

# Machine learning for set-identified linear models

Vira Semenova

MIT

## Abstract

This paper provides estimation and inference methods for an identified set where the selection among a very large number of covariates is based on modern machine learning tools. I characterize the boundary of the identified set (i.e., support function) using a semiparametric moment condition. Combining Neyman-orthogonality and sample splitting ideas, I construct a root-N consistent, uniformly asymptotically Gaussian estimator of the support function and propose a weighted bootstrap procedure to conduct inference about the identified set. I provide a general method to construct a Neyman-orthogonal moment condition for the support function. Applying my method to Lee (2008)'s endogenous selection model, I provide the asymptotic theory for the sharp (i.e., the tightest possible) bounds on the Average Treatment Effect in the presence of high-dimensional covariates. Furthermore, I relax conventional monotonicity assumption and allow the sign of the treatment effect on the selection (e.g., employment) to be determined by covariates. Using JobCorps data set with very rich baseline characteristics, I substantially tighten the bounds on the JobCorps effect on wages under weakened monotonicity assumption.

# 1 Introduction

When dealing with endogenous selection, economists often use Lee (2008) approach to bound average treatment effect from above and below because this parameter is not point-identified. In population, the sharp (i.e. the tightest possible) version of Lee (2008) bounds is defined conditionally on all available pre-treatment covariates. However, modern data sets have so many covariates that using all of them in estimation may be computationally difficult or can invalidate inference. Therefore, economists choose to estimate non-sharp population bounds and end up with the estimates that may not be sufficiently informative. For example, the upper and the lower bound often have opposite signs and cannot determine whether treatment helps or hurts the outcome. This paper gives a method to estimate and conduct inference in set-identified models in the presence of high-dimensional covariates, including the model of Lee (2008) as a special case.

The main contribution of this paper is to provide construct a root- $N$  consistent, uniformly asymptotically Gaussian estimator of the identified set's boundary where the selection among high-dimensional covariates is based on modern machine learning tools. The paper focuses on identified sets whose boundaries can be characterized by a semiparametric moment equation. In this equation, the parametric component gives the description of the boundary (i.e., support function) and the nonparametric component is a nuisance parameter, for example, a conditional mean function. A naive approach would be to plug-in a machine learning estimate of the nuisance parameter into the moment equation and solve the moment equation for the boundary. However, modern regularized methods (machine learning techniques) have bias converging slower than the parametric rate. As a result, plugging such estimates into the moment equation produces a biased, low-quality estimate of the identified set's boundary.

The major challenge of this paper is to overcome the transmission of the biased estimation of the first-stage nuisance parameter into the second stage. A basic idea, previously proposed in the point-identified case, is to make the moment equation insensitive, or, formally, Neyman-orthogonal, to the biased estimation of the first-stage parameter (Neyman (1959)). Combining Neyman-orthogonality and sample splitting, (Chernozhukov et al. (2017a), Chernozhukov et al. (2017b)) derive a root- $N$  consistent and asymptotically normal estimator of the low-dimensional parameter identified by a semiparametric moment equation. However, extending this idea to a

set-identified case presents additional challenges.

The main distinction between the point- and set-identified cases is that the target parameter is no longer a finite-dimensional vector, but a boundary that consists of continuum points. Therefore, in addition to pointwise inference, economists are interested in uniform statistical properties over the identified set's boundary. Establishing uniform inference is not trivial. To control the speed at which an empirical sample average concentrates around the population mean, I invoke empirical process theory instead of Markov inequality that is typically sufficient in the point-identified case. Second, because the moment condition for the boundary depends on the nuisance parameter in a discontinuous way, establishing Neyman-orthogonality is a non-trivial exercise. I develop high-level sufficient conditions for Neyman-orthogonality and derive a uniformly root- $N$  consistent, uniformly asymptotically Gaussian estimator of the identified set's boundary.

To make the orthogonal approach useful, I provide a general recipe to construct a Neyman-orthogonal moment equation starting from a non-orthogonal one, extending the previous work on orthogonal estimation (Hardle and Stoker (1989), Newey and Stoker (1993), Newey (1994), Ichimura and Newey (2017), Chernozhukov et al. (2017b)) from a point- to a set-identified case. I also provide a weighted bootstrap algorithm to conduct inference about the identified set's boundary. The procedure simplifies the weighted bootstrap algorithm from Chandrasekhar et al. (2011): instead of re-estimating the first-stage parameter in each bootstrap repetition, I estimate the first-stage parameter once on an auxiliary sample. My algorithm is faster to compute because only the second stage is repeated in the simulation. I show that the simpler weighted bootstrap procedure is valid when the moment equation is Neyman-orthogonal.

My main contribution to the applied work is the estimator of sharp Lee (2008) bounds in the presence of high-dimensional covariates. Reporting nonparametric bounds on the average treatment effect in addition to the point estimates derived under stronger identification assumptions is a common robustness check in labor and education studies (Angrist et al. (2006), Lee (2008), Engberg et al. (2014), Huber et al. (2017), Abdulkadiroglu et al. (2018), Sieg and Wang (2018))<sup>1</sup>. In contrast to Lee (2008) approach adopted in these papers, I represent each bound as a solution

---

<sup>1</sup>For example, Engberg et al. (2014) reports the effect of attending a magnet program on the Mathematics test score lies between  $-24.22(148.06)$  and  $87.09(57.62)$ . The results are taken from Table 8 of Engberg et al. (2014), which reports the ATE of attending a magnet program in a mid-sized urban school district on the high school achievement in Mathematics, as measured by a standardized achievement test score. The standard errors are indicated in parentheses.

to a semiparametric moment equation that depends on conditional expectation function of selection (e.g., employment) and conditional quantile function of the outcome (e.g., wage). I derive the Neyman-orthogonal moment equations for each bound and give the low-level conditions for the asymptotic theory. Furthermore, I relax the conventional monotonicity assumption adopted in these papers and allow the sign of the treatment effect on the selection to be determined by the covariates.

My two other applications consider the case when an outcome variable is recorded in intervals. In the second application, I study the partially linear model from Robinson (1988) in the presence of high-dimensional covariates and interval-valued outcome. I characterize the identified set for the causal parameter in this model and provide estimation and inference methods for the identified set's boundary. I provide primitive conditions on the problem design that allow to incorporate machine learning tools to conduct uniform inference about the boundary. Because Robinson (1988)'s model may be misspecified in practice, I introduce a new parameter, called a partially linear predictor, to measure the predictive effect of a variable of interest on an outcome variable in the presence of high-dimensional controls. I show that the identified set for the causal parameter in Robinson (1988) is the sharp identified set for the partially linear predictor. In the third application, I study the average partial derivative (Härdle and Stoker (1989), Newey and Stoker (1993)) and provide primitive sufficient conditions that allow to incorporate machine learning tools to conduct uniform inference about the boundary.

As an empirical application, I bound the wage effects of JobCorps training program that has been originally studied by Lee (2008). The data consists of 9145 subjects that have been randomly split into treatment (enroll as usual) groups and control (embargo from the program) groups and had their wages recorded on a weekly basis during the following four years. The pre-treatment information includes employment, health, criminal, welfare receipt history during one year before random assignment and is represented by more than 5000 covariates. I find that standard monotonicity assumption (Lee (2008)) almost never holds. Indeed, enrollment in JobCorps helps noone's employment in the first week since random assignment, helps employment for 50% of the population by the end of the second year, and helps almost everyone's employment by the end of fourth year. I also find that JobCorps has small positive effects on the wage earned at week 90 through 208 because the all-covariate estimate of lower bound is positive almost always. In con-

trast, standard Lee (2008) no-covariate lower bound is always negative and cannot determine the direction of the wage effect<sup>2</sup>.

The paper is organized as follows. Section 2 provides motivating examples and constructs an estimator for the support function in a one-dimensional case of the partially linear predictor. Section 3 introduces a general set-identified linear model with high-dimensional covariates and establishes theoretical properties of the support function estimator. Section 4 describes the applications of the proposed framework to bounds analysis and to models where an outcome variable is recorded in intervals. Section 5 revisits the effectiveness of JobCorps training program. Section 6 states my conclusions.

## 1.1 Literature Review

This paper is related to two lines of research: estimation and inference in set-identified models and Neyman-orthogonal semiparametric estimation. This paper contributes to the literature by introducing Neyman-orthogonal semiparametric estimation to the set-identified literature.

Set-identification is a vast area of research (Manski (1989), Manski and Tamer (2002), Beresteanu and Molinari (2008), Bontemps et al. (2012), Beresteanu et al. (2011), Ciliberto and Tamer (2009), Chen et al. (2011), Kaido and White (2014), Kaido and Santos (2014), Chandrasekhar et al. (2011), Kaido (2016), Kaido (2017)), see e.g. Tamer (2010) or Molinari and Molchanov (2018) for a review. There are two approaches to estimate and conduct inference on identified sets: the moment inequalities approach (Chernozhukov et al. (2007), Kaido and White (2014)) and the support function approach (Beresteanu and Molinari (2008), Bontemps et al. (2012)), which applies only to convex and compact identified sets. A framework to unify these approaches was proposed by Kaido (2016). In this paper, I extend the support function approach, allowing the moment equation for the identified set's boundary to depend on a nuisance parameter that can be high-dimensional and is estimated by machine learning methods. In Semenova (2018), I introduce the same dependence in moment inequalities.

Within the first line of research, my empirical applications are most connected to work that derives nonparametric bounds on the average treatment effect in the presence of endogenous sample

---

<sup>2</sup>The data, replication code, and the package to estimate bounds are available at <https://github.com/vsemenova/leebounds>.

selection and non-compliance. This literature (Angrist et al. (2002), Angrist et al. (2006), Engberg et al. (2014), Huber et al. (2017), Abdulkadiroglu et al. (2018), Sieg and Wang (2018)) derives non-parametric bounds on the average treatment effect. Specifically, I build on Lee (2008), who derived sharp bounds on the average treatment effect and highlighted the role of covariates in achieving sharpness. However, Lee (2008)’s estimator only applies to a small number of discrete covariates. In this paper, I permit a large number of both discrete and continuous covariates and leverage the predictive power of machine learning tools to estimate and perform inference on sharp bounds.

The second line of research obtains a  $\sqrt{N}$ -consistent and asymptotically normal estimator of a low-dimensional target parameter  $\theta$  in the presence of a high-dimensional nuisance parameter  $\eta$  (Neyman (1959), Neyman (1979), Hardle and Stoker (1989), Newey and Stoker (1993), Newey (1994), Robins and Rotnitzky (1995), van der Vaart (1998), Robinson (1988), Chernozhukov et al. (2017a), Chernozhukov et al. (2017b)). It is common to estimate the target parameter in two stages, where a first-stage estimator of the nuisance  $\hat{\eta}$  is plugged into a sample analog of a mathematical relation that identifies the target, such as a moment condition, a likelihood function, etc. A statistical procedure is called Neyman-orthogonal (Neyman (1959), Neyman (1979)) if it is locally insensitive with respect to the estimation error of the first-stage nuisance parameter. In a point-identified problem, the orthogonality condition is defined at the true value of the target  $\theta_0$ . Since the notion of unique true value  $\theta_0$  no longer exists in a set-identified framework, I extend the orthogonality condition to hold on a slight expansion of the boundary of the identified set.

## 2 Setup and Motivation

### 2.1 General Framework

I focus on identified sets that can be represented as weighted averages of an outcome variable that is known to lie within an interval. Let  $Y$  be an outcome and  $Y_L, Y_U$  be random variables such that

$$Y_L \leq Y \leq Y_U \text{ a.s.} \tag{2.1}$$

Consider an identified set of the following form

$$\mathcal{B} = \{\beta = \Sigma^{-1}\mathbb{E}VY, \quad Y_L \leq Y \leq Y_U\}, \quad (2.2)$$

where  $V \in \mathcal{R}^d$  is a  $d$ -vector of weights and  $\Sigma \in \mathcal{R}^{d \times d}$  is a full-rank normalizing matrix.  $\Sigma$  can be either known or unknown, covering a variety of cases. For example,  $\Sigma = V = 1$  corresponds to the expectation of an outcome  $Y$ . For another example,  $\Sigma = (\mathbb{E}VV^\top)^{-1}$  corresponds to the set-valued best linear predictor of the outcome  $Y$  when  $V$  is used as a predictive covariate. I have adopted this structure because it allows me to cover a wide class of set-identified models that are usually studied separately.

A key innovation of my framework is that the bounds  $Y_L, Y_U$  and the weighting variable  $V$  can depend on an identified nuisance parameter that I allow to be high-dimensional. To fix ideas, let  $W$  be a vector of observed data and  $P_W$  denote its distribution. Then, I allow each coordinate of the weighting vector  $V$  and the bounds  $Y_L, Y_U$  to depend on an identified parameter of the data distribution  $P_W$ . The examples below demonstrate the importance of this innovation.

## 2.2 Motivating Examples

**Example 1. Endogenous Sample Selection.** In this example I revisit the model of endogenous sample selection from Lee (2008). I use the following notation for the potential outcomes. Let  $D \in \{1, 0\}$  denote an indicator for whether an unemployed subject has won a lottery to participate in a job training program. Let  $S_0 = 1$  be a dummy for whether the subject would have been employed after losing the lottery, and  $S_1 = 1$  be a dummy for whether the subject would have been employed after winning the lottery. Similarly, let  $\{Y_d, d \in \{1, 0\}\}$  represent the potential wages in case of winning and losing the lottery, respectively. The object of interest is the average effect on wages  $\beta = \mathbb{E}[Y_1 - Y_0 | S_1 = 1, S_0 = 1]$  for the group of people who would have been employed regardless of lottery's outcome, or, briefly, the always-employed. The data consist of the admission outcome  $D$ , the observed employment status  $S = DS_1 + (1 - D)S_0$ , wages for employed subjects  $S \cdot Y = S \cdot (DY_1 + (1 - D)Y_0)$ , and the baseline covariates  $X$  (e.g., age, gender, race). As discussed in Lee (2008), the average difference in wages between the treatment and control group contains selection bias and cannot be used to estimate  $\beta$ .

I start with brief description of Lee (2008) model and its assumptions.

**ASSUMPTION 1** (Unconditional monotonicity of Lee (2008)). *The following assumptions hold.*

(1) *The program admission  $D$  must be independent of the potential employment and wage outcomes, as well as the baseline covariates:  $D \perp (S_1, S_0, Y_1, Y_0, X)$ .* (2) *Either  $S_1 \geq S_0$  with probability 1 or  $S_0 \geq S_1$  with probability 1.*

I derive the sharp upper bound  $\beta_U$  on the treatment effect under Assumption 1 assuming  $S_1 \geq S_0$  a.s. Define the following functions:

$$s(D, X) = \mathbb{E}[S = 1 | D, X], p_0(X) = \frac{s(0, X)}{s(1, X)}, \quad (2.3)$$

$$y_{\{u, X\}} : \Pr(Y \leq y_{\{u, X\}} | X, D = 1, S = 1) = u, \quad u \in [0, 1].$$

Specifically,  $s(1, X), s(0, X)$  are the conditional probabilities of employment in the treatment and control group,  $p_0(X) = \frac{s(0, X)}{s(1, X)}$ , and  $y_{\{u, X\}}$  is the quantile function of wage in the treated group  $\{D = 1, S = 1\}$ . Since  $S_0 = 1$  implies  $S_1 = 1$ , the group  $\{D = 0, S = 1\}$  is a random sample from the always-employed. In contrast, the group  $\{D = 1, S = 1\}$  consists of the always-employed and the compliers, and the fraction of the always-employed conditional on  $X$  is equal to  $p_0(X) = \frac{s(0, X)}{s(1, X)}$ . Therefore, the sharp upper bound on ATE  $\Delta_{X,0}^{\text{UB}}$  is attained when the always-employed comprise the top  $p_0(X)$ -quantile of wages in  $\{D = 1, S = 1\}$  group and is equal to

$$\Delta_{X,0}^{\text{UB}} = \mathbb{E}[Y | D = 1, S = 1, Y \geq y_{\{1-p_0(X), X\}}] - \mathbb{E}[Y | D = 0, S = 1, X]. \quad (2.4)$$

Integrating the conditional bound  $\Delta_{X,0}^{\text{UB}}$  against the density of always-employed, we obtain the unconditional bound

$$\beta_U = \int_{x \in \mathcal{X}} \Delta_{X,0}^{\text{UB}} p(x | D = 0, S = 1) dx. \quad (2.5)$$

Lee (2008)'s estimator of  $\beta_U$  is the empirical analog of (2.5) and is restricted only to the settings with a small number of discrete covariates.

