (10.2)$$

Definition of $p_0(X)$ and Bayes' rule imply

$$\begin{aligned}
p_0(x)&=\frac{\Pr(S=1|D=0,X=x)}{\Pr(S=1|D=1,X=x)} \quad (10.3)\\
&=\frac{\Pr(S=1,D=0,X=x)\Pr(D=1|X=x)}{\Pr(S=1,D=1,X=x)\Pr(D=0|X=x)}
\end{aligned}$$

Plugging (10.2) and (10.3) into (10.1) gives (2.4)

$$\begin{aligned}
&\int_{x\in\mathcal{X}} \frac{f(x|D=0,S=1)}{f(x|D=1,S=1)}\frac{1}{p_0(x)}\mathbb{E}[Y1_{\{Y\leq y_{\{p_0(x),x\}}\}}|D=1,S=1,X=x]f(x|D=1,S=1) \quad (10.4)\\
&=\mathbb{E}\frac{D\cdot S\cdot Y\cdot 1_{\{Y\leq y_{\{p_0(x),x\}}\}}\Pr(D=0|X=x)}{\Pr(D=0,S=1)\Pr(D=1|X=x)} \quad (10.5)
\end{aligned}$$

The proof for the upper bound is similar.

□

Lemma 15 (Equivalence of Long and Short Definitions of the Partially Linear Predictor)**.** *Suppose the matrix* $\Sigma = \mathbb{E}[D - \eta_0(X)][D - \eta_0(X)]'$ *is invertible. Then, the identified set* $\mathcal{B}$ *given by:*

$$\mathcal{B} = \{\beta = \arg \min_{b \in \mathcal{R}^d, f \in \mathcal{M}} \mathbb{E}(Y - D^\top b - f(X))^2, \quad Y_L \leq Y \leq Y_U\} \tag{10.6}$$

coincides with the identified set given by (2.8).

Proof. Fix a random variable Y in a random interval $[Y_L, Y_U]$. Let us show that the minimizer β_0 of (10.6) coincides with the minimizer β_0^s defined as:

$$\beta_0^s = \arg \min_{b \in \mathcal{R}^d, f \in \mathcal{M}} \mathbb{E}(Y - (D - \eta_0(X))'b)^2, \tag{10.7}$$

where $\eta_0(X) = \mathbb{E}[D|X]$. For each b in (10.6) we solve for $f(X) = f_b(X)$ as a function of b. The solution $f_b(X)$ is a conditional expectation function:

$$f_b(X) = \mathbb{E}[Y - Db|X] = \mathbb{E}[Y|X] - \eta_0(X).$$

Substituting $f_b(X)$ into (10.6) gives:

$$\beta = \arg \min_{b \in \mathcal{R}^d} \mathbb{E}(Y - \mathbb{E}[Y|X] - (D - \eta_0(X))'b)^2. \tag{10.8}$$

Expanding $(m+n)^2 = m^2 + 2mn + n^2$ with $m = Y - (D - \eta_0(X))'b$ and $n = \mathbb{E}[Y|X]$ gives:

$$\begin{aligned}
\beta =^{i} & \arg \min_{b \in \mathcal{R}^d} \mathbb{E}(Y - (D - \eta_0(X))'b)^2 \\
& - 2\mathbb{E}(Y - (D - \eta_0(X))'b)\mathbb{E}[Y|X] \\
& + \mathbb{E}(\mathbb{E}[Y|X])^2 \\
=^{ii} & \arg \min_{b \in \mathcal{R}^d} \mathbb{E}(Y - (D - \eta_0(X))'b)^2 - \mathbb{E}(\mathbb{E}[Y|X])^2 \\
=^{iii} & \arg \min_{b \in \mathcal{R}^d} \mathbb{E}(Y - (D - \eta_0(X))'b)^2
\end{aligned}$$

Since $\mathbb{E}[(D-\eta_0(X))'b]\mathbb{E}[Y|X]=0$ and $\mathbb{E}(Y-(D-\eta_0(X))'b)\mathbb{E}[Y|X]=\mathbb{E}[Y|X]^2$, *ii* follows. Since $\mathbb{E}[Y|X]^2$ does not depend on b, *iii* follows. The solution to the minimization problem in *iii* coincides with β_0^s in (2.8). According to Bontemps et al. (2012) (Proposition 2), the set (2.8) is a sharp identified set for β_0.

□

10.2 Proofs of Section 9

Below we present examples of bias correction terms $\alpha(W,p,\xi(p))$ for various types of functional nuisance parameter η. In order to guess a general form of these bias correction terms we have relied on the previous work that used semiparametric efficiency theory to produce an efficient (and, therefore, Neyman-orthogonal) score. In particular, a general form of a bias correction term for conditional expectation functions is given in Newey (1994) and for average partial derivatives in Hardle and Stoker (1989).