To make progress, I characterize the boundary of identified set

$$\mathcal{B} = [\beta_L, \beta_U] \quad (2.6)$$

using semiparametric moment equations. Invoking Bayes theorem (see Supplementary Appendix, Lemma 14), I obtain  $\beta_U = \mathbb{E}Y_U(X) - \mathbb{E}[Y|D = 1, S = 0]$ , where  $Y_U$  is

$$Y_U = Y_U(X) := \frac{D \cdot S \cdot Y \cdot 1_{\{Y \geq y_{\{1-p_0(X), X\}}\}} \Pr(D = 0)}{\Pr(D = 0, S = 1) \Pr(D = 1)}. \quad (2.7)$$

Likewise,  $\beta_L = \mathbb{E}Y_L(X) - \mathbb{E}[Y|D = 1, S = 0]$ , where the lower bound  $Y_L(X)$  is equal to

$$Y_L = Y_L(X) := \frac{D \cdot S \cdot Y \cdot 1_{\{Y \leq y_{\{p_0(X), X\}}\}} \Pr(D = 0)}{\Pr(D = 0, S = 1) \Pr(D = 1)}. \quad (2.8)$$

Having derived the orthogonal moments for each of these equations (see Section 4.1), I estimate and conduct inference on the sharp bounds  $[\beta_L, \beta_U]$  while allowing the first-stage nuisance parameters  $s(1, X), s(0, X), y_{\{p_0(X), X\}}, y_{\{1-p_0(X), X\}}$  to be estimated by machine learning tools. Since the proposed estimator of the bounds is asymptotically equivalent to sample average, I can also conduct inference about the true treatment effect based on Imbens and Manski (2004) approach.

Finally, I relax the original monotonicity Assumption 1 by allowing the sign of the treatment effect on employment to be determined by the covariates.

**ASSUMPTION 2** (Conditional monotonicity). *The following assumptions hold. (1) The program admission  $D$  must be independent of the potential employment and wage outcomes, as well as the baseline covariates:  $D \perp (S_1, S_0, Y_1, Y_0, X)$ . (2) There exists a partition of covariate space  $\mathcal{X} = \mathcal{X}_0 \sqcup \mathcal{X}_1$  so that  $S_1 \geq S_0$  a.s. on  $\mathcal{X}_0$  and  $S_0 \geq S_1$  a.s. on  $\mathcal{X}_1$ . (3) The treatment effect on the response is sufficiently strong:  $\inf_{x \in \mathcal{X}} |s(1, x) - s(0, x)| \geq \varepsilon > 0$  for some  $\varepsilon > 0$  that does not depend on sample size.*

Under Assumption 2, one can identify the set  $\mathcal{X}_0 = \{p_0(X) < 1\}$  and  $\mathcal{X}_1 = \{p_0(X) > 1\}$ . The sharp upper bound  $\bar{\Delta}_X^{\text{UB}}$  conditional on  $X$  takes the form

$$\bar{\Delta}_X^{\text{UB}} = 1_{\{p_0(X) < 1\}} \Delta_{X,0}^{\text{UB}} + 1_{\{p_0(X) > 1\}} \Delta_{X,1}^{\text{UB}}, \quad (2.9)$$

where  $\Delta_{X,0}^{\text{UB}}$  is defined in (2.4) and  $\Delta_{X,1}^{\text{UB}} = \mathbb{E}[Y|D = 1, S = 1, X] - \mathbb{E}[Y|D = 0, S = 1, Y \leq y_{\{1/p_0(X), X\}}]$  is the analog of  $\Delta_{X,0}^{\text{UB}}$  for the case  $p_0(X) > 1$ . Integrating the conditional bound  $\bar{\Delta}_X^{\text{UB}}$  against the density of always-employed, we obtain the unconditional bound.

**Example 2. Average Partial Derivative.** An important parameter in economics is the average partial derivative. This parameter shows the average effect of a small change in a variable of interest  $D$  on the outcome  $Y$  conditional on the covariates  $X$ . To describe this change, define the conditional expectation function of an outcome  $Y$  given the variable  $D$  and exogenous variable  $X$  as  $\mu(D, X) := \mathbb{E}[Y|D, X]$  and its partial derivative with respect to  $D$  as  $\partial_D \mu(D, X) := \partial_d \mu(d, X)|_{d=D}$ . Then, the average partial derivative is defined as  $\beta = \mathbb{E} \partial_D \mu(D, X)$ . For example, when  $Y$  is the logarithm of consumption,  $D$  is the logarithm of price, and  $X$  is the vector of other demand attributes, the average partial derivative stands for the average price elasticity.

Assume that the variable  $D$  has bounded support  $\mathcal{D} \subset \mathcal{R}^d$  and has positive density on this support. Hardle and Stoker (1989) have shown that the average partial derivative can be represented as  $\beta = \mathbb{E} V Y$ , where  $V = -\partial_D \log f(D|X) = -\frac{\partial_D f(D|X)}{f(D|X)}$  is the negative partial derivative of the logarithm of the density  $f(D|X)$ .

Suppose the outcome  $Y$  is interval-censored. As discussed in Kaido (2017), the sharp identified set  $\mathcal{B}$  for the average partial derivative can be represented as

$$\mathcal{B} = \{\mathbb{E} V Y, \quad Y_L \leq Y \leq Y_U\}, \quad (2.10)$$

which is a special case of model (2.1)-(2.2) with  $\Sigma = 1$ ,  $V = -\frac{\partial_D f(D|X)}{f(D|X)}$ , and the nuisance parameter  $\eta_0(X) = \frac{\partial_D f(D|X)}{f(D|X)}$ . In contrast to Kaido (2017), I allow the vector of covariates  $X$  to be high-dimensional.

**Example 3. Partially Linear Predictor.** A widely used approach to measure the causal effect of a variable of interest  $D$  on an outcome variable  $Y$  is to adopt the partially linear model (Robinson (1988)):

$$Y = D\beta_0 + f_0(X) + U, \quad \mathbb{E}[U|D, X] = 0,$$

where  $X$  is a vector of covariates. However, when the model is misspecified, the parameter  $\beta_0$  has no interpretation. An alternative parameter, which is robust to misspecification of the partially

linear model, is a partially linear predictor. This parameter is defined as

$$\beta = \arg \min_{b \in \mathbb{R}^d, f \in \mathcal{M}} \mathbb{E}(Y - D^\top b - f(X))^2,$$

or the linear component of the projection of  $Y$  on a partially linear combination of the variable  $D$  and the covariates  $X$ , where  $\mathcal{M}$  is a set of functions of  $X$ . Equivalently, the parameter  $\beta$  can be represented as the best linear predictor of variable  $Y$  in terms of the first-stage residual  $V$  (see Lemma 15) as

$$\beta = \arg \min_{b \in \mathbb{R}^d} \mathbb{E}(Y - V^\top b)^2,$$

where the first-stage residual  $V$  is defined as  $V = D - \mathbb{E}[D|X]$ .

Suppose the outcome  $Y$  is interval-censored. Then, the sharp identified set  $\mathcal{B}$  for the partially linear predictor is

$$\mathcal{B} = \{(\mathbb{E}VV^\top)^{-1}\mathbb{E}VY, \quad Y_L \leq Y \leq Y_U\}, \quad (2.11)$$

which is a special case of model (2.1)-(2.2) with  $V = D - \mathbb{E}[D|X]$ ,  $\Sigma = \mathbb{E}VV^\top$ , and the nuisance parameter  $\eta_0(X) = \mathbb{E}[D|X]$ . Moreover, the identified set  $\mathcal{B}$  is a non-sharp identified set for the causal parameter  $\beta_0$  when the partially linear model is correctly specified.

### 2.3 Main ideas

This section is a demonstration of two well-known ideas, Neyman-orthogonality (Neyman (1959), Chernozhukov et al. (2017a), Chernozhukov et al. (2017b)) and sample splitting, for the partially linear predictor (Example 3) when the variable  $D$  is one-dimensional. Combining these two ideas, I obtain a root- $N$  consistent, asymptotically Gaussian estimator of the lower and the upper bound on the partially linear predictor.

### 2.3.1 Neyman-orthogonality

When the variable  $D$  is one-dimensional, the identified set (2.11) is a closed interval  $[\beta_L, \beta_U]$ . The moment function for the upper bound  $\beta_U$  takes the form

$$m(\text{data}, \beta_U, \eta_0) := (Y^{\text{UBG}}(\eta_0) - (D - \eta_0(X))\beta_U)(D - \eta_0(X)), \quad (2.12)$$

where the *upper bound generator*  $Y^{\text{UBG}}(\eta)$  is defined as  $Y^{\text{UBG}}(\eta) = \begin{cases} Y_L, & D - \eta(X) \leq 0 \\ Y_U, & D - \eta(X) \geq 0. \end{cases}$ . A

naive estimator  $\hat{\beta}_U^{\text{NAIVE}}$ , based on plugging in a regularized estimator  $\hat{\eta}$  into the moment condition (8.2), converges at a rate slower than root- $N$  and cannot be used to conduct inference about  $\beta_U$  using standard asymptotic theory. Because  $\hat{\eta}$  is a regularized estimator, its own bias converges at a rate slower than root- $N$  and carries over into the second stage. The second-stage bias is non-zero  $\partial_{\eta_0} \mathbb{E}m(W, \beta_U, \eta_0)(\hat{\eta} - \eta_0) = -\mathbb{E}[Y^{\text{UBG}}(\eta_0)(\hat{\eta}(X) - \eta_0(X))] \neq 0$ .

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. Specifically, I replace the upper bound generator  $Y^{\text{UBG}}(\eta)$  by its residual after projection on the set of covariates  $Y^{\text{UBG}}(\eta) - \mathbb{E}[Y^{\text{UBG}}(\eta)|X]$ , which makes the derivative above equal to zero. This replacement corresponds to a new moment function

$$g(\text{data}, \beta_U, \xi_0) = (Y^{\text{UBG}}(\eta_0) - \mathbb{E}[Y^{\text{UBG}}(\eta_0)|X] - (D - \eta_0(X))\beta_U)(D - \eta_0(X)),$$

which is orthogonal with respect to its first-stage parameter  $\xi_0 = \{\eta_0, \mathbb{E}[Y^{\text{UBG}}(\eta_0)|X]\}$  (see Theorem 5). By orthogonality, the first-stage bias of  $\hat{\xi}$  does not carry over into the second-stage.

### 2.3.2 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. 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. 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.

### 3 Main Results

In this section, I introduce a general set-identified linear model with a high-dimensional nuisance parameter. I describe the boundary of the identified set (support function) by a semiparametric moment equation. I introduce a new sufficient condition for a moment equation - uniform near-orthogonality - and establish uniform asymptotic theory for the support function estimator and bootstrap support function process under that condition. Finally, I provide a general recipe to construct a uniformly near-orthogonal moment equation starting from a non-orthogonal one.

**Notation.** I use the following standard notation. Let  $\mathcal{S}^{d-1} = \{q \in \mathcal{R}^d, \|q\| = 1\}$  be the  $d$ -dimensional unit sphere and  $q \in \mathcal{S}^{d-1}$  be a generic vector on the unit sphere. I use the following notation

$$\Gamma(t_1, t_2 - t_1, t_3) := t_1 + (t_2 - t_1)1_{\{t_3 \geq 0\}}, \quad (3.1)$$

where  $1_{\{t_3 \geq 0\}} = 1$  if  $t_3$  is non-negative and  $1_{\{t_3 \geq 0\}} = 0$  otherwise. I use standard notation for numeric and stochastic dominance. For two numeric sequences  $\{a_n, n \geq 1\}$  and  $\{b_n, n \geq 1\}$ , let  $a_n \lesssim b_n$  stand for  $a_n = O(b_n)$ . For two sequences of random variables  $\{a_n, n \geq 1\}$  and  $\{b_n, n \geq 1\}$ , let  $a_n \lesssim_P b_n$  stand for  $a_n = O_P(b_n)$ . For a random variable  $\xi$ ,  $(\xi)^0 := \xi - \mathbb{E}[\xi]$ . Let  $L^\infty(\mathcal{S}^{d-1})$  be the space of absolutely surely bounded functions defined on the unit sphere  $\mathcal{S}^{d-1}$ . Define an  $L_{P,c}$  norm of a vector-valued random variable  $W$  as:  $\|W\|_{L_{P,c}} := (\int_{W \in \mathcal{W}} \|W\|^c)^{1/c}$ . Let  $\mathcal{W}$  be the support of the data vector  $W$  of the distribution  $P_W$ . Let  $(W_i)_{i=1}^N$  be an i.i.d sample from the distribution  $P_W$ . Denote the sample average of a function  $f(\cdot)$  as  $\mathbb{E}_N[f(W_i)] := \frac{1}{N} \sum_{i=1}^N f(W_i)$  and the centered, root- $N$  scaled sample average as  $\mathbb{G}_N[f(W_i)] := \frac{1}{\sqrt{N}} \sum_{i=1}^N [f(W_i) - \mathbb{E}f(W_i)]$ .

### 3.1 High-Level Assumptions

One of this paper's key innovations is to allow the identified set  $\mathcal{B}$ , given in (2.2), to depend on an identified parameter of data distribution  $P_W$ . Definition 1 formalizes this dependence.

**Definition 1** (Constructed Random Variable). *Let  $W$  be the vector of the observed data,  $P_W$  its distribution, and  $\mathcal{W}$  its support. I refer to  $V$  as a constructed random variable if there exists an identified parameter  $\eta_0, \eta_0 \in \mathcal{T}$  from a linear and convex set  $\mathcal{T}$  and a known measurable map  $H(W, \eta) : \mathcal{W} \times \mathcal{T} \rightarrow \mathcal{R}$  such that  $V = H(W, \eta_0)$  a.s. I refer to  $\eta$  as a nuisance parameter and  $\eta_0$  as the true value of  $\eta$ .*

**ASSUMPTION 3** (Constructed Random Variables). *Each coordinate of the random vector  $(V, Y_L, Y_U)$  in the identified set (2.2) is either an observed or constructed random variable.*

To complete the model, I need to identify matrix  $\Sigma$  in (2.2) when  $\Sigma$  is unknown. If this is the case, I assume that  $\Sigma$  is identified by a semiparametric moment condition (Assumption 4).