Adding a level of generality helps to understand the derivation. Suppose one is interested in the function $M(p)$ defined by the following equation:

$$M(p)-\mathbb{E}m(W,p,\eta_0)=0, \tag{10.9}$$

where $m(W,p,\eta):\mathcal{W}\bigtimes\mathcal{P}\bigtimes\mathcal{T}\to\mathcal{R}^{\dim(m)}$ is a measurable moment function. We constructed a bias correction term $\alpha(W,p,\xi(p))$ such that the function:

$$g(W,p,\xi(p))=m(W,p,\eta)+\alpha(W,p,\xi(p) \tag{10.10}$$

obeys orthogonality condition with respect to ξ at $\mathbf{x}_0$ for all $p\in\mathcal{P}$.

Proof of Lemma 10. Fix a vector p in $\mathcal{P}$. Let us show that the Gateaux derivative of $\mathbb{E}[g(W,p,\xi(p)]$

with respect to $\xi(p)$ at $\xi_0(p)$ is equal to zero. The derivative with respect to η at η_0 is:

$$\begin{aligned}\partial_{\eta_0}\mathbb{E}g(W,p,\xi_0(p)) &= \partial_{\eta_0}\mathbb{E}m(W,p,\eta_0(X))[\eta(X)-\eta_0(X)] \\ &\quad - \gamma_0(p,X)I_0(X)^{-1}\partial_{\eta_0}\mathbb{E}R(W,\eta_0)[\eta(X)-\eta_0(X)] \\ &=^{i} \mathbb{E}_X\left[\partial_{\eta_0}\mathbb{E}[m(W,p,\eta_0(X))|X] - \gamma_0(p,X)I_0(X)^{-1}I_0(X)[\eta(X)-\eta_0(X)]\right] \\ &=^{ii} 0,\end{aligned}$$

where equality i follows from the definition of $I_0(X) = \partial_{\eta_0}\mathbb{E}[R(W,\eta_0)|X]$ and equality ii follows from the definition of $\gamma_0(p,X) = \partial_{\eta_0}\mathbb{E}[m(W,p,\eta_0(X))|X]$. The derivative with respect to $I(X)$ at I_0 is:

$$\begin{aligned}\partial_{I_0}\mathbb{E}g(W,p,\xi_0(p)) &= -\mathbb{E}_X I_0(X)^{-2}\gamma_0(p,X)\mathbb{E}R(W,\eta_0(X))[I(X)-I_0(X)] \\ &= 0\end{aligned}$$

by Equation (9.2). The derivative with respect to $\gamma(p,\cdot)$ at $\gamma_0(p,\cdot)$ is:

$$\begin{aligned}\partial_{\gamma_0}\mathbb{E}g(W,p,\xi_0(p)) &= -\mathbb{E}_X I_0(X)^{-1}\mathbb{E}R(W,\eta_0(X))[\gamma(p,X)-\gamma_0(p,X)] \\ &= 0\end{aligned}$$

by Equation (9.2). □

Lemma[7] 10 derives a general form of a bias correction term for the case $\eta_0(X)$ is defined via the conditional exogeneity restriction (9.2). In our applications, we consider two important cases of this Lemma: a conditional expectation function (Lemma 11) and a conditional quantile function (Lemma 12), respectively.

Lemma 11 is a special case of Lemma 10 with $R(W,\eta(X)) = U - \eta(X)$.

Proof of Lemma 11. Consider the setup of Lemma 10 with

$$R(W,\eta(X)) := U - \eta(X).$$

[7]Lemma 10 was co-developed in the co-authored project "Plug-in Regularized Estimation of High-Dimensional Parameters in Nonlinear Semiparametric Models" with Vasilis Syrgkanis, Denis Nekipelov, and Victor Chernozhukov.

Then, $I_0(X) := \partial_{\eta_0}\mathbb{E}[R(W,\eta_0(X))] = -1$ and is bounded away from zero a.s. in X. Therefore, by Lemma 10 the bias correction term is equal to

$$\alpha(W,p,\xi(p)) := \gamma(p,X)(U-\eta(X)),$$

where $\gamma_0(p,X) := \partial_{\eta_0}\mathbb{E}[m(W,p,\eta_0)|X]$.

□

Lemma 12 is a special case of Lemma 10 with $R(W,\eta(X)) = 1_{\{U\leq\eta(X)\}} - u_0$, where $u_0 \in (0,1)$ is a given quantile level.

Proof of Lemma 12. Let $u_0 \in (0,1)$ be a given quantile level. Suppose the true value $\eta_0(X)$ of the nuisance parameter $\eta(X)$ is the conditional quantile function $\eta_0(X) = Q_{U|X=x}(u_0,x)$ of level u_0. Consider the setup of Lemma 10 with

$$R(W,\eta) := 1_{U\leq\eta(X)} - u_0.$$

Then,

$$I_0(X) := \partial_{\eta_0}\mathbb{E}[1_{U\leq\eta_0(X)}|X] = f_{U|X}(\eta_0(X))$$

where $f_{U|X}(\eta_0(X))$ is the conditional density of U given X evaluated at $\eta_0(X)$. By Assumptions of Lemma 12, $f_{U|X}(\eta_0(X))$ is bounded away from zero a.s. in X. Therefore, by Lemma 10 the bias correction term is equal to

$$\alpha(W,p,\xi(p)) := -\gamma(p,X)\frac{1_{U\leq\eta(X)} - u_0}{l(X)},$$

where $\gamma_0(p,X) := \partial_{\eta_0}\mathbb{E}[m(W,p,\eta_0)|X]$ and $l_0(X) = f_{U|X}(\eta_0(X))$.

□

Lemma 16 (Bias Correction term for Average Partial Derivative)**.** *Suppose the true value $\eta_0(D,X)$ of the functional parameter $\eta(D,X)$ is the gradient of the logarithm of the conditional density of D given X: $\eta_0(D,X) = \partial_D \log f_0(D|X) = \partial_d \log f(D=d|X)$. Let the moment function be:*

$$m(W,p,\eta(D,X)) := p^\top \eta(D,X)Y.$$

Define the bias correction term

$$\alpha(W,p,\xi(p)) := p^\top[-\eta(D,X)\mu(D,X) + \partial_D\mu(D,X)],$$

where $\xi(p) = \xi(D,X) := \{\eta(D,X), \mu(D,X)\}$ *consists of two P-square-integrable function of* D,X *which do not depend on p. Moreover, the true value of* $\eta(D,X)$ *is* $\eta_0(D,X) = \partial_d \log f(D=d|X)$*, and that of* $\mu(D,X)$ *is* $\mu_0(D,X) := \mathbb{E}[Y|D,X]$*. Then, the moment function* $g(W,p,\xi(p))$ *in (10.10) has a zero Gateaux derivative with respect to* $\xi(p)$ *at* $\xi_0(p)$ *uniformly on* $\mathcal{P}$*:*

$$\partial_\xi \mathbb{E}g(W,p,\xi_0)[\xi - \xi_0] = 0, \quad \forall p \in \mathcal{P}.$$

Lemma 16 is the extension of Hardle and Stoker (1989) to set-identified case.

Proof of Lemma 13. Consider the setup of Lemma 10. Since each component $\eta_l(X), l \in \{1,2,\ldots,L\}$ is defined by a separate exclusion restriction, the matrix $I_0(X)$ is a diagonal matrix whose l'th element on the diagonal is equal to $I_{ll,0}(X) := \partial_{\eta_{l,0}}\mathbb{E}[R_l(W,\eta_{l,0}(X))|X]$. Therefore,

$$\begin{aligned}
\alpha(W,p,\xi(p)) &= \gamma(p,X)I(X)^{-1}R(W,\eta(X)) \\
&= \sum_{l=1}^{L} \gamma_l(p,X)I_{ll}(X)^{-1}R_l(W,\eta_l(X)) \\
&= \sum_{l=1}^{L} \alpha(W,p,\xi_l(p)).
\end{aligned}$$

□

Proof of Lemma 16. Let $\xi(p,X) = \xi(X) = \{\eta(D,X), \mu(D,X)\}$ be a P-square integrable vector-valued function that does not depend on p.

$$g(W,p,\xi) = p^\top \eta(D,X)Y + p^\top[-\eta(D,X)\mu(D,X) + \partial_D\mu(D,X)].$$

The first-order Gateaux derivative of $\mathbb{E}g(W,p,\xi)$ w.r.t ξ at ξ_0 is equal to:

$$\begin{aligned}\partial_{\xi_0}\mathbb{E}g(W,p,\xi_0)[\xi-\xi_0] &= \begin{bmatrix} p^\top \mathbb{E}[\eta(D,X)-\partial_D \log f_0(D|X)][Y-\mu_0(D,X)] \\ \mathbb{E}[\partial_D[\mu(D,X)-\mu_0(D,X)]-\eta_0(D,X)[\mu(D,X)-\mu_0(D,X)]]\end{bmatrix} \\ &= \begin{bmatrix}0\\0\end{bmatrix},\end{aligned}$$

where the second equality follows from integration by parts:

$$\begin{aligned}&\mathbb{E}[\partial_D[\mu(D,X)-\mu_0(D,X)]-\eta_0(D,X)[\mu(D,X)-\mu_0(D,X)]] \\ &= \mathbb{E}[\eta_0(D,X)[\mu(D,X)-\mu_0(D,X)]-\eta_0(D,X)[\mu(D,X)-\mu_0(D,X)] \\ &= 0.\end{aligned}$$

□

Proof of Lemma 9. Fix an element η in the realization set $\mathcal{T}_N$. Define an event of an incorrectly chosen sign in the indicator function $1_{\{\cdot\}}$:

$$\mathcal{E} := \{1_{p^\top V_\eta>0} \neq 1_{p^\top V_{\eta_0}>0}\} = \{p^\top V_\eta > 0 > p^\top V_{\eta_0}, \quad p^\top V_\eta < 0 < p^\top V_{\eta_0}\}.$$