**ASSUMPTION 4** (Identification of  $\Sigma$ ). *1. There exists an identified parameter  $\eta$  of the distribution  $P_W$  and a known measurable map  $A(W, \eta) : \mathcal{W} \times \mathcal{T} \rightarrow \mathcal{R}$  such that*

$$\Sigma = \mathbb{E}A(W, \eta_0). \quad (3.2)$$

*2. There exist constants  $\lambda_{\min} > 0$  and  $\lambda_{\max} < \infty$  that bound the eigenvalues of  $\Sigma$  from above and below  $0 < \lambda_{\min} \leq \min \text{eig}(\Sigma) \leq \max \text{eig}(\Sigma) < \lambda_{\max}$ .*

According to Beresteanu and Molinari (2008) and Bontemps et al. (2012), the identified set  $\mathcal{B}$  is a compact and convex set. Therefore, it can be equivalently represented by its support function. Fix a direction  $q$  on a unit sphere  $\mathcal{S}^{d-1} := \{q \in \mathcal{R}^d, \|q\| = 1\}$ . Define the support function

$$\sigma(q, \mathcal{B}) := \sup_{b \in \mathcal{B}} q^\top b \quad (3.3)$$

as the (signed) distance from the origin to the hyperplane tangent to  $\mathcal{B}$  in the direction  $q$ . Using the arguments of (Beresteanu and Molinari (2008), Bontemps et al. (2012)), the function  $\sigma(q, \mathcal{B})$

is equal to the expectation of the product of two random variables  $z_q$  and  $Y_q$

$$\sigma(q, \mathcal{B}) = \mathbb{E}z_q Y_q, \quad (3.4)$$

where  $z_q = q^\top \Sigma^{-1} V$  is a normalized projection of the covariate  $V$  onto the direction  $q$  and the variable  $Y_q$  is defined as

$$Y_q = Y_L + (Y_U - Y_L) 1_{\{z_q > 0\}} := \Gamma(Y_L, Y_U - Y_L, z_q). \quad (3.5)$$

Namely,  $Y_q$  is equal to the lower bound  $Y_L$  when  $z_q$  is non-positive and equal to the upper bound  $Y_U$  otherwise. To highlight the dependence of  $z_q$  and  $Y_q$  on  $\eta$ , I will rewrite (3.4) as a semiparametric moment equation for  $\sigma(q, \mathcal{B})$

$$\mathbb{E}[\sigma(q, \mathcal{B}) - z_q(\eta_0) Y_q(\eta_0)] = 0. \quad (3.6)$$

### 3.2 Orthogonality and near-orthogonality

As discussed in the introduction, the moment equation (3.6) produces a low-quality estimator of the support function. The problem arises because the moment equation (3.6) is sensitive with respect to the biased estimation of  $\eta_0$ . To overcome this problem, I replace (3.6) with an orthogonal moment function  $g(W, p, \xi(p)) : \mathcal{W} \times \mathcal{P} \times \Xi \rightarrow \mathcal{R}$  which depends on a functional nuisance parameter  $\xi = \xi(p)$ , defined on a compact subset  $\mathcal{P} \subset \mathcal{R}^d$ <sup>3</sup>. The orthogonal moment for support function takes place

$$\mathbb{E}[\sigma(q, \mathcal{B}) - g(W, p_0(q), \xi_0(p_0(q)))] = 0, \quad (3.7)$$

where  $p_0(q) = (\Sigma^{-1})^\top q$ .

Let  $\Xi$  be a convex subset of a normed vector space that contains the functional parameter  $\xi_0 = \xi_0(p)$ . Define the pathwise (Gateaux) derivative map on the set  $\Xi - \xi_0$  as  $D_r : \Xi - \xi_0 \rightarrow \mathcal{R}$

$$D_r[\xi - \xi_0] := \partial_r \mathbb{E}g(W, p, r(\xi - \xi_0) + \xi_0), \quad \xi \in \Xi, \quad p \in \mathcal{P}, \quad r \in [0, 1)$$

---

<sup>3</sup>For example, the set  $\mathcal{P}$  is taken to be  $\mathcal{P} = \{(\Sigma^{-1})^\top q, q \in S^{d-1}\}$  when  $\Sigma$  is known and  $\mathcal{P} = \{p \in \mathcal{R}^d, 0.5\lambda_{\min} \leq \|p\| \leq 2\lambda_{\max}\}$  when  $\Sigma$  is unknown.

which I assume exists. I will also use the notation  $\partial_0 \mathbb{E}g(W, p, \xi_0)[\xi - \xi_0] = D_0[\xi - \xi_0]$  for the Gateaux derivative at  $\xi_0$ . Let  $\{\Xi_N, N \geq 1\}$  be a sequence of subsets of  $\Xi$  (i.e.,  $\Xi_N \subseteq \Xi$ ) and  $\{\mathcal{T}_N, N \geq 1\}$  be a sequence of subsets of  $\mathcal{T}$  (i.e.,  $\mathcal{T}_N \subseteq \mathcal{T}$ ).

**Definition 2** (Neyman-orthogonality). *The moment function  $g(W, p, \xi)$  obeys the orthogonality condition at  $\xi_0$  with respect to the nuisance realization set  $\Xi_N \subset \Xi$  if the following conditions hold. (1) Equation (3.7) holds. (2) The pathwise derivative map  $D_r[\xi - \xi_0]$  exists for all  $r \in [0, 1)$  and  $\xi \in \Xi_N$  and vanishes at  $r = 0$  for each  $p \in \mathcal{P}$*

$$\partial_\xi \mathbb{E}g(W, p, \xi_0)[\xi - \xi_0] = 0 \quad \forall p \in \mathcal{P}. \quad (3.8)$$

Definition 2 requires the expectation of the moment function  $g(W, p, \xi_0)$  to have zero Gateaux derivative with respect to  $\xi$  at  $\xi_0$  at each vector  $p$  in the projection set  $\mathcal{P}$ . To accommodate the moment function (3.4) that depends on  $\eta$  in a non-smooth way, I relax the requirement of Definition 2 using the notion of uniform near-orthogonality.

**Definition 3** (Uniform near-orthogonality). *The moment function  $g(W, p, \xi)$  obeys the near-orthogonality condition at  $\xi_0$  with respect to the nuisance realization set  $\Xi_N \subset \Xi$  uniformly over  $\mathcal{P}$  if the following conditions hold. (1) Equation (3.7) holds. (2) The pathwise derivative map  $D_r[\xi - \xi_0]$  exists for all  $r \in [0, 1)$  and  $\xi \in \Xi_N$ . (3) There exists a sequence of positive constants  $\mu_N = o(N^{-1/2})$  such that the pathwise derivative  $D_r[\xi - \xi_0]$  at  $r = 0$  is uniformly small over the set  $\mathcal{P}$*

$$\sup_{p \in \mathcal{P}} |\partial_\xi \mathbb{E}g(W, p, \xi_0)[\xi - \xi_0]| \leq \mu_N.$$

**ASSUMPTION 5** (Near-orthogonality). *(1) There exists a measurable moment function  $g(W, p, \xi) : \mathcal{W} \times \mathcal{P} \times \Xi \rightarrow \mathcal{R}$  that obeys (3.7) and the near orthogonality condition uniformly over  $\mathcal{P}$ . (2) When  $\Sigma$  is unknown, there exists a moment matrix-valued function  $A(W, \eta)$  that obeys (3.2) and the orthogonality condition  $\partial_\eta \mathbb{E}A(W, \eta_0)[\eta - \eta_0] = 0$ .*

Assumption 5 is the key assumption of my paper. Assumption 5 (1) states that there exists a moment function  $g(W, p, \xi)$  for the support function  $\sigma(q, \mathcal{B})$  that is approximately insensitive with respect to the biased estimation of the nuisance parameter  $\xi$  at  $\xi_0$ . I show how to achieve this condition in Section 9. The second assumption states that the moment function  $A(W, \eta)$  is

insensitive with respect to the biased estimation of the parameter  $\eta$  at  $\eta_0$ . This assumption holds in practical applications (e.g., in Example 3). To sum up, the uniform near-orthogonal moment equation for  $\sigma(q, \mathcal{B})$  is

$$\mathbb{E}\psi(W, \theta(q), \xi_0(p)) := \mathbb{E} \begin{bmatrix} \sigma(q, \mathcal{B}) - g(W, p(q), \xi_0(p(q))) \\ A(W, \eta_0)p(q) - q \end{bmatrix} = 0. \quad (3.9)$$

**Definition 4** (Support Function Estimator when  $\Sigma$  is Known). *Let  $(W_i)_{i=1}^N$  be an i.i.d sample of the distribution  $P_W$  and  $\xi_0(p)$  be an identified parameter of  $P_W$ . Let  $\hat{\xi}(p)$  be the estimate of  $\xi_0(p)$ . Define an estimate of the support function  $\hat{\sigma}(q, \mathcal{B})$  as follows*

$$\hat{\sigma}(q, \mathcal{B}) = \frac{1}{N} \sum_{i=1}^N g(W_i, p_0(q), \hat{\xi}_i(p)), \quad (3.10)$$

where  $p_0(q) = (\Sigma^{-1})^\top q$ .

**Definition 5** (Support Function Estimator when  $\Sigma$  is Unknown). *Let  $(W_i)_{i=1}^N$  be an i.i.d. sample from a distribution  $P_W$ . Let  $\hat{\xi}(p), p \in \mathcal{P}$  be the estimate of  $\xi_0(p)$ . Define an estimate of the support function  $\hat{\sigma}(q, \mathcal{B})$  as follows*

$$\hat{\sigma}(q, \mathcal{B}) = \frac{1}{N} \sum_{i=1}^N g(W_i, \hat{p}(q), \hat{\xi}_i(\hat{p}(q))), \quad (3.11)$$

where  $\hat{p}(q) = (\hat{\Sigma}^{-1})^\top q$  and  $\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^N A(W_i, \hat{\eta}_i)$ .

### 3.3 Additional regularity conditions

Assumption 6 formalizes the speed of convergence of the estimated nuisance parameter  $\hat{\xi}(p)$ . It introduces the sequence of neighborhoods  $\{\Xi_N, N \geq 1\}$  around the true value  $\xi_0(p)$  that contain the estimate  $\hat{\xi}(p)$  with probability approaching one. As the sample size  $N$  increases, the neighborhoods shrink. The following rates  $r_N, r'_N, r''_N, \delta_N$  control the speed at which these neighborhoods shrink around  $\xi_0(p)$ .

**ASSUMPTION 6** (Quality of the First-Stage Estimation and Regularity of the Moment Function). *There exists a sequence  $\{\Xi_N, N \geq 1\}$  of subsets of  $\Xi$  (i.e.,  $\Xi_N \subseteq \Xi$ ) such that the following conditions*

hold. (1) The true value  $\xi_0$  belongs to  $\Xi_N$  for all  $N \geq 1$ . There exists a sequence of numbers  $\phi_N = o(1)$  such that the first-stage estimator  $\hat{\xi}(p)$  of  $\xi_0(p)$  belongs to  $\Xi_N$  with probability at least  $1 - \phi_N$ . There exist sequences  $r_N, r'_N, r''_N, \delta_N$  :  $r''_N \log^{1/2}(1/r''_N) = o(1)$ ,  $r'_N \log^{1/2}(1/r'_N) = o(1)$ ,  $r_N = o(N^{-1/2})$ , and  $\delta_N = o(N^{-1/2})$  such that the following bounds hold

$$\begin{aligned} & \sup_{\xi \in \Xi_N} \sup_{p \in \mathcal{P}} (\mathbb{E}(g(W, p, \xi(p)) - g(W, p, \xi_0(p)))^2)^{1/2} \lesssim r''_N, \\ & \sup_{\xi \in \Xi_N} \sup_{q \in \mathcal{S}^{d-1}} \sup_{p \in \mathcal{P}: \|p - p_0(q)\| \lesssim RN^{-1/2}} (\mathbb{E}(g(W, p, \xi_0(p)) - g(W, p_0, \xi_0(p_0)))^2)^{1/2} \lesssim r'_N, \\ & \sup_{r \in [0,1]} \sup_{p \in \mathcal{P}} (\partial_r^2 \mathbb{E}g(W, p, r(\xi(p) - \xi_0(p)) + \xi_0(p))) \leq r_N, \\ & \sup_{r \in [0,1]} \|\partial_r^2 \mathbb{E}A(W, r(\eta - \eta_0) + \eta_0)\| \leq r_N, \\ & \sup_{\eta \in \mathcal{T}_N} (\mathbb{E}\|A(W, \eta) - A(W, \eta_0)\|^2)^{1/2} \lesssim \delta_N. \end{aligned}$$

(2) The following conditions hold for the function class  $\mathcal{F}_\xi = \{g(W, p, \xi(p)), p \in \mathcal{P}\}$ . There exists a measurable envelope function  $F_\xi = F_\xi(W)$  that absolutely surely bounds all elements in the class  $\sup_{p \in \mathcal{P}} |g(W, p, \xi(p))| \leq F_\xi(W)$  a.s.. There exists  $c > 2$  such that  $\|F_\xi\|_{L_{p,c}} := (\int_{w \in \mathcal{W}} (F_\xi(w))^c)^{1/c} < \infty$ . There exist constants  $a, v$  that do not depend on  $N$  such that the uniform covering entropy of the function class  $\mathcal{F}_\xi$  is bounded  $\log \sup_Q N(\varepsilon \|F_\xi\|_{Q,2}, \mathcal{F}_\xi, \|\cdot\|_{Q,2}) \leq v \log(a/\varepsilon)$ , for all  $0 < \varepsilon \leq 1$ .

Assumption 7 is a standard requirement for a differentiable support function  $\sigma(q, \mathcal{B})$  (see, e.g. Chandrasekhar et al. (2011)). The differentiability of the support function ensures that the identified set  $\mathcal{B}$  is strongly convex. This property rules out the presence of the exposed faces of the identified set  $\mathcal{B}$  where the bias accumulates non-trivially. When the random variable  $V$  in (2.2) is sufficiently continuous, Assumption 7 holds. When the distribution of  $V$  is discrete, adding a small amount of continuously distributed noise suffices to achieve this requirement.

**ASSUMPTION 7** (Differentiable Support Function). *Let  $\bar{P} \subset \mathcal{R}^d$  be an open set that contains  $\mathcal{P} : \mathcal{P} \subset \bar{P}$ . Assume that the function  $p \rightarrow \mathbb{E}[g(W, p, \xi_0(p))]$ ,  $p \in \bar{P}$  is differentiable on  $\bar{P}$ . Define its gradient*

$$G(p) := \nabla_p \mathbb{E}g(W, p, \xi_0(p)). \quad (3.12)$$

In addition, if  $\Sigma$  is unknown, assume that the following bound holds uniformly on  $p_0 \in \mathcal{P}$

$$\mathbb{E}[g(W, p, \xi_0(p)) - g(W, p_0, \xi_0(p_0))] = G(p_0)(p - p_0) + o(\|p - p_0\|), \quad \|p - p_0\| \rightarrow 0. \quad (3.13)$$

### 3.4 Main Result

**Theorem 1** (Limit Theory for the Support Function Estimator). *Suppose Assumptions 3, 4, 5, 6, 7 hold. Let  $p_0(q) = (\Sigma^{-1})^\top q$ . Let  $\hat{\sigma}(q, \mathcal{B})$  be the Support Function Estimator provided in Definition 4 when  $\Sigma$  is known and Definition 5 when  $\Sigma$  is unknown. Define an influence function*

$$h(W, q) := g(W, p_0(q), \xi_0(p_0(q))) - \mathbb{E}[g(W, p_0(q), \xi_0(p_0(q)))]$$

when  $\Sigma$  is known and

$$h(W, q) := g(W, p_0(q), \xi_0(p_0(q))) - \mathbb{E}[g(W, p_0(q), \xi_0(p_0(q)))] - q^\top \Sigma^{-1} (A(W, \eta_0) - \Sigma) \Sigma^{-1} G(p_0(q))$$

when  $\Sigma$  is unknown. Then, the Support Function Estimator  $\hat{\sigma}(q, \mathcal{B})$  is uniformly asymptotically linear over the unit sphere  $\mathcal{S}^{d-1}$

$$\sqrt{N}(\hat{\sigma}(q, \mathcal{B}) - \sigma_0(q, \mathcal{B})) = \mathbb{G}_N[h(W_i, q)] + o_P(1). \quad (3.14)$$

Moreover, the empirical process  $\mathbb{G}_N[h(W_i, q)]$  converges to a tight Gaussian process  $\mathbb{G}[h(W_i, q)]$  in  $L^\infty(\mathcal{S}^{d-1})$  with the non-degenerate covariance function

$$\Omega(q_1, q_2) = \mathbb{E}[h(W, q_1)h(W, q_2)] - \mathbb{E}[h(W, q_1)]\mathbb{E}[h(W, q_2)], \quad q_1, q_2 \in \mathcal{S}^{d-1}.$$

Theorem 1 is my first main result. It says that the Support Function Estimator is asymptotically equivalent to a Gaussian process and can be used for pointwise and uniform inference about the support function. By orthogonality, the first-stage estimation error does not contribute to the total uncertainty of the two-stage procedure. In particular, orthogonality allows me to avoid relying on any particular choice of the first-stage estimator, and, in particular, employ modern regularized techniques to estimate the first-stage.