Recognize that the event $\mathcal{E}$ is included into the event

$$\mathcal{E}_{\eta\eta_0} := \{0 < |p^\top V_{\eta_0}| < |p^\top V_\eta - p^\top V_{\eta_0}|\},$$

that is: $\mathcal{E} \subseteq \mathcal{E}_{\eta\eta_0}$ a.s. . Therefore, the contribution of an incorrectly chosen sign in the indicator function admits the following bound, where we dropped the dependence of Y_L and Y_U on η in the notation:

$$
\begin{aligned}
&\left|\mathbb{E}p^\top V_\eta (Y_U - Y_L)(1_{\{p^\top V_\eta > 0\}} - 1_{\{p^\top V_{\eta_0} > 0\}})\right| \\
&= \left|\mathbb{E}p^\top V_\eta (Y_U - Y_L) 1_{\{\mathcal{E}\}}\right| && \text{(Definition of } \mathcal{E}) \\
&\leq \left|\mathbb{E}p^\top V_\eta (Y_U - Y_L) 1_{\{\mathcal{E}_{\eta\eta_0}\}}\right| && (\mathcal{E} \subseteq \mathcal{E}_{\eta\eta_0} \text{ a.s. }) \\
&\leq B_{UL}\left|\mathbb{E}p^\top V_\eta 1_{\{\mathcal{E}_{\eta\eta_0}\}}\right| && \text{(Assumption (a))} \\
&\leq B_{UL}\left|\mathbb{E}p^\top V_{\eta_0} 1_{\{\mathcal{E}_{\eta\eta_0}\}}\right| + B_{UL}\left|\mathbb{E}p^\top (V_\eta - V_{\eta_0}) 1_{\{\mathcal{E}_{\eta\eta_0}\}}\right| && (V_\eta = V_{\eta_0} + V_\eta - V_{\eta_0}) \\
&:= B_{UL}(i + ii)
\end{aligned}
$$

The first-order term is bounded by an L_2-bound of the error $V_\eta - V_{\eta_0}$

$$
i = \mathbb{E}|p^\top V_{\eta_0}| 1_{\{0 < |p^\top V_{\eta_0}| \leq |p^\top V_\eta - p^\top V_{\eta_0}|\}} \leq \mathbb{E}\|p^\top (V_\eta - V_{\eta_0})\|^2 \leq \sup_{p\in\mathcal{P}} \|p\| \mathbb{E}(V_\eta - V_{\eta_0})\|^2.
$$

The second-order term is bounded by Assumption (2):

$$
ii = \mathbb{E}|p^\top V_\eta - p^\top V_{\eta_0}| 1_{\{0 < |p^\top V_{\eta_0}| \leq |p^\top V_\eta - p^\top V_{\eta_0}|\}} \leq \mathbb{E}\|V_\eta - V_{\eta_0}\|^2
$$

Therefore, $i + ii \lesssim \mathbb{E}\|V_\eta - V_{\eta_0}\|^2 = o(N^{-1/2})$. □

Lemma 17 (From Zero Gateaux Derivative to Uniform Near Orthogonality)**.** *Let $M(p)$ be a target function defined by (10.9) on a compact set $\mathcal{P}$. Suppose Assumption 4 holds on $\mathcal{P}$. Moreover, the moment condition (10.10) has zero Gateaux derivative for all vectors p in $\mathcal{P}$. Then, the moment condition (10.10) satisfies Assumption 4 uniformly on $\mathcal{P}$.*

Proof. We repeat the proof of Step 2, Lemma 6.3 in Chernozhukov et al. (2017a). Consider a Taylor expansion of the function $r \to \mathbb{E}[g(W, p, r(\xi(p) - \xi_0(p)) + \xi_0(p))], \quad r \in (0,1)$.

$$
\begin{aligned}
&\sup_{p\in\mathcal{P}} \sup_{\xi\in\Xi_n} |\mathbb{E}[g(W,p,\xi(p)) - g(W,p,\xi_0(p))]| \\
&\qquad \leq \sup_{p\in\mathcal{P}} \sup_{\xi\in\Xi_n} |\partial_{\xi_0} \mathbb{E}g(W,p,\xi_0(p))[\xi(p) - \xi_0(p)]| \\
&\qquad + \sup_{p\in\mathcal{P}} \sup_{\xi\in\Xi_n} \sup_{r\in[0,1)} \left| \int_0^1 2^{-1} \partial_r^2 \mathbb{E}g(W,p,r(\xi(p) - \xi_0(p)) + \xi_0(p)) dr \right| \\
&\qquad \leq r_N = o(N^{-1/2}).
\end{aligned}
$$

□

Lemma 18 (Achieving Small Bias Assumption for Support Function)**.** *Suppose the conditions of Lemma 9 hold. Let $m_0(W,p,\eta)$ be a smoothened analog of the support function moment (3.6) defined as:*

$$m_0(W,p,\eta) := p^\top V_\eta \Gamma(Y_{L,\eta}, Y_{U,\eta} - Y_{L,\eta}, p^\top V_{\eta_0}).$$

Let $\alpha_0(W,p,\xi(p))$ be a bias correction term for $m_0(W,p,\eta)$ such that $m_0(W,p,\eta)+\alpha_0(W,p,\xi(p))$ obeys orthogonality condition with respect to ξ at ξ_0 for each $p \in \mathcal{P}$. Then, the moment function

$$g(W,p,\xi(p)) := m(W,p,\eta) + \alpha_0(W,p,\xi(p))$$

satisfies Assumption 3.