**Definition 6** (Weighted Bootstrap). *Let  $B$  represent a number of bootstrap repetitions. For each  $b \in \{1, 2, \dots, B\}$ , repeat*

1. *Draw  $N$  i.i.d. exponential random variables  $(e_i)_{i=1}^N : e_i \sim \text{Exp}(1)$ . Let  $\bar{e} = \mathbb{E}_N e_i$ .*
2. *When the matrix  $\Sigma$  is known, set  $\tilde{\Sigma} = \Sigma$ . Otherwise, estimate  $\tilde{\Sigma} = \mathbb{E}_N \frac{e_i}{\bar{e}} A(W_i, \hat{\eta}_i)$ .*
3. *Estimate  $\tilde{\sigma}^b(q, \mathcal{B}) = \mathbb{E}_N \frac{e_i}{\bar{e}} g(W_i, (\tilde{\Sigma}^{-1})^\top q, \hat{\xi}_i((\tilde{\Sigma}^{-1})^\top q))$ .*

**Theorem 2** (Limit Theory for the Bootstrap Support Function Process). *The bootstrap support function process  $\tilde{S}_N(q) := \sqrt{N}(\tilde{\sigma}(q, \mathcal{B}) - \hat{\sigma}(q, \mathcal{B}))$  admits the following approximation conditional on the data  $\tilde{S}_N(q) = \mathbb{G}_N[e_i^0 h_i^0(q)] + o_{P^e}(1)$  in  $L^\infty(\mathcal{S}^{d-1})$ . Moreover, the support function process admits an approximation conditional on the data*

$$\tilde{S}_N(q) = \tilde{\mathbb{G}}[h(q)] + o_{P^e}(1) \text{ in } L^\infty(\mathcal{S}^{d-1}),$$

where  $\tilde{\mathbb{G}}[h(q)]$  is a sequence of tight  $P$ -Brownian bridges in  $L^\infty(\mathcal{S}^{d-1})$  with the same distributions as the processes  $\mathbb{G}_N[h(q)]$ , and independent of  $\mathbb{G}_N[h(q)]$ .

Theorem 2 is my second main result. It says that weighted bootstrap of Theorem 1 can be used to approximate the support function process of Theorem 1 and conduct uniform inference about the support function. By orthogonality, I do not need to re-estimate the first-stage parameter in each bootstrap repetition. Instead, I estimate the first-stage parameter once on an auxiliary sample, and plug the estimate into the bootstrap sampling procedure.

Theorems 1 and 2 cover the result of Chandrasekhar et al. (2011) as a special case, extending it in several important directions. First, my framework allows for a richer set of parameters, described by Assumptions 3 and 4. Second, the asymptotic theory provided here does not rely on the choice of the series estimator in the first stage. Instead, it allows any machine learning method obeying Assumption 6 to estimate the first-stage parameters, which makes the theory accommodating many covariates.

## 4 Applications

In this section, I apply the asymptotic theory of Section 3 to three empirically relevant settings: Endogenous Sample Selection of Lee (2008), Average Partial Derivative of Kaido (2017), and Partially Linear Predictor. For each setting, I derive an orthogonal moment equation for the support function and provide primitive sufficient conditions for the theoretic results of Section 3 to hold.

### 4.1 Endogenous Sample Selection

Consider the setup of Example 1 under Assumption 1. The orthogonal moment equation for the upper bound  $\beta_U$  is

$$\mathbb{E}[\beta_U - \frac{D \cdot S \cdot Y 1_{\{Y \geq Q_{\{Y|D=1, S=1, X\}}(p_0(X), X)\}} \Pr(D=0)}{\Pr(D=0, S=1) \Pr(D=1)} + \sum_{j=1}^3 \alpha_j(W, \eta)] = 0, \quad (4.1)$$

where  $W = (D, S, SY, X)$  is the vector of data and  $\alpha_j(W, \eta)$ ,  $j \in \{1, 2, 3\}$  is the function that corrects bias of  $j$ 'th component of  $\eta$ . The true values of each bias correction term are

$$\begin{aligned} \alpha_1(W, \eta) &= y_{\{1-p_0(X), X\}} \frac{\Pr(D=0)}{\Pr(D=0, S=1)} \left( \frac{(1-D)S}{\Pr(D=0)} - \eta_1(X) \right), \\ \alpha_2(W, \eta) &= -y_{\{1-p_0(X), X\}} \frac{\Pr(D=0)s(0, X)}{\Pr(D=0, S=1)s(1, X)} \left( \frac{DS}{\Pr(D=1)} - \eta_2(X) \right), \\ \alpha_3(W, \eta) &= y_{\{1-p_0(X), X\}} \frac{\Pr(D=0)s(1, X)}{\Pr(D=0, S=1)} \left( 1_{\{Y \leq Q_{\{Y|D=1, S=1, X\}}(1-p_0(X), X)\}} - 1 + p_0(X) \right). \end{aligned}$$

**ASSUMPTION 8** (Regularity Conditions for Endogenous Sample Selection). *The following assumptions hold. (1) There exist both a lower bound  $\underline{s} > 0$  and an upper bound  $\bar{s} < \infty$  such that the conditional employment probability  $s(d, x)$  is bounded from above and below:  $0 < \underline{s} \leq s(D, X) \leq \bar{s} < \infty$  a.s. (2) Let  $\mathcal{U} \subset \mathcal{R}$  be an open set that contains the support of  $s(0, X)/s(1, X)$  and  $1 - s(0, X)/s(1, X)$  a.s. in  $X$ . Assume that the conditional quantile function  $u \rightarrow Q_{Y|D=1, S=1, X=x}(u, x)$  is differentiable on  $\mathcal{U}$  absolutely surely in  $X$ , and its derivative is bounded by some  $K_Q < \infty$   $\Pr(\sup_{u \in cl(\mathcal{U})} |\partial_u Q_{Y|D=1, S=1, X=x}(u, X)| \leq K_Q) = 1$ . (3) Assume that there exists conditional density  $\rho(y)$  of  $Y$  given  $D = 1, S = 1, X$  a.s. in  $X$  that is bounded and has bounded first derivative. (4) There exists a rate  $g_N = o(N^{-1/4})$  such that the first-stage estimates  $\hat{s}(D, X)$  and  $\hat{Q}_{Y|D=1, S=1, X}(u, x)$*

converge at  $g_N$  rate:  $\forall d \in \{1, 0\} \quad \|\hat{s}(d, X) - s(d, X)\|_{\{P, 2\}} \lesssim g_N, \sup_{u \in \mathcal{U}} |\hat{Q}_{Y|D=1, S=1, X}(u, x) - Q_{Y|D=1, S=1, X=x}(u, X)| \lesssim g_N.$

**Theorem 3** (Asymptotic Theory for Endogenous Sample Selection). *Suppose Assumption 8 holds. Then, the estimator  $(\hat{\beta}_L, \hat{\beta}_U)$  of Definition 5 obeys:*

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

where  $\Omega$  is a positive-definite covariance matrix.

Theorem 3 is my third main result. It establishes that the bounds defined by (11.12) are consistent and asymptotically normal. It extends the bounds estimator from Lee (2008), defined for a small number of discrete covariates, to the case of high-dimensional covariates that can be either discrete and continuous. Assumption 8 (4) about the first-stage rate is stated at a high level in order to incorporate various first-stage methods. For example, Assumption 8 (4) holds in high-dimensional sparse design when  $\hat{s}(d, x)$  is estimated by  $\ell_1$ -regularized logistic regression with data-driven choice of penalty (Belloni et al. (2016)) and  $\hat{Q}_{Y|D=1, S=1, X}(u, x)$  is estimated by  $\ell_1$ -penalized quantile regression (Belloni et al. (2013)). Because the estimators  $(\hat{\beta}_L, \hat{\beta}_U)$  are asymptotically equivalent to sample averages, they can be also used to construct the Imbens and Manski (2004)-type confidence interval for the treatment effect parameter under no additional assumptions.

It is worth considering the case when selection is exogenous conditional on controls. Then,  $p_0(X) = \frac{s(0, X)}{s(1, X)} = 1$  and  $\beta_L = \beta_U$ . Nevertheless, the sample estimators  $\hat{\beta}_L \neq \hat{\beta}_U$ , because the estimators of  $\hat{\beta}_L$  and  $\hat{\beta}_U$  are sample averages of two distinct orthogonal moment functions that have the same population mean. Therefore,  $\Omega$  remains to be positive-definite.

## 4.2 Average Partial Derivative

Consider the setup of Example 2. The constructed random variable  $V$  is equal to the derivative of the log conditional density  $V = -\partial_D \log f(D|X) = \eta_0(D, X)$ , the orthonormalized projection  $z_q(\eta) := q^\top \eta(D, X)$  and the  $q$ -generator  $Y_q = \Gamma(Y_L, Y_U - Y_L, q^\top \eta(D, X))$ . The orthogonal moment

equation (3.9) takes the form

$$g(W, q, \xi(q)) = z_q(\eta)Y_q(\eta) - q^\top V\gamma_q(D, X) + q^\top \partial_D \gamma_q(D, X), \quad (4.2)$$

where  $z_q(\eta) = q^\top \eta(D, X)$ ,  $Y_q(\eta) = \Gamma(Y_L, Y_U - Y_L, q^\top \eta(D, X))$ , and  $\gamma_q(D, X) = \gamma_L(D, X) + \gamma_{U-L}(D, X) \cdot 1_{\{q^\top \partial_D \log f(D|X) > 0\}}$ . The first-stage parameter of this problem is  $\xi(D, X) = \{\eta(D, X), \gamma_L(D, X), \gamma_{U-L}(D, X)\}$  whose true value  $\xi_0(D, X)$  is  $\xi_0(D, X) = \{\partial_D \log f(D|X), \mathbb{E}[Y_L|D, X], \mathbb{E}[Y_{U-L}|D, X]\}$ .

**ASSUMPTION 9** (Regularity Conditions for Average Partial Derivative). *The following conditions hold. (1) There exists a bound  $B_{UL} < \infty$  such that the interval width  $Y_U - Y_L \leq B_{UL}$  is bounded a.s. (2) The distribution of the gradient of the log-density  $\eta_0(D, X) = \partial_D \log f(D|X)$  is sufficiently continuous. For the event  $\mathcal{E}_q := \{0 < |q^\top \partial_D \eta_0(D, X)| < |q^\top \partial_D(\eta(D, X) - \eta_0(D, X))|\}$ , the following bound holds*

$$\sup_{q \in \mathcal{S}^{d-1}} \mathbb{E}|q^\top \partial_D(\eta(D, X) - \eta_0(D, X))| 1_{\{\mathcal{E}_q\}} \leq \|\eta(D, X) - \eta_0(D, X)\|_{L_{P,2}}.$$

(3) *There exist a rate  $g_N = o(N^{-1/4})$  such that each component of  $\hat{\xi}(D, X)$  converge at  $g_N$  rate:  $\|\hat{\xi}(D, X) - \xi_0(D, X)\|_{\{P,2\}} \lesssim g_N$ .*

**Theorem 4** (Asymptotic Theory for Average Partial Derivative with an Interval-Valued Outcome). *Suppose Assumption 9 holds. Then, Theorems 1 and 2 hold for the Support Function Estimator of Definition 4 with the influence function equal to*

$$h(W, q) = g(W, q, \xi(q)) - \mathbb{E}[g(W, q, \xi(q))],$$

where  $g(W, q, \xi(q))$  is given in (4.2).

Theorem 4 says that the Support Function Estimator given in (4) is uniformly consistent and asymptotically Gaussian. It extends the support function estimator of Kaido (2017), defined for a small number of covariates, to the case of high-dimensional covariates.

**Remark 1.** *Suppose the first-stage residual  $V = D - \mathbb{E}[D|X]$  is normally distributed independently from  $X$  (i.e,  $V \sim N(0, \Lambda)$ ). Then, Assumption 9 (2) holds. Furthermore, the estimator of the support function can be simplified as shown in the algorithm below.*

---

**Algorithm 1** Support Function Estimator for Average Partial Derivative
 

---

Let  $m_0(X) = \mathbb{E}[D|X]$ . Input: a direction  $q$  on a unit sphere  $\mathbb{S}^{d-1}$ , 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  $(\hat{m}(X_i), \hat{\gamma}_L(D_i, X_i), \hat{\gamma}_{U-L}(D_i, X_i))_{i=1}^N$ .

- 1: Estimate the first-stage residual for every  $i \in \{1, 2, \dots, N\}$ :  $\hat{V}_i := D_i - \hat{m}(X_i)$ .
- 2: Compute the sample covariance matrix of the first-stage residuals:  $\hat{\Lambda} := \frac{1}{N} \sum_{i=1}^N \hat{V}_i \hat{V}_i^\top$ .
- 3: Estimate the  $q$ -generator for every  $i \in \{1, 2, \dots, N\}$

$$\hat{Y}_{q,i} := Y_{L,i} + (Y_{U,i} - Y_{L,i}) \mathbf{1}_{\{q^\top \hat{\Lambda}^{-1} \hat{V}_i > 0\}}.$$

- 4: Compute the second-stage reduced form  $\hat{\gamma}_q(D_i, X_i) := \hat{\gamma}_L(D_i, X_i) + \hat{\gamma}_{U-L}(D_i, X_i) \mathbf{1}_{\{q^\top \hat{\Lambda}^{-1} \hat{V}_i > 0\}}$
- 5: Estimate  $\hat{\beta}_q$  by Ordinary Least Squares with the second-stage residual of the  $q$ -generator as the dependent variable and the first-stage residual  $V$  as the regressor

$$\hat{\beta}_q = \hat{\Lambda}^{-1} \frac{1}{N} \sum_{i=1}^N \hat{V}_i [\hat{Y}_{q,i} - \hat{\gamma}_q(D_i, X_i)].$$

Return: the projection of  $\hat{\beta}_q$  on the direction  $q$ :  $\hat{\sigma}(q, \mathcal{B}) = q^\top \hat{\beta}_q$ .

---

### 4.3 Partially Linear Predictor

Consider the setup in Example 3. The constructed variable  $V_\eta = D - \eta(X)$  is equal to the first-stage residual, the orthonormalized projection  $z_q(\eta)$  is equal to the inner product of this residual and the vector  $p(q) = (\Sigma^{-1})^\top q$ ,  $z_q(\eta) = q^\top \Sigma^{-1} (D - \eta(X))$ , and the  $q$ -generator  $Y_q$  is equal to  $Y_q(\eta) = Y_L + (Y_U - Y_L) \mathbf{1}_{\{q^\top \Sigma^{-1} (D - \eta(X)) > 0\}}$ . Equation (3.6) does not satisfy Assumption 5(1). The Neyman orthogonal moment equation (3.9) is

$$\psi(W, \theta(q), \xi(p(q))) = \begin{bmatrix} \sigma(q, \mathcal{B}) - p(q)^\top (D - \eta(X)) (Y_q(\eta) - \mathbb{E}[Y_q(\eta)|X]) \\ (D - \eta(X)) (D - \eta(X))^\top p(q) - q \end{bmatrix}, \quad (4.3)$$

where the true value of  $\theta(q)$  is  $\theta_0(q) = [\sigma(q, \mathcal{B}), p(q)]$  and the true value of the nuisance parameter  $\xi(p) = \{\eta(X), \gamma(p, X)\}$  is

$$\xi_0(p(q)) = \{\eta_0(X), \mathbb{E}[Y_q(\eta_0)|X]\}.$$

Assumption 10 gives the regularity conditions for the Support Function Estimator.

**ASSUMPTION 10** (Regularity Conditions for Partially Linear Predictor). *The following regular-*

ity conditions hold. (1) There exists  $\bar{D} < \infty$  such that  $\max(\|D\|, |Y_L|, |Y_U|) \leq \bar{D}$  holds absolutely surely. (2) For all vectors  $p \in \mathcal{P}$  there exists a conditional density  $\rho_{\{p^\top V|X=x\}}(\cdot, x)$  absolutely surely in  $X$ . (3) A bound  $K_h < \infty$  exists such that the collection of the densities in (2)  $\{\rho_{\{p^\top V|X=x\}}(\cdot, x), p \in \mathcal{P}\}$  is uniformly bounded over  $p \in \mathcal{P}$  a.s. in  $X$ . (4) Let  $\mathcal{F}_\gamma = \{\gamma(p, x) : \mathcal{P} \times \mathcal{X} \rightarrow \mathcal{R}\}$  be a class of functions in  $p, x$  that satisfy Conditions 6(1,2). Moreover, there exists a sequence of realization sets  $\mathcal{G}_N$  that are subsets of  $\mathcal{F}_\gamma$   $\mathcal{G}_N \subseteq \mathcal{F}_\gamma$  such that the estimator  $\hat{\gamma}(\cdot, \cdot)$  belongs to  $\mathcal{G}_N$  with probability at least  $1 - \phi_N$ . Moreover, the nuisance realization set shrinks at a statistical rate uniformly in  $p \in \mathcal{P}$   $\sup_{p \in \mathcal{P}} \sup_{\gamma(\cdot, \cdot) \in \mathcal{G}_N} \|\gamma(p, X) - \gamma_0(p, X)\|_{L_{P,2}} \leq g_N = o(N^{-1/4})$ . (5) Let  $\gamma_0(p, x)$  be the conditional expectation on  $X$  of the  $q$ -generator  $\gamma_0(p, x) = \mathbb{E}[Y_q|X] = \mathbb{E}[\Gamma(Y_L, Y_U - Y_L, p^\top V_{\eta_0})|X]$ . Assume that there exists a sequence  $g'_N$  such that  $\gamma_0(p, x)$  is continuous uniformly on  $\mathcal{P}$ , on average in  $X$ :  $\sup_{q \in \mathcal{S}^{d-1}} \sup_{p \in \mathcal{P}: \|p - p_0(q)\| \leq RN^{-1/2}} \|\gamma_0(p, X) - \gamma_0(p_0(q), X)\|_{L_{P,2}} \leq g'_N : g'_N \log(1/g'_N) = o(1)$ . (6) The first-stage residual  $D - \eta_0(X)$  has a uniformly sufficiently smooth distribution on  $\mathcal{P}$ . Namely, for some  $m$  such that  $0 < m \leq 1$ , the following bound holds:  $\sup_{p \in \mathcal{P}} \Pr(0 < \frac{p^\top (D - \eta_0(X))}{\|D - \eta_0(X)\|} < \delta) = O(\delta^m), \delta \rightarrow 0$ .

**Theorem 5** (Asymptotic Theory for Partially Linear Predictor with an Interval-Valued Outcome). *Suppose Assumption 4 and 10 holds. Then, Theorem 1 and 2 hold for the Support Function Estimator with the influence function  $h(W, q)$  equal to*

$$h(W, q) = g(W, p_0(q), \xi_0(p_0(q), X)) - \mathbb{E}[g(W, p_0(q), \xi_0(p_0(q), X))] \\ - q^\top \Sigma^{-1} ((D - \eta_0(X))(D - \eta_0(X))' - \Sigma) \Sigma^{-1} \mathbb{E}(D - \eta_0(X)) Y_q,$$

where  $p_0(q) = (\Sigma^{-1})^\top q$  and  $Y_q = Y_L + (Y_U - Y_L) 1_{q^\top \Sigma^{-1} (D - \eta_0(X)) > 0}$ .

Theorem 5 is my fifth main result. It establishes that the Support Function Estimator given in Algorithm 2 is uniformly consistent and asymptotically normal.

I describe the computation steps of the Support Function Estimator  $\hat{\sigma}(q, \mathcal{B})$  in the following algorithm.

---

**Algorithm 2** Support Function Estimator for Partially Linear Predictor
 

---

Input: a direction  $q$  on a unit sphere  $\mathcal{S}^{d-1}$ , 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  $(\hat{\eta}(X_i), \hat{\gamma}(p, X_i))_{i=1}^N, p \in \mathcal{P}$ .

- 1: Estimate the first-stage residual for every  $i \in \{1, 2, \dots, N\}$ :  $\hat{V}_i := D_i - \hat{\eta}(X_i)$ .
- 2: Compute the sample covariance matrix of the first-stage residuals:  $\hat{\Sigma} := \frac{1}{N} \sum_{i=1}^N \hat{V}_i \hat{V}_i^\top$ .
- 3: Estimate the  $q$ -generator for every  $i \in \{1, 2, \dots, N\}$

$$\hat{Y}_{q,i} := Y_{L,i} + (Y_{U,i} - Y_{L,i}) \mathbf{1}_{\{q^\top \hat{\Sigma}^{-1} \hat{V}_i > 0\}}.$$

- 4: Estimate  $\hat{\beta}_q$  by Ordinary Least Squares with the second-stage residual of the  $q$ -generator as the dependent variable and the first-stage residual  $V$  as the regressor

$$\hat{\beta}_q = \hat{\Sigma}^{-1} \frac{1}{N} \sum_{i=1}^N \hat{V}_i [\hat{Y}_{q,i} - \hat{\gamma}(\hat{\Sigma}^{-1} q^\top, X_i)].$$

Return: the projection of  $\hat{\beta}_q$  on the direction  $q$ :  $\hat{\sigma}(q, \mathcal{B}) = q^\top \hat{\beta}_q$ .

---

**ASSUMPTION 11** (Simple Sufficient Conditions for Partially Linear Predictor). *The following conditions hold. (1) The first-stage residual is independent from the covariates  $X$  and has a symmetric continuous distribution around zero (i.e.,  $\Pr(p^\top V > 0) = \frac{1}{2} \quad \forall p \in \mathcal{R}^d$ ). (2) Conditional on  $X$ , the interval width  $Y_U - Y_L$  and the covariate  $D$  are mean independent*

$$\mathbb{E}[(Y_U - Y_L) \mathbf{1}_{\{p^\top V > 0\}} | X] = \mathbb{E}[Y_U - Y_L | X] \Pr(p^\top V > 0 | X) = \frac{1}{2} \mathbb{E}[Y_U - Y_L | X].$$

(3) *There exists a sequence of realization sets  $\{Z_{L,N}, N \geq 1\}$  and  $Z_{U-L,N}, N \geq 1\}$  that are shrinking neighborhoods of  $\gamma_{L,0}(X) := \mathbb{E}[Y_L | X]$  and  $\gamma_{U-L,0}(X) := \mathbb{E}[Y_U - Y_L | X]$  obeying the following conditions. For some sequence  $\phi_N = o(1)$  the estimate  $\hat{\gamma}_L$  of  $\gamma_{L,0}$  and  $\hat{\gamma}_{U-L}$  of  $\gamma_{U-L,0}$  belong to respective sets  $Z_{L,N}$  and  $Z_{U-L,N}$  with probability at least  $1 - \phi_N$ . There exist rates  $\zeta_{L,N} = o(N^{-1/4})$  and  $\zeta_{U-L,N} = o(N^{-1/4})$  such that the following bounds hold*

$$\sup_{\gamma_L \in Z_{L,N}} (\mathbb{E}(\gamma_L(X) - \gamma_{L,0}(X))^2)^{1/2} \lesssim \zeta_{L,N}, \quad \sup_{\gamma_{U-L} \in Z_{U-L,N}} (\mathbb{E}(\gamma_{U-L}(X) - \gamma_{U-L,0}(X))^2)^{1/2} \lesssim \zeta_{U-L,N}.$$

**Remark 2.** *Simple Sufficient Conditions for Partially Linear Predictor Suppose Assumption 11*

holds. Then, the conditional expectation function  $\gamma_0(q, X)$  no longer depends on  $q$ :

$$\gamma_0(p, X) := \gamma_0(X) = \mathbb{E}[Y_L|X] + \frac{1}{2}\mathbb{E}[Y_U - Y_L|X] = \gamma_{L,0}(X) + \frac{1}{2}\gamma_{U-L,0}(X).$$

Therefore, Assumptions 10 (4)-(6) are satisfied as long as the functions  $\gamma_{L,0}(X)$  and  $\gamma_{U-L,0}(X)$  are estimated at the  $o(N^{-1/4})$  rate.

## 5 Empirical application

In this section, I re-examine the effectiveness of JobCorps, a federally funded educational and vocational training program founded at 1964 and aimed at helping young people aged 16 through 24 to improve their labor market outcomes. In mid-1990s, the program conducted a study that randomly split applicants into the treatment and the control groups. The control group of 5977 individuals was embargoed from the program for following three years, while the applicants in the treatment group could enroll in the JobCorps as usual. Following Lee (2008), I am interested in the effect of having access to JobCorps training today on wage earned several years later.

I will use the notation of Example 1 to formalize the problem. The variable  $D = 1$  stands for the indicator of treatment receipt,  $S_{0,t} = 1$  and  $S_{1,t} = 1$  stand for the potential employment outcomes in the week  $t$  in case of being in the control and the treatment group, respectively. Similarly,  $Y_{0,t} = 1$  and  $Y_{1,t} = 1$  stand for the potential wages. I am interested in the average JobCorps effect on wages of the individuals who would have been employed at week  $t$  regardless of JobCorps enrollment

$$\mathbb{E}[Y_{1,t} - Y_{0,t} | S_{1,t} = 1, S_{0,t} = 1],$$

or briefly, the always-employed. The research sample contains the treatment status, the number of hours worked each week, and the weekly earnings. In addition, the data set contains rich baseline information including demographics, employment, training, welfare, criminal, health history collected throughout a year before random assignment on a weekly basis and represented by more than 5000 covariates. Following Lee (2008), the sample consists of 9145 individuals whose number of working hours and weekly earnings have been recorded (non-missing) each week since random assignment.

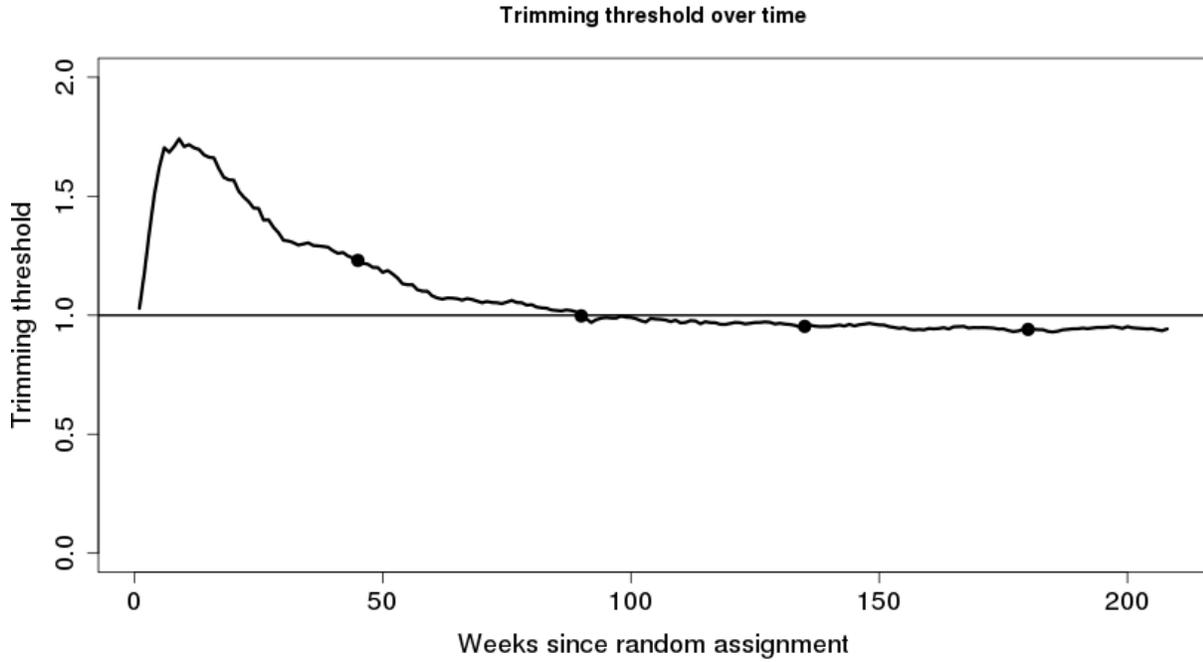


Figure 1: Trimming threshold against time. The trimming threshold is the ratio of employment rates in the control group (in the numerator) to the treatment group (in the denominator). The black dots represent the findings reported in Lee (2008).

I will first present the basic Lee (2008) bounds without using any covariates. As described in Example 1, I assume that JobCorps training affects everyone’s employment in the same direction at each week  $t$ , but this direction may differ across weeks (i.e., Assumption 1 holds for every week  $t$ ). Then, the control-to-treat ratio of employment rates (the trimming threshold)  $\hat{p}_{0,t} = \frac{\frac{1}{\sum_{i:D_i=0} S_{i,t}}}{\frac{1}{\sum_{i:D_i=1} S_{i,t}}}$  is less than one if JobCorps encourages employment and is greater than one otherwise. Figure 1 plots the trimming threshold  $\hat{p}_{0,t}$ . Under Assumption 1, JobCorps must deter employment in the first 89 weeks and encourage employment starting week 90. Figure 2 plots the basic (no-covariate) Lee (2008) bounds and the 95-% confidence intervals for the bounds.