Proof of Lemma 18. Let Ξ be a space of P-square integrable functions, $\xi_0(p,X)$ be a functional nuisance parameter that depends on p, and Ξ_n be a sequence of realization sets around $\xi_0(p,X)$.

$$\begin{aligned}
&\sup_{p\in\mathcal{P}} \sup_{\xi\in\Xi_n} \mathbb{E}[g(W,p,\xi(p)) - g(W,p,\xi_0(p))] \\
&\leq \sup_{p\in\mathcal{P}} \sup_{\xi\in\Xi_n} |\mathbb{E}[p^\top V(\eta)\Gamma(Y_L, Y_U - Y_L, p^\top V(\eta)) - p^\top V(\eta)\Gamma(Y_L, Y_U - Y_L, p^\top V(\eta_0))]| \\
&+ \sup_{p\in\mathcal{P}} \sup_{\xi\in\Xi_n} |\mathbb{E}[p^\top V(\eta)\Gamma(Y_L, Y_U - Y_L, p^\top V(\eta_0)) + \alpha_0(W,p,\xi(p))]| \\
&\leq \mathbb{E}\|V_\eta - V_{\eta_0}\|^2 + r_N
\end{aligned}$$

where the bound on the first summand is by Lemma 9 and on the second one by Lemma 17. □

Lemma 19 (Maximal Inequality for Support Function)**.** *Let R and C be positive constants. Assume there exist $c > 2$ such that the vector $(V_\eta, Y_{L,\eta}, Y_{U,\eta})$ is $L_{P,c}$-integrable: $\|(V_\eta, Y_{L,\eta}, Y_{U,\eta})\|_{L_{P,c}} \leq C < \infty$. Then, the function class $\mathcal{F}_\eta = \{p^\top V(\eta)\Gamma(Y_{L,\eta}, Y_{U,\eta} - Y_{L,\eta}, p^\top V(\eta)), \quad \|p - p_0(q)\| \leq RN^{-1/2}, q \in \mathcal{S}^{d-1}\}$ satisfies Assumption 4(1,2,3(a)). If, in addition, the function class*

$$\mathcal{A}_\xi = \{\alpha_0(W,p,\xi(p)), \quad \|p - p_0(q)\| \leq RN^{-1/2}, q \in \mathcal{S}^{d-1}\}$$

satisfies Assumption 4, so does the class $\mathcal{R}_g = \{g(W,p,\xi(p)), \quad \|p - p_0(q)\| \leq RN^{-1/2}, q \in \mathcal{S}^{d-1}\}$.

Proof. Let $\eta \in \Xi_n$ be a fixed element of the nuisance realization set. Let $F_\eta = \lambda_{\max}\|V\|_\eta(|Y_{L,\eta}| + |Y_{U,\eta}|)$ be a measurable envelope. By the condition of the Lemma, there exists $c > 2$ such that $\|F\|_{L_{P,c}} \leq C$. Next, recognize that the function class

$$\mathcal{F}_\eta = \mathcal{F}_{L,\eta} + \mathcal{F}_{U-L,\eta},$$

where

$$\mathcal{F}_{L,\eta} = \{p^\top V_\eta Y_{L,\eta}, \quad \|p - p_0(q)\| \leq RN^{-1/2}, q \in \mathcal{S}^{d-1}\}$$

and

$$\mathcal{F}_{U-L,\eta} = \{p^\top V_\eta (Y_{U,\eta} - Y_{L,\eta}) 1_{\{p^\top V_\eta > 0\}}, \quad \|p - p_0(q)\| \leq RN^{-1/2}, q \in \mathcal{S}^{d-1}\}.$$

First, the linear class $\mathcal{L}_\eta := \{p^\top V_\eta, \quad \|p - p_0(q)\| \leq RN^{-1/2}, q \in \mathcal{S}^{d-1}\}$ and the class of the indicators:

$$\mathcal{I}_\eta := \{1_{\{p^\top V_\eta > 0\}}, \quad \|p - p_0(q)\| \leq RN^{-1/2}, q \in \mathcal{S}^{d-1}\}$$

have bounded uniform covering entropy (UCE), respectively:

$$\log \sup_Q N(\varepsilon, \mathcal{L}_\eta, \|\cdot\|_{Q,2}) \leq d \log(a/\varepsilon), \quad \text{for all } 0 < \varepsilon \leq 1,$$

$$\log \sup_Q N(\varepsilon, \mathcal{I}_\eta, \|\cdot\|_{Q,2}) \leq d \log(a/\varepsilon), \quad \text{for all } 0 < \varepsilon \leq 1.$$