I will now allow the sign of the JobCorps effect on the employment to depend on the covariates (i.e., I will replace Assumption 1 by Assumption 2). For each week  $t$ , I estimate the employment equation

$$S_{it} = \begin{cases} 1, & X_i' \gamma_t + D_i X_i' \bar{w}_t + \varepsilon_{it} > 0 \\ 0, & \text{otherwise} \end{cases}, \quad i = 1, 2, \dots, N, \quad (5.1)$$

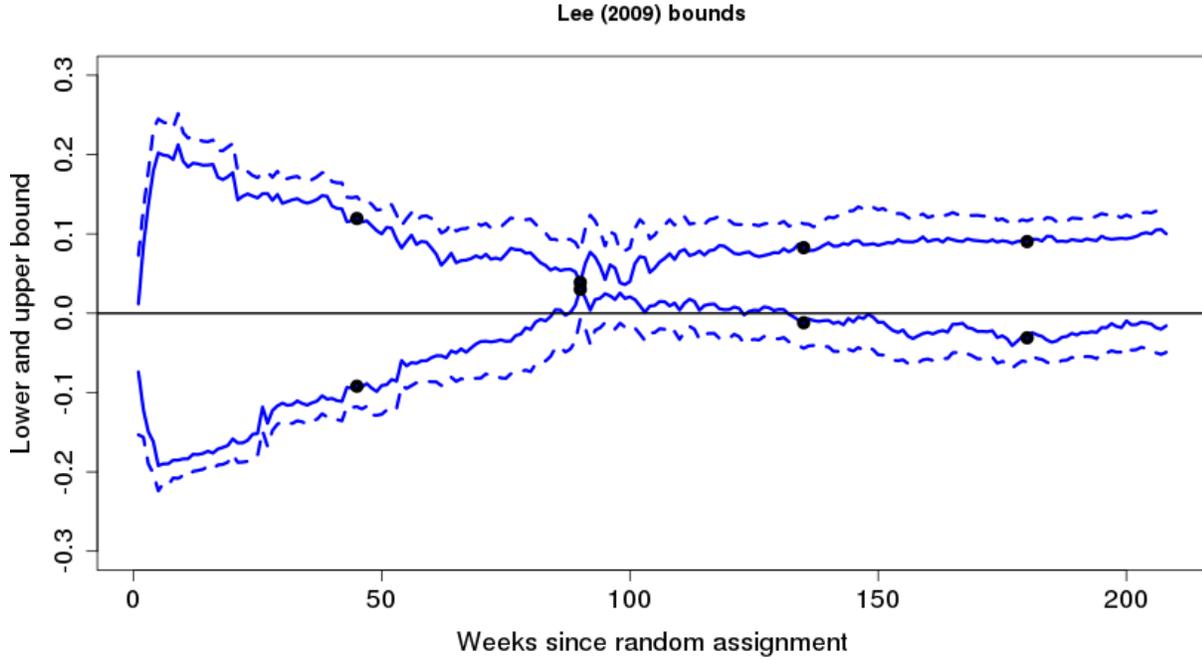


Figure 2: Lee (2008) no-covariate bounds and 95% pointwise confidence bands. The Y-axis corresponds to log hourly wage defined for a subject with a positive number of working hours in a given week. The blue solid lines are the Lee (2008) no-covariate upper and lower bound on the JobCorps wage effect based on standard monotonicity assumption (Assumption 1). The blue dashed lines are 95% pointwise confidence interval for the identified set parameter, computed by bootstrap with  $B = 500$  repetitions. The black dots correspond to Lee (2008) empirical findings.

where  $\gamma_t$  and  $\gamma_t + \bar{w}_t$  are the effects of the pre-treatment covariates  $X_i = (1, \bar{X}_i)$  in the control group and treatment groups, respectively. The shock  $\varepsilon_{it}$  is the type 1 extreme value shock that is independent across time. I assume that few of the covariates  $X_i$  have a non-zero effect on the selection, but the identities of those covariates are unknown. I estimate Equation (5.1) using logistic lasso with the data-driven choice of penalty as in Belloni et al. (2016). As a result, the set  $\mathcal{X}_0^t = \{ \frac{s(0, X)}{s(1, X)} < 1 \}$  can be estimated as

$$\widehat{\mathcal{X}}_0^t := \left\{ i, \frac{\widehat{s}(0, X_i)}{\widehat{s}(1, X_i)} < 1 \right\}$$

as the set of all observations where the conditional trimming threshold is less than one, and  $\widehat{\mathcal{X}}_1^t = \{ \cup_{i=1}^N i \setminus \widehat{\mathcal{X}}_0^t \}$  can be taken to be its complement.

Figure 3 plots the treatment-control difference in employment rates computed for the subjects

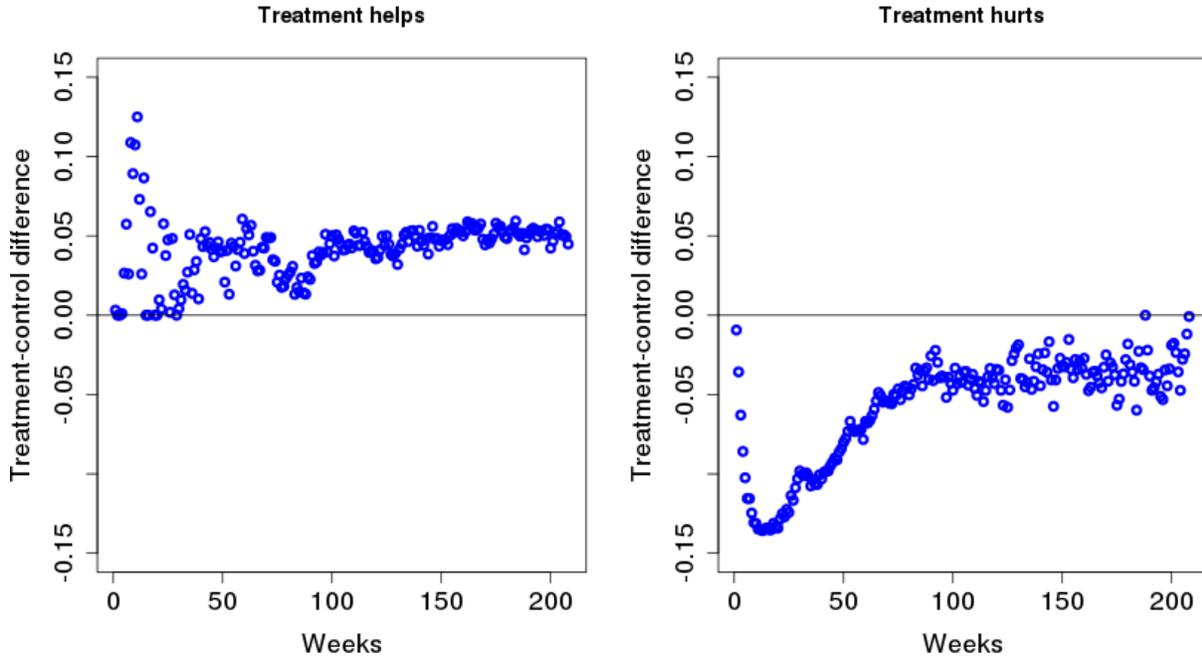


Figure 3: Treatment-control difference in employment rates against time. The trimming threshold (5.1) is estimated using standard logistic regression with 27 covariates pre-selected by Lee (2008). For each week  $t \in \{1, 2, \dots, 208\}$ , the set  $\hat{\mathcal{X}}_0^t$  (left panel, "Treatment helps") consists of the individuals whose trimming threshold is less than one. The set  $\hat{\mathcal{X}}_1^t$  (right panel) is taken to be its complement.

in the set  $\hat{\mathcal{X}}_0^t$  (left panel) and  $\hat{\mathcal{X}}_1^t$  (right panel) against time. For almost all weeks, the treatment-control difference is strictly positive in  $\hat{\mathcal{X}}_0^t$  and strictly negative in  $\hat{\mathcal{X}}_1^t$ . Therefore, the standard monotonicity assumption (Assumption 1) cannot be true. Figure 4 plots the fraction of subjects in  $\hat{\mathcal{X}}_0^t$  as a function of time. In the first weeks after random assignment, JobCorps deters employment for nearly all individuals during the program course. By the end of year 2, JobCorps helps to be employed for nearly half of the individuals, and this fraction goes to 0.8 by the end of the fourth year where tracking wages is stopped. These findings are consistent with JobCorps program description. Enrolling in JobCorps provides full-time training and benefits (board, meals, and medical care) and thus discourages employment. By the end of the training, JobCorps graduates may have gained useful experience that helps them in finding job, leading to higher employment rate in the treatment group relative to the control one. Figure 5 plots the no-covariate Lee (2008) bounds under standard monotonicity (in blue), which we refer to as monotone, and conditional monotonicity (black), which we refer to as non-monotone. As expected, the discrepancy between monotone and non-monotone bounds is the largest at week 104, where the unconditional monotonicity is far

from the truth for half of the subjects.

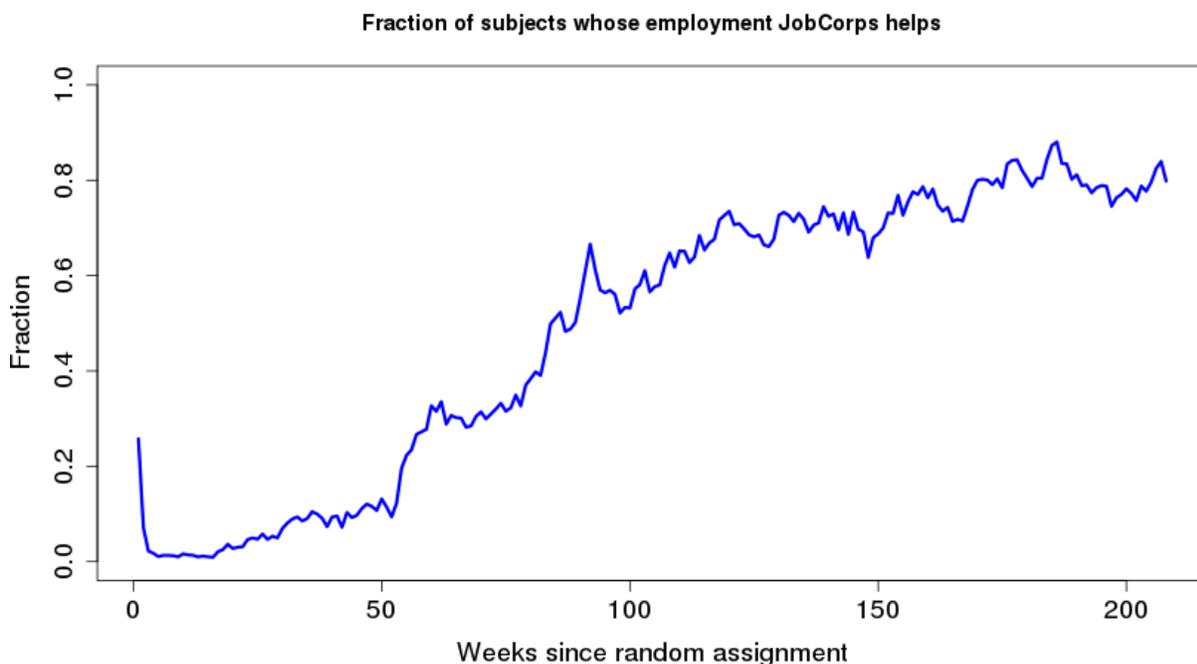


Figure 4: Fraction of subjects in the set  $\hat{\mathcal{X}}_0^t$  against time. The trimming threshold (5.1) is estimated using standard logistic regression with 27 covariates pre-selected by Lee (2008). The set  $\hat{\mathcal{X}}_0^t$  consists of the individuals whose trimming threshold is less than one (i.e., whose estimated treatment effect on employment is positive).

Figure 7 compares the estimates of non-sharp (in blue, without covariates) and sharp (in red, with covariates) non-monotone bounds for two specifications of the covariates. The first specification (Figure 7a) is based on 27 demographic covariates selected by Lee (2008) for his point-identified parametric (i.e., Heckman (1979)) specification. Since some of those covariates are continuous (e.g., baseline work experience, earnings), they could not be properly used to estimate sharp bounds by Lee (2008)’s approach. The reported sharp bounds (in red) are the orthogonal estimates based on standard logistic regression and quantile regression estimates of the first-stage functions. On average, the width of the sharp 27-covariate bounds is 80% of the width of the no-covariate bounds and is equal to 0.13. I use these bounds as a benchmark for what can be achieved by classic nonparametric, as opposed to modern regularized methods.

Figure 7b shows the estimates of non-sharp (in blue, without covariates) and sharp (in red, with all 5177 covariates) non-monotone bounds. The sharp bounds (in red) are the orthogonal estimates

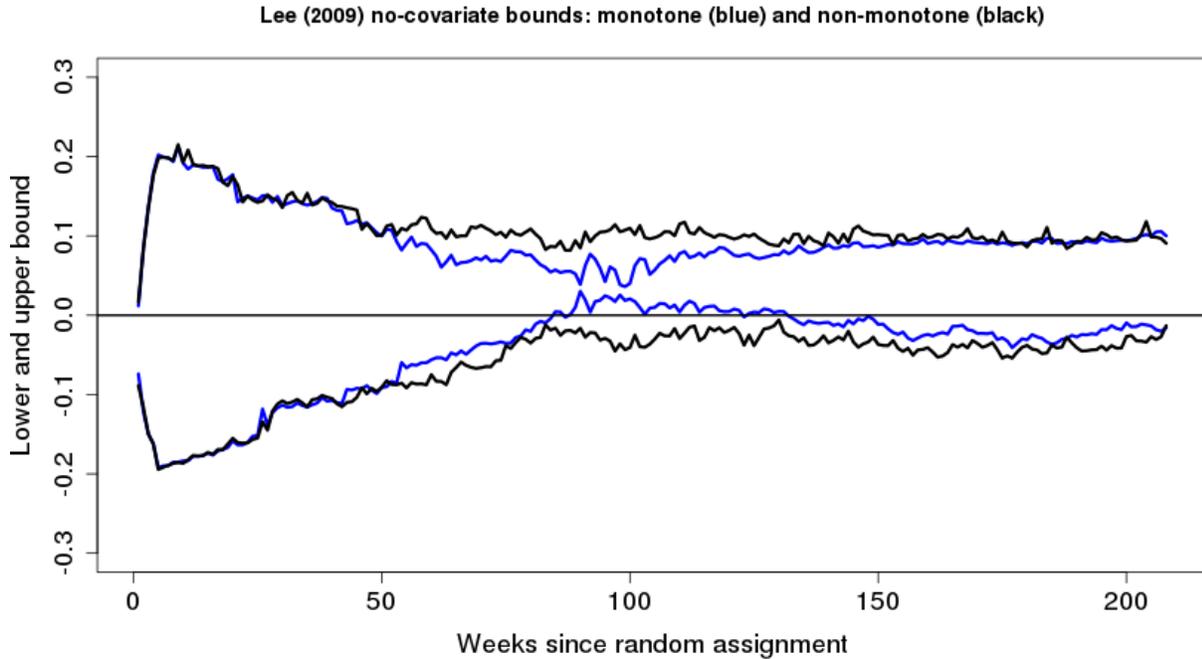
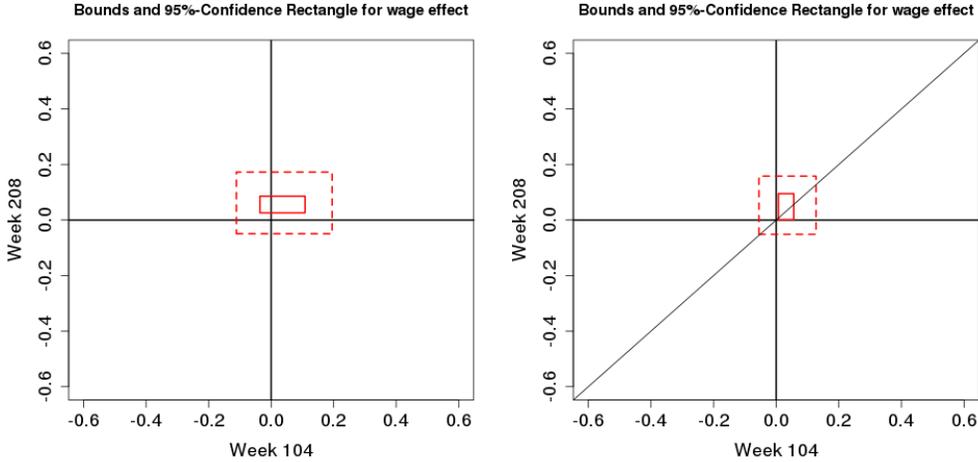


Figure 5: The no-covariate Lee (2008) bounds under standard monotonicity (blue) and conditional monotonicity (black). The trimming threshold (5.1) is estimated using standard logistic regression with 27 covariates pre-selected by Lee (2008).

based on  $\ell_1$ -penalized logistic regression and  $\ell_1$ -penalized conditional quantile of the first-stage functions with the data-driven choice of penalties as in Belloni et al. (2016) and Belloni et al. (2013), respectively. Using all baseline covariates and regularized methods in the first stage allows for substantial improvements relative to the classic nonparametric specification. First, all-covariate lower bound on wage effect is positive in 66% of the weeks 90 through 208. In contrast, the no-covariate bound is always negative and the 27-covariate lower bound is positive only in 19% of those weeks. Second, the average width of the all-covariate bounds is only 58% of the width of the no-covariate bounds and 76% of the width of the 27-covariate bounds. Finally, the simultaneous confidence rectangle (Figure 6) for the two-dimensional wage effect at the weeks 104 and 208 is substantially smaller based on the all-covariate specification.