The conclusions below follow from Lemma 8.3 of Chandrasekhar et al. (2011). The multiplication of the class $\mathcal{L}_\eta$ by a random variable $Y_{L,\eta}$ preserves UCE. Third, the product of the classes $\mathcal{L}_\eta \cdot \mathcal{I}_\eta$ has bounded UCE. Therefore, $\mathcal{F}_{L,\eta}$ and $\mathcal{F}_{U-L,\eta}$ have bounded UCE. Finally, the sum of the classes $\mathcal{F}_{L,\eta}, \mathcal{F}_{U-L,\eta}, A_\xi$ has bounded UCE. □

10.3 Proofs of Section 4

Proof of Theorem 3. The nuisance parameter $\eta = \{\eta_1(X), \eta_2(X), \eta_3(u,X)\}$ whose true value

$$\eta_0 = \{s(0,X), s(1,X), Q_{Y|D=1,S=1,X}(u,X)\}.$$

The notation $\eta_{-k,0}, k \in \{1,2,3\}$ stands for the true value of η_{-k} obtained from η by excluding the k'th component of the nuisance parameter. Step 1. Derivation of $\alpha_1(W,\eta)$. The nuisance parameter $s(0,X)$ appears in (10.12) inside the quantile $y_{\{1-s(0,X)/s(1,X),X\}}$ and in the denominator of (10.12). The Gateaux derivative of the function $\mathbb{E}[m_U(W,\eta)|X]$ with respect to η_1 at $\eta_{1,0} = s(0,X)$ is equal to:

$$\partial_{\eta_1}\mathbb{E}[m_U(W,\eta_{1,0};\eta_{-1,0})|X] =$$
$$\partial_{\eta_1}\mathbb{E}[Y 1_{Y \geq Q_{U|X}(1-\eta_{1,0})/s(1,X),X)}|X, D=1, S=1]\frac{s(1,X)\Pr(D=0)}{\Pr(S=1,D=0)}$$

$$= y_{\{1-s(0,X)/s(1,X)\}}\frac{\Pr(D=0)}{\Pr(S=1,D=0)}$$

The application of Lemma 11 gives the bias correction term:

$$\alpha_1(W,\eta) = \gamma_1(X)\left(\frac{(1-D)S}{\Pr(D=0)} - \eta_1(X)\right),$$

where the true value of $\gamma_{1,0}(X)$ equals to:

$$\gamma_{1,0}(X) = y_{\{1-s(0,X)/s(1,X)\}}\frac{\Pr(D=0)}{\Pr(S=1,D=0)}.$$

Step 2. Derivation of $\alpha_2(W,\eta)$. The nuisance parameter $s(1,X)$ appears inside the quantile function in Equation (10.12). The Gateaux derivative of the function $\mathbb{E}[m_U(W,\eta)|X]$ with respect to η_2 at $\eta_{2,0} = s(1,X)$ is equal to:

$$\partial_{\eta_2}\mathbb{E}m_U(W,\eta_{2,0})|X] = -\frac{y_{\{1-s(0,X)/s(1,X),X\}}s(0,X)\Pr(D=0)}{s(1,X)\Pr(S=1,D=0)}.$$

The application of Lemma 11 gives the bias correction term:

$$\alpha_2(W,\eta) = \gamma_2(X)\left(\frac{DS}{\Pr(D=1)} - s(1,X)\right),$$

where the true value of $\gamma_{2,0}(X)$ is equal to:

$$\gamma_{2,0}(X) = -\frac{y_{\{1-s(0,X)/s(1,X),X\}} s(0,X)\Pr(D=0)}{s(1,X)\Pr(S=1,D=0)}.$$

Step 3. Derivation of $\alpha_3(W,\eta)$. The nuisance parameter $\eta_3(u,x) = Q_{Y|D=1,S=1,X=x}(u,x)$ appears in the numerator of (10.12). The application of Lemma 12 gives the bias correction term:

$$\alpha_3(W,\eta) = -\gamma_3(X)\frac{1_{Y\leq\eta_3(1-s(0,X)/s(1,X),X)} - 1 + s(0,X)/s(1,X)}{f_{Y|D=1,S=1,X}(y_{\{1-s(0,X)/s(1,X),X\}})},$$

where the true value of $\gamma_{3,0}(X)$ is equal to:

$$\gamma_{3,0}(X) = -y_{\{1-s(0,X)/s(1,X),X\}} f_{Y|D=1,S=1,X}(y_{\{1-s(0,X)/s(1,X)\}})\frac{s(1,X)\Pr(D=0)}{\Pr(D=0,S=1)}.$$

Therefore,

$$\alpha_3(W,\eta) = y_{\{1-s(0,X)/s(1,X),X\}}\frac{s(1,X)\Pr(D=0)}{\Pr(D=0,S=1)}\left(1_{\{Y\leq\eta_3(1-s(0,X)/s(1,X),X)\}} - 1 + s(0,X)/s(1,X)\right)$$

Step 4. Define the moment function $m_L(W,\eta)$ for the lower bound as:

$$m_L(W,\eta) := \frac{D\cdot S\cdot Y 1_{\{Y\leq\eta_3(\eta_1/\eta_2,X)\}}\Pr(D=0)}{\Pr(D=0,S=1)\Pr(D=1)}. \tag{10.11}$$