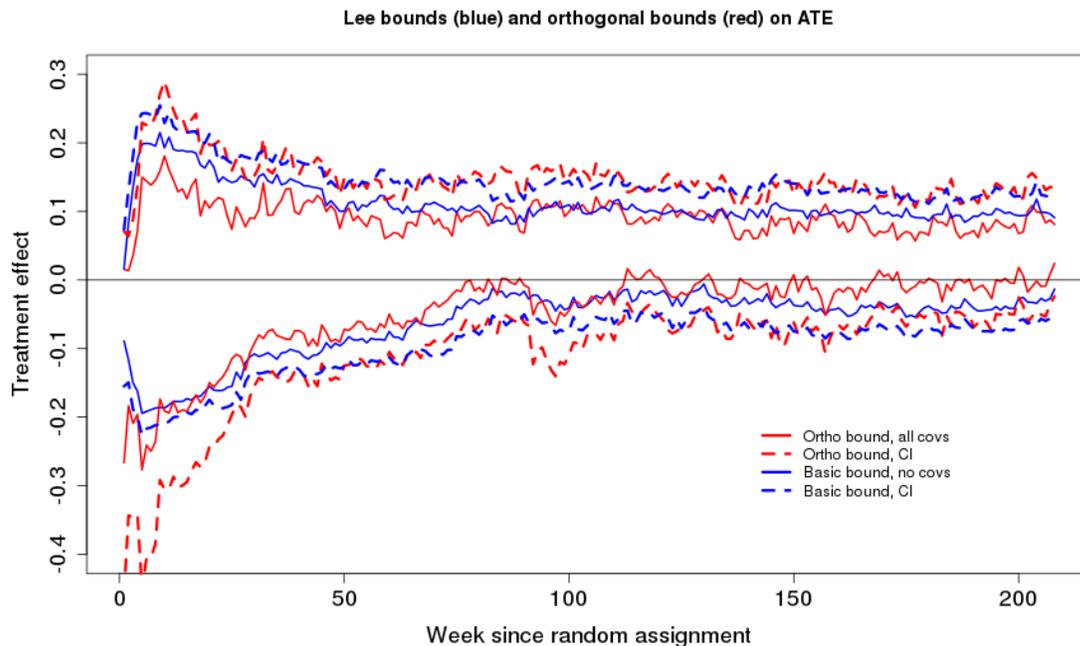
Figure 6 shows the estimate and the 95%-simultaneous confidence rectangle for the two-dimensional wage effect at the weeks 104 and 208, based on the 27-covariate (Figure 6a) and all-covariate estimate (Figure 6b). While one cannot rule out the effect  $(0,0)$  by either method, one can rule out the hypothesis of large negative effects less than  $-0.1$  for both weeks and large



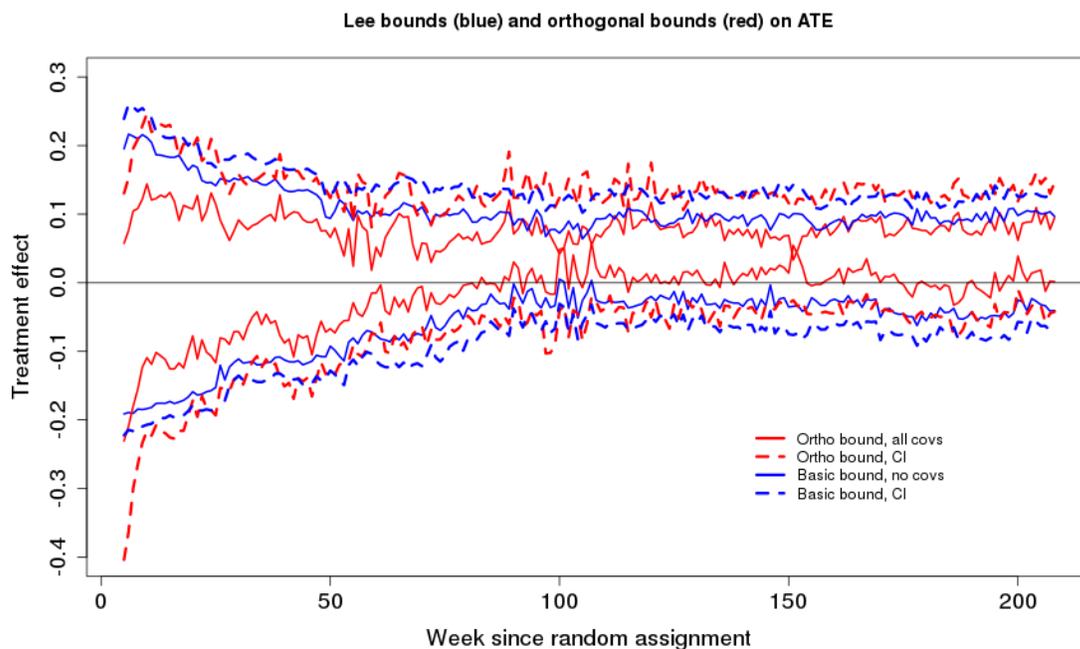
(a) Logistic regression based on 27 pre-selected/ covariates (b) Logistic  $\ell_1$ -penalized regression based on 5177 covariates

Figure 6: Identified set and the 95% pointwise confidence rectangle for this set, where the target parameter is the two-dimensional wage effect at weeks 104 and 208 . Figure 6a is based on 27 covariates pre-selected by Lee (2008) and the logistic regression estimate of the trimming threshold. Figure 6b is based on all 5177 covariates and  $\ell_1$ -regularized logistic regression (Belloni et al. (2016)) with a data-driven choice of penalty. The boundary of the identified set (red solid line) is estimated by orthogonal estimate. The boundary of the confidence rectangle (red dashed line) is estimated by weighted orthogonal bootstrap (Definition 6) with  $B = 500$  repetitions and exponential weights.

positive effects exceeding 0.2 for both weeks using the model based on all covariates. Figure 6 highlights the use of asymptotic theory for the support function, as well as the novelty of this work relative to the pointwise asymptotic results (Chernozhukov et al. (2017a), Chernozhukov et al. (2017b)). Indeed, the reported confidence rectangle allows to conduct simultaneous inference about the projections of the multi-dimensional effect on any direction of interest, while the pointwise results are applicable only when the treatment effect is one-dimensional.



(a) Logistic regression based on 27 pre-selected covariates



(b) Logistic  $\ell_1$ -penalized regression based on 5177 covariates

Figure 7: Bounds on the wage effect and 95% pointwise confidence bands. Figure 7a and 7b use different choices of the covariates. Figure 7a is based on 27 covariates pre-selected by Lee (2008) and the logistic regression estimate of the trimming threshold. Figure 7b is based on all 5177 covariates and  $\ell_1$ -regularized logistic regression (Belloni et al. (2016)) with a data-driven choice of penalty. The blue solid line is the no-covariate non-monotone upper and lower bound on the wage effect. The blue dashed line is the 95% pointwise confidence bands for the identified set, based on regular bootstrap with  $B = 500$  repetitions. The red solid line is the all-covariate non-monotone orthogonal estimate of the bound on the wage effect. The red dashed line is the 95% pointwise confidence bands for the identified set, based on weighted orthogonal bootstrap (Definition 6) with  $B = 500$  repetitions and exponential weights.

## 6 Conclusion

In this paper, I incorporate machine learning tools into set-identification and harness their predictive power to tighten an identified set. I focus on the set-identified models with high-dimensional covariates and provide two-stage estimation and inference methods for an identified set. In the first stage, I select covariates (or estimate a nonparametric function of them) using machine learning tools. In the second stage, I plug the estimates into the moment equation for the identified set's boundary that is insensitive, or, formally, Neyman-orthogonal, to the bias in the first-stage estimates. I establish the uniform limit theory for the proposed estimator and the weighted bootstrap procedure and provide a general recipe to construct a Neyman-orthogonal moment function starting from a non-orthogonal one.

My method's main application is to estimate Lee (2008) nonparametric bounds on the average treatment effect in the presence of endogenous selection. I derive a Neyman-orthogonal moment equation for Lee (2008)'s bounds and provide primitive sufficient conditions for their validity. Moreover, I substantially tighten Lee (2008)'s bounds on the JobCorps wage effect in Lee (2008)'s empirical setting with 5000+ covariates (Section 5) and Angrist et al. (2006)'s empirical setting (Supplementary Appendix). In addition, I also provide the low-level sufficient conditions to estimate sharp identified sets for two other parameters - the causal parameter in the partially linear model and the average partial derivative when the outcome variable is interval-censored.

## 7 Appendix

**Notation.** We use the standard notation for vector and matrix norms. For a vector  $v \in \mathcal{R}^d$ , denote the  $\ell_2$  norm of  $a$  as  $\|v\|_2 := \sqrt{\sum_{j=1}^d v_j^2}$ . Denote the  $\ell_1$  norm of  $v$  as  $\|v\|_1 := \sum_{j=1}^d |v_j|$ , the  $\ell_\infty$  norm of  $v$  as  $\|v\|_\infty := \max_{1 \leq j \leq d} |v_j|$ , and  $\ell_0$  norm of  $v$  as  $\|v\|_0 := \sum_{j=1}^d 1_{\{a_j \neq 0\}}$ . Denote a unit sphere as  $\mathcal{S}^{d-1} = \{\alpha \in \mathcal{R}^d : \|\alpha\| = 1\}$ . For a matrix  $M$ , denote its operator norm by  $\|M\|_2 = \sup_{\alpha \in \mathcal{S}^{d-1}} \|M\alpha\|$ . We use standard notation for numeric and stochastic dominance. For two numeric sequences  $\{a_n, b_n\}, n \geq 1$   $a_n \lesssim b_n$  stands for  $a_n = O(b_n)$ . For two sequences of random variables  $\{a_n, b_n, n \geq 1\}$ :  $a_n \lesssim_P b_n$  stands for  $a_n = O_P(b_n)$ . Finally, let  $a \wedge b = \min\{a, b\}$  and  $a \vee b = \max\{a, b\}$ . For a random variable  $\xi$ ,  $(\xi)^0 := \xi - \mathbb{E}[\xi]$ .

Fix a partition  $k$  in a set of partitions  $[K] = \{1, 2, \dots, K\}$ . Define the sample average of a function  $f(\cdot)$  within this partition as:  $\mathbb{E}_{n,k}[f] = \frac{1}{n} \sum_{i \in J_k} f(x_i)$  and the scaled normalized sample average as:

$$\mathbb{G}_{n,k}[f] = \frac{\sqrt{n}}{n} \sum_{i \in J_k} [f(x_i) - \mathbb{E}[f(x_i) | J_k^c]],$$

where  $[\cdot | J_k^c] := [\cdot | (W_i, i \in J_k^c)]$ . For each partition index  $k \in [K]$  define an event  $\mathcal{E}_{n,k} := \{\hat{\xi}_k \in \Xi_n\}$  as the nuisance estimate  $\hat{\xi}_k$  belonging to the nuisance realization set  $\mathcal{G}_N$ . Define  $\mathcal{E}_N = \bigcap_{k=1}^K \mathcal{E}_{n,k}$  as the intersection of such events.

## 7.1 Proofs

This section contains the proofs of main results of this paper. Section 7.1.1 contains the proofs of Theorem 1 and Theorem 2 from Section 3. Section 7.1.2 contains the proofs of Theorem 5. Other Lemmas can be found in Supplementary Appendix.

### 7.1.1 Proofs of Section 3

*Proof of Theorem 1,  $\Sigma$  is known.* To simplify notation, we assume  $\Sigma = I_d$ . The proof holds for any invertible matrix  $\Sigma$ . Let us focus on the partition  $k \in [K]$ .

$$\begin{aligned} \sqrt{n} |\mathbb{E}_{n,k}[g(W_i, q, \hat{\xi}(q)) - g(W_i, q, \xi_0(q))]| &\leq \sqrt{n} |\mathbb{E}[g(W_i, q, \hat{\xi}(q)) - g(W_i, q, \xi_0(q))]| \\ &\quad + |\mathbb{G}_{n,k}[g(W_i, q, \hat{\xi}(q)) - g(W_i, q, \xi_0(q))]| \\ &=: |i(q)| + |ii(q)|. \end{aligned}$$

**Step 1.** Recognize that  $|i(q)|$  converges to zero conditionally on the partition complement  $J_k^c$  and the event  $\mathcal{E}_N$ :

$$\begin{aligned}
|i(q)| &:= \sup_{q \in \mathcal{S}^{d-1}} \sqrt{n} |\mathbb{E}[g(W_i, q, \hat{\xi}(q)) - g(W_i, q, \xi_0(q)) | \mathcal{E}_N \cup J_k^c]| \\
&\leq \sup_{q \in \mathcal{S}^{d-1}} \sup_{\xi \in \Xi_n} \sqrt{n} |\mathbb{E}[g(W_i, q, \xi(q)) - g(W_i, q, \xi_0(q)) | \mathcal{E}_N \cup J_k^c]| \\
&\leq \sup_{q \in \mathcal{S}^{d-1}} \sup_{\xi \in \Xi_n} \sqrt{n} |\mathbb{E}[g(W_i, q, \xi(q)) - g(W_i, q, \xi_0(q))]| \\
&\leq \sqrt{n} \mu_n = o(1).
\end{aligned}$$

By Lemma 6.1 of Chernozhukov et al. (2017a), the term  $i(q) = O(\mu_n) = o(1)$  unconditionally.

**Step 2.** To bound the second quantity, consider the function class

$$\mathcal{F}_{\hat{\xi}\xi_0} = \{g(W_i, q, \hat{\xi}(q)) - g(W_i, q, \xi_0(q)), \quad q \in \mathcal{S}^{d-1}\}.$$

for some fixed  $\hat{\xi}$ . By definition of the class,

$$\mathbb{E} \sup_{q \in \mathcal{S}^{d-1}} |ii(q)| := \mathbb{E} \sup_{f \in \mathcal{F}} |\mathbb{G}_{n,k}[f]|.$$

We apply Lemma 6.2 of Chernozhukov et al. (2017a) conditionally on the hold-out sample  $J_k^c$  and the event  $\mathcal{E}_N$  so that  $\hat{\xi}(q) = \hat{\xi}_k$  can be treated as a fixed member of  $\Xi_n$ . The function class  $\mathcal{F}_{\hat{\xi}\xi_0}$  is obtained as the difference of two function classes:  $\mathcal{F}_{\hat{\xi}\xi_0} := \mathcal{F}_{\hat{\xi}} - \mathcal{F}_{\xi_0}$ , each of which has an integrable envelope and bounded logarithm of covering numbers by Assumption 6. In particular, one can choose an integrable envelope as  $F_{\hat{\xi}\xi_0} := F_{\hat{\xi}} + F_{\xi_0}$  and bound the covering numbers as:

$$\begin{aligned}
\log \sup_Q N(\varepsilon \|F_{\hat{\xi}\xi_0}\|_{Q,2}, \mathcal{F}_{\hat{\xi}\xi_0}, \|\cdot\|) &\leq \log \sup_Q N(\varepsilon \|F_{\hat{\xi}}\|_{Q,2}, \mathcal{F}_{\hat{\xi}}, \|\cdot\|) + \log \sup_Q N(\varepsilon \|F_{\xi_0}\|_{Q,2}, \mathcal{F}_{\xi_0}, \|\cdot\|) \\
&\leq 2\nu \log(a/\varepsilon), \quad \text{for all } 0 < \varepsilon \leq 1.
\end{aligned}$$

Finally, we can choose the speed of shrinkage  $(r'_n)^2$  such that

$$\sup_{q \in \mathcal{S}^{d-1}} \sup_{\xi \in \Xi_n} \left( \mathbb{E}[g(W_i, q, \xi(q)) - g(W_i, q, \xi_0(q))]^2 \right)^{1/2} \leq r'_n,$$

the application of Lemma 6.2 of Chernozhukov et al. (2017a) gives with  $M := \max_{i \in I_k^c} F_{\hat{\xi}, \xi_0}(W_i)$

$$\begin{aligned} \sup_{q \in \mathcal{S}^{d-1}} |ii(q)| &\leq \sup_{q \in \mathcal{S}^{d-1}} |\mathbb{G}_{n,k}[g(W_i, q, \hat{\xi}(q)) - g(W_i, q, \xi_0(q))]| \\ &\leq \sqrt{v(r'_n)^2 \log(a \|F_{\hat{\xi}, \xi_0}\|_{P,2}/r'_n) + v \|M\|_{P,c'} / \sqrt{n} \log(a \|F_{\hat{\xi}, \xi_0}\|_{P,2}/r'_n)} \\ &\lesssim_P r'_n \log^{1/2}(1/r'_n) + n^{-1/2+1/c'} \log^{1/2}(1/r'_n) \end{aligned}$$

where a constant  $\|M\|_{P,c'} \leq n^{1/c'} \|F\|_{P,c'}$  for the constant  $c' \geq 2$  in Assumption 6.