The bias correction term $\alpha_L(W,\eta)$ for β_L is:

$$\alpha_L(W,\eta) = \sum_{i=4}^{6}\alpha_i(W,\eta),$$

where the bias correction terms for individual components are:

$$\begin{aligned}
\alpha_4(W,\eta) &= \gamma_4(X)\left(\frac{(1-D)S}{\Pr(D=0)} - \eta_1(X)\right),\\
\alpha_5(W,\eta) &= \gamma_5(X)\left(\frac{DS}{\Pr(D=1)} - \eta_2(X)\right),\\
\alpha_6(W,\eta) &= -\gamma_6(X)\left(1_{\{Y\leq\eta_3(p_0(X),X)\}} - p_0(X)\right),
\end{aligned}$$

and the true values of the nuisance parameters above are:

$$\gamma_{4,0}(X) = y_{\{s(0,X)/s(1,X),X\}} \frac{\Pr(D=0)}{\Pr(S=1,D=0)}$$

$\gamma_{5,0}(X) = -\frac{y_{\{s(0,X)/s(1,X),X\}} s(0,X)}{s(1,X)} \frac{\Pr(D=0)}{\Pr(S=1,D=0)}$ and $\gamma_{6,0}(X) = y_{p_0(X)} \frac{s(1,X)\Pr(D=0)}{\Pr(D=0,S=1)}$. Define the Neyman-orthogonal moment functions for the upper and the lower bound as

$$g_U(W,\xi_U) = m_U(W,\eta) + \alpha_U(W,\xi_U), \tag{10.12}$$

$$g_L(W,\xi_L) = m_L(W,\eta) + \alpha_L(W,\xi_L), \tag{10.13}$$

where the nuisance parameter ξ_U consists of the original nuisance parameter η and the functions $(\gamma_i)_{i=1}^3$: $\xi_U = \{\eta, (\gamma_i)_{i=1}^3\}$; the nuisance parameter ξ_L consists of the original nuisance parameter η and the functions $(\gamma_i)_{i=4}^6$: $\xi_U = \{\eta, (\gamma_i)_{i=4}^6\}$.

Step 5. To apply Lemma 17 and conclude that Assumption 3 holds, we should verify that the conditional Hessian of the moment function with respect to $\eta(X)$ is bounded a.s. in X.

Step 6. Verification of Assumption 4(3(b)). The first derivative $\partial_\eta \mathbb{E}[m_U(W,\eta_0)|X]$ is a composition of the functions $(t_1,t_2) \to t_1/t_2$, $t \to tf(t)$, where $f(\cdot)$ is the conditional density of Y given $D=1, S=1, X=x$, and the functional elements of Ξ_N^U: $u \to \eta_3(u,x)$. By Assumption 7, each of these functions is bounded and has a bounded first derivative. Therefore, Assumption 4 holds.

Having established the estimates of the bounds on the expected wage $\mathbb{E}[Y_1|S_1=1,S_0=1]$, we proceed to deriving the bounds on the actual Average Treatment Effect on the always-employed:

$$\theta := \mathbb{E}[Y_1 - Y_0 | S_1 = 1, S_0 = 1].$$

The sharp bounds θ_L, θ_U on θ are given by the following moment conditions:

$$\theta_L = g_L(W,\eta) - \frac{(1-D)SY}{\mathbb{E}[D=0|X]s(0,X)} - \mathbb{E}[Y|S=1,D=0,X]\left(\frac{(1-D)S}{\mathbb{E}[D=0|X]} - s(0,X)\right),$$
$$\theta_U = g_U(W,\eta) - \frac{(1-D)SY}{\mathbb{E}[D=0|X]s(0,X)} - \frac{\mathbb{E}[Y|S=1,D=0,X]}{s^2(0,X)}\left(\frac{(1-D)S}{\mathbb{E}[D=0|X]} - s(0,X)\right),$$

where the term $\frac{\mathbb{E}[Y|S=1,D=0,X]}{s^2(0,X)}(\frac{(1-D)S}{1-\mathbb{E}[D=1|X]} - s(0,X))$ is correcting the bias of the estimation of the

function $s(0,X)$. □

Remark 3 (Asymptotic Theory for the Average Treatment Effect in Endogenous Sample Selection of Lee (2008))**.** *Suppose Assumption 7 holds. In addition, suppose the function* $\gamma_0(X) := \mathbb{E}[Y|S = 1, D = 0, X]$ *is a P-square integrable function* $\gamma(X)$ *such that* $\|\gamma - \gamma_0\|_{L_{P,2}} \leq o(N^{-1/4})$*. Then, the Bounds Estimator obeys:*

$$\sqrt{N}\begin{pmatrix} \hat{\theta}_L - \theta_L \\ \hat{\theta}_U - \theta_U \end{pmatrix} \Rightarrow N(0,\Omega), \tag{10.14}$$

where Ω *is a positive-definite covariance matrix.*

Proof of Theorem 4. Step 1. Verification of Assumption 3(1). Consider the original moment function:

$$m(W,q,\eta) = q^\top \eta(D,X)\Gamma(Y_L, Y_U, q^\top \eta(D,X)).$$

First, we apply Lemma 9 (verified in Step 2) to shift to a smoothened moment function

$$m_0(W,q,\eta) = q^\top \eta(D,X)\Gamma(Y_L, Y_U, q^\top \eta_0(D,X)).$$

Second, we apply Lemma 16 to the moment function $m_0(W,q,\eta)$ and derive the bias correction term

$$\alpha_0(W,q,\xi) = -q^\top \eta(D,X)\mu_q(D,X) + q^\top \partial_D \mu_q(D,X),$$

where the nuisance parameter $\xi = \xi(X) = \{\eta(D,X), \gamma_L(D,X), \gamma_{U-L}(D,X)\}$ is a vector-valued *P*-square integrable function that does not depend on q. The true value ξ_0 is:

$$\xi_0(q,X) = \{\partial_D \log f(D|X), \gamma_{L,0}(D,X), \gamma_{U-L,0}(D,X)\}.$$

For each $q \in \mathcal{S}^{d-1}$ define the function

$$\mu_q(D,X) := \gamma_L(D,X) + \gamma_{U-L}(D,X)1_{\{q^\top \eta(D,X)>0\}}.$$

Therefore,

$$g(W,p,\xi) = m_0(W,q,\eta) + \alpha_0(W,q,\xi)$$

has zero Gateaux derivative with respect to ξ at ξ_0. Since the function $m_0(W,q,\eta)$ is linear in $\eta(D,X)$, its second derivative w.r.t $\eta(D,X)$ is equal to zero. Therefore, Assumption 4 holds. Assumption 17 implies that Assumption 3(1) is satisfied.

Step 2. Verification of the conditions of Lemma 9. Assumptions 1-4 of Lemma 9 follows from Assumptions 8 (1)-(3). Step 3. Verification of Assumption 4. Consider the setting of Lemma 19. Consider the function class

$$\mathcal{A}_\xi = \{-q^\top V_\eta \mu_L(D,X) - q^\top V_\eta 1_{\{q^\top V_\eta\}>0}(\mu_U(D,X) - \mu_L(D,X)) + q^\top \partial_D \mu(D,X), q \in \mathcal{S}^{d-1}\}.$$

This class is obtained in three steps:

1. The classes $\mathcal{L}_\eta$ and $\mathcal{I}_\eta$ are multiplied by random variables $\mu_L(D,X)$ and $\mu_U(D,X) - \mu_L(D,X)$, respectively.

2. The elements of the classes $\mathcal{L}_\eta \cdot \mu_L(D,X)$ and $\mathcal{I}_\eta \cdot (\mu_U(D,X) - \mu_L(D,X))$ are summed into $\mathcal{L}_\eta \cdot \mu_L(D,X) + \mathcal{I}_\eta \cdot (\mu_U(D,X) - \mu_L(D,X))$.

3. The class $\mathcal{L}_\eta \cdot \mu_L(D,X) + \mathcal{I}_\eta \cdot (\mu_U(D,X) - \mu_L(D,X))$ is added to a linear class $\{q^\top \partial_D \mu(D,X)\}$, $q \in \mathcal{S}^{d-1}\}$.

By Lemma 8 of Chandrasekhar et al. (2011), Steps 1, 2 and 3 do not change the order of the uniform covering entropy of the function class.

Step 4. Verification of Assumption 4.

$$g(W,q,\xi(q)) - g(W,q,\xi_0(q)) = \underbrace{q^\top(\partial_D \mu_q(D,X) - \partial_D \mu_{q,0}(D,X))}_{K_1} + \underbrace{q^\top(\eta(D,X) - \eta_0(D,X))(Y_{q,\eta(D,X)} - \mu_q(D,X)}_{K_2}$$
$$+ \underbrace{q^\top \eta_0(D,X)(Y_{q,\eta(D,X)} - Y_{q,\eta_0(D,X)})}_{K_3} + \underbrace{q^\top \eta_0(D,X)(\mu_{q,0}(D,X)) - \mu_q(D,X)}_{K_4}.$$

By Assumption 8, the following bounds apply: $\|K_1\|_{L_{P,2}} \leq g_N$, $\|K_2\|_{L_{P,2}} \leq g_N$ and $\|K_4\|_{L_{P,2}} \leq g_N$

where $g_N = o(N^{-1/4})$. Furthermore, the bound on $\|K_3\|_{L_{P,2}}$ is as follows:

$$\begin{aligned}
\|K_3\|_{L_{P,2}} &\leq (\mathbb{E}|q^\top \eta_0(D,X)|^2 1_{0<|q^\top \eta_0(D,X)|<|q^\top(\eta(D,X)-\eta_0(D,X))|})^{1/2} \\
&\leq (\mathbb{E}|q^\top(\eta(D,X)-\eta_0(D,X))|^2)^{1/2} \\
&\lesssim \|\eta(D,X)-\eta_0(D,X)\|_{L_{P,2}} = o(N^{-1/4}).
\end{aligned}$$

This step completes the proof of Theorem 4. □