**Step 3. Asymptotic Normality.** By Theorem 19.14 from van der Vaart (1998), Assumption 6 implies that the function class  $\mathcal{F}_{\xi_0} = \{g(W, q, \xi_0(q)), q \in \mathcal{S}^{d-1}\}$  is  $P$ -Donsker. Therefore, the asymptotic representation follows from the Skorohod-Dudley-Whichura representation, assuming the space  $L^\infty(\mathcal{S}^{d-1})$  is rich enough to support this representation.  $\square$

*Proof of Theorem 1,  $\Sigma$  is unknown.* Step 1.  $\sqrt{n}$ -Convergence of Matrix Estimator. Let us show that there exists  $\phi_N = o(1)$  and a constant  $R$  such that with probability at least  $1 - \phi_N$ ,

$$\|\hat{\Sigma} - \Sigma\| \leq RN^{-1/2},$$

where

$$\hat{\Sigma} := \mathbb{E}_N A(W_i, \hat{\eta}_i) = \frac{1}{K} \sum_{k=1}^K \underbrace{\mathbb{E}_{n,k} A(W_i, \hat{\eta}_i) - \mathbb{E}_{n,k} A(W_i, \eta_0)}_{I_{1,k}} + \mathbb{E}_N A(W_i, \eta_0) - \mathbb{E} A(W_i, \eta_0).$$

Recognize that the first and the second moments of  $\sqrt{N}I_{1,k}$  converge to zero conditionally on the

partition complement  $J_k^c$  and the event  $\mathcal{E}_N$ . The first moment is bounded as:

$$\begin{aligned} \sqrt{n} \|\mathbb{E} I_{1,k} | \mathcal{E}_N \cup J_k^c\| &:= \sqrt{n} \|\mathbb{E}[A(W_i, \hat{\eta}) - A(W_i, \eta_0)] | \mathcal{E}_N \cup J_k^c\| \\ &\leq \sup_{\eta \in \mathcal{T}_N} \sqrt{n} \|\mathbb{E}[A(W_i, \eta) - A(W_i, \eta_0)] | \mathcal{E}_N \cup J_k^c\| \\ &\leq \sqrt{n} \mu_n = o(1). \end{aligned}$$

By Assumption 6, the bound on the second moment  $\|I_{1,k}\|^2$  applies:

$$n \mathbb{E}[\|I_{1,k}\|^2 | \mathcal{E} \cup J_k^c] \leq \sup_{\eta \in \mathcal{T}_n} \mathbb{E} \|A(W_i, \eta) - A(W_i, \eta_0)\|^2 \leq \delta_n = o(1).$$

Applying the Markov inequality conditionally on  $J_k^c, \mathcal{E}_N$  yields:  $ii = o_P(\delta_n)$ . By Lemma 6.1 of Chernozhukov et al. (2017a), conditional convergence to zero implies unconditional convergence. Therefore, for each  $k \in [K]$ ,  $I_{1,k} = o_P(1)$ . Since the number of partitions  $K$  is finite,  $\frac{1}{K} \sum_{k=1}^K I_{1,k} = o_P(1)$ . The application of the Law of Large Numbers for Matrices to the term  $\mathbb{E}_N[A(W_i, \eta_0) - \Sigma]$  yields:  $\|\mathbb{E}_N A(W_i, \eta_0) - \Sigma\| = O_P(N^{-1/2})$ .

Step 2. Decomposition of the error. Fix a generic element of this set  $p \in \mathcal{P}$  and  $\xi \in \Xi_N$ .

$$\begin{aligned} \mathbb{E}_N g(W_i, p, \xi(p)) &= \mathbb{E}_N g(W_i, p_0(q), \xi_0(p_0(q))) \\ &\quad + \underbrace{\mathbb{E}_N [g(W_i, p, \xi(p)) - g(W_i, p_0(q), \xi_0(p_0(q)))]}_{I_3} \\ &\quad + \underbrace{\mathbb{E} [g(W_i, p, \xi(p)) - g(W_i, p_0(q), \xi_0(p_0(q)))]}_{I_4}. \end{aligned} \tag{7.1}$$

Consider the following expansion:

$$\begin{aligned} I_4 &= \mathbb{E}[g(W_i, p, \xi(p)) - g(W_i, p, \xi_0(p))] + \mathbb{E}[g(W_i, p, \xi_0(p)) - \mathbb{E}g(W_i, p_0(q), \xi_0(p_0(q)))] \\ &= \underbrace{\mathbb{E}[g(W_i, p, \xi(p)) - \mathbb{E}g(W_i, p, \xi_0(p))]}_{I_{4,1}} + J_0(p_0(q))[p - p_0(q)] + R(p, p_0(q)), \end{aligned}$$

where the gradient  $J_0(p_0(q)) = \nabla_{p_0(q)} \mathbb{E}g(W_i, p_0(q), \xi_0(p))$ . By Assumption 7, the remainder term  $R(p, p_0) = o(\|p - p_0(q)\|)$  uniformly over  $q \in \mathcal{S}^{d-1}$ . By Assumption 6 applied on  $\mathcal{P}$ , the term  $I_{4,1}$

in (7.1) is bounded as:

$$\begin{aligned}
\sqrt{n}I_{4,1} &= \sqrt{n}|\mathbb{E}[g(W_i, p, \xi(p)) - g(W_i, p, \xi_0(p))]| \\
&\leq \sqrt{n} \sup_{p \in \mathcal{P}} \sup_{\xi \in \Xi_n} |\mathbb{E}[g(W_i, p, \xi(p)) - g(W_i, p, \xi_0(p))]| \\
&\leq \sqrt{n}r_n'' = o(1).
\end{aligned}$$

Step 4. The bound on  $I_3$ . We have

$$\mathcal{F}_{\xi \xi_0} := \{g(\cdot, p, \xi(p)) - g(\cdot, p_0(q), \xi(p_0(q))), p \in \mathcal{P}, q \in \mathcal{S}^{d-1}, \|p - p_0(q)\| \leq RN^{-1/2}\}.$$

We apply Lemma 6.2 of Chernozhukov et al. (2017a) conditionally on  $J_k^c$  and  $\mathcal{E}_N$ , so that  $\hat{\xi}$  can be treated as fixed. By Assumption 6, the class  $\mathcal{F}_{\hat{\xi} \xi_0}$  has a measurable envelope  $F_{\hat{\xi} \xi_0} := F_{\hat{\xi}} + F_{\xi_0}$ :

$$\begin{aligned}
&\sup_{q \in \mathcal{S}^{d-1}} \sup_{p \in \mathcal{P}, \|p - p_0(q)\| \leq RN^{-1/2}} |g(W_i, p, \xi(p)) - g(W_i, p_0(q), \xi(p_0(q)))| \\
&\leq \sup_{q \in \mathcal{S}^{d-1}} \sup_{p \in \mathcal{P}, \|p - p_0(q)\| \leq RN^{-1/2}} |g(W_i, p, \xi(p))| + \sup_{q \in \mathcal{S}^{d-1}} \sup_{p \in \mathcal{P}, \|p - p_0(q)\| \leq RN^{-1/2}} |g(W_i, p, \xi(p))| \\
&\leq F_{\hat{\xi}} + F_{\xi_0}.
\end{aligned}$$

Moreover, the uniform covering entropy of the function class  $F_{\hat{\xi} \xi_0} := F_{\hat{\xi}} + F_{\xi_0}$  is bounded by the sum of the entropies of  $F_{\hat{\xi}}$  and  $F_{\xi_0}$ . Finally,

$$\begin{aligned}
&\sup_{q \in \mathcal{S}^{d-1}} \sup_{p \in \mathcal{P}, \|p - p_0(q)\| \leq RN^{-1/2}} \sup_{\xi \in \Xi_n} (\mathbb{E}(g(W_i, p, \xi(p)) - g(W_i, p_0(q), \xi_0(p_0(q))))^2)^{1/2} \\
&\lesssim r_n' \log^{1/2}(1/r_n') + n^{-1/2+1/c} \log^{1/2} n.
\end{aligned}$$

□

**Lemma 6.** Let  $\{D_i(p), p \in \mathcal{P}\}$  be a function class with a bounded uniform covering entropy and  $\mathbb{E}D_i(p) = 0 \quad \forall p \in \mathcal{P}$ . Let  $(D_i(p))_{i=1}^N$  be an i.i.d sequence of random functions. Let  $(e_i)_{i=1}^N$  be an

i.i.d sequence of  $\text{Exp}(1)$  random variables independent from  $(D_i(p))_{i=1}^N$ . Then uniformly in  $\mathcal{P}$

$$\sqrt{N}\mathbb{E}_N \frac{e_i}{\bar{e}} D_i(p) = \sqrt{N}\mathbb{E}_N e_i D_i(p) + \underbrace{\sqrt{N}\mathbb{E}_N e_i D_i(p)}_{o_P(1)} \frac{1 - \bar{e}}{\bar{e}} = \sqrt{N}\mathbb{E}_N e_i D_i(p) (1 + o_P(1)).$$

**Lemma 7.** *Let Assumptions 5 - 6 hold. Then*

$$\sqrt{N}\mathbb{E}_N (g(W_i, (\tilde{\Sigma}^{-1})^\top q, \hat{\xi}((\tilde{\Sigma}^{-1})^\top q)) - g(W_i, (\tilde{\Sigma}^{-1})^\top q, \xi_0((\tilde{\Sigma}^{-1})^\top q))) \frac{e_i}{\bar{e}} = o_P(1).$$

*Proof.* Step 1. Decompose the sample average into the sample averages within each partition:

$$\begin{aligned} & \mathbb{E}_N (g(W_i, (\tilde{\Sigma}^{-1})^\top q, \hat{\xi}(p)) - g(W_i, (\tilde{\Sigma}^{-1})^\top q, \xi_0((\tilde{\Sigma}^{-1})^\top q))) \frac{e_i}{\bar{e}} \\ &= \frac{1}{K} \sum_{k=1}^K \mathbb{E}_{n,k} (g(W_i, (\tilde{\Sigma}^{-1})^\top q, \hat{\xi}(p)) - g(W_i, (\tilde{\Sigma}^{-1})^\top q, \xi_0((\tilde{\Sigma}^{-1})^\top q))) \frac{e_i}{\bar{e}}. \end{aligned}$$

Since the number of partitions  $K$  is finite, it suffices to show that the bound holds on every partition:

$$\mathbb{E}_{n,k} (g(W_i, (\tilde{\Sigma}^{-1})^\top q, \hat{\xi}(p)) - g(W_i, (\tilde{\Sigma}^{-1})^\top q, \xi_0((\tilde{\Sigma}^{-1})^\top q))) \frac{e_i}{\bar{e}} = o_P(1).$$

Let  $\mathcal{E}_N := \cap_{k=1}^K \{\hat{\xi}_k^c \in \Xi_n\}$ . By Assumption 6  $\Pr(\mathcal{E}_N) \geq 1 - K\phi_N = 1 - o(1)$ . The analysis below is conditionally on  $\mathcal{E}_N$  for some fixed element  $\hat{\xi}_k^c \in \Xi_n$ . Since the probability of  $\mathcal{E}_N$  approaches one, the statements continue to hold unconditionally, which follows from the Lemma 6.1 of Chernozhukov et al. (2017a).

Step 2. Consider the function class  $\mathcal{F}_{\xi\xi_0} := \{(g(W_i, p, \xi(p)) - g(W_i, p, \xi_0(p))), p \in \mathcal{P}\}$ . Consider the function class  $\mathcal{F}_{\xi\xi_0}^e := \{(g(W_i, p, \xi(p)) - g(W_i, p, \xi_0(p)))e_i, p \in \mathcal{P}\}$ . The function class is obtained by the multiplication of a random element of class  $\mathcal{F}_{\xi\xi_0}$  by an integrable random variable  $e_i$ . Therefore,  $\mathcal{F}_{\xi\xi_0}^e$  is also  $P$ -Donsker and has bounded uniform covering entropy. The expectation of the random element of the class  $\mathcal{F}_{\xi\xi_0}^e$  is bounded as:

$$\begin{aligned} \sqrt{n} \sup_{p \in \mathcal{P}} |\mathbb{E}[g(W_i, p, \hat{\xi}(p)) - g(W_i, p, \xi_0(p)) | \mathcal{E}_N]| &\lesssim \sup_{\xi \in \Xi_n} \mathbb{E}[g(W_i, p, \xi(p)) - g(W_i, p, \xi_0(p))] \\ &\lesssim \sqrt{n}\mu_n = o(1). \end{aligned}$$

The variance of each element of the class  $\mathcal{F}_{\xi, \xi_0}^e$  is bounded as:

$$\begin{aligned} \sup_{p \in \mathcal{P}} \sup_{\xi \in \Xi_n} \mathbb{E}((g(W_i, p, \xi(p)) - g(W_i, p, \xi_0(p)))^0 e_i)^2 &= \sup_{p \in \mathcal{P}} \sup_{\xi \in \Xi_n} \mathbb{E}((g(W_i, p, \xi(p)) - g(W_i, p, \xi_0(p)))^0)^2 \mathbb{E}e_i^2 \\ &\leq 2 \sup_{p \in \mathcal{P}} \sup_{\xi \in \Xi_n} \mathbb{E}((g(W_i, p, \xi(p)) - g(W_i, p, \xi_0(p))))^2 \\ &\lesssim r_n'', \end{aligned}$$

where the bound follows from the conditional independence of  $e_i$  from  $W_i$ ,  $\mathbb{E}e_i^2 = 2$  for  $e_i \sim \text{Exp}(1)$ , and Assumption 6. By Lemma 6

$$\mathbb{E}[\sup_{p \in \mathcal{P}} \mathbb{G}_{n,k}[(g(W_i, p, \xi(p)) - g(W_i, p, \xi_0(p)))^0] \frac{e_i(1-\bar{e})}{\bar{e}} | \mathcal{E}_N] = o_P(1). \text{Q.E.D.}$$

□

**Lemma 8.** *Let Assumptions 5 - 6 hold. Consider the function class  $\mathcal{F}_{pp_0} = \{g(W_i, p, \xi_0(p)) - g(W_i, p_0, \xi_0(p_0))\}$ ,  $\|p - p_0(q)\| \leq N^{-1/2}$ ,  $q \in \mathcal{S}^{d-1}$ ,  $p \in \mathcal{P}\}$ . Then uniformly on  $\mathcal{F}_{pp_0}$ ,*

$$\begin{aligned} &\sqrt{N} \mathbb{E}_N(g(W_i, (\tilde{\Sigma}^{-1})^\top q, \xi_0((\tilde{\Sigma}^{-1})^\top q)) - g(W_i, (\Sigma^{-1})^\top q, \xi_0((\Sigma^{-1})^\top q))) \frac{e_i}{\bar{e}} \\ &= \sqrt{N} q' \Sigma^{-1} (\tilde{\Sigma} - \Sigma) \Sigma^{-1} G(\Sigma^{-1} q) + o_P(1), \end{aligned}$$

where the gradient  $G(p)$  is defined in Assumption 7.

*Proof.* Consider the function class  $\mathcal{F}_{pp_0}^e = \{e_i(g(W_i, p, \xi_0(p)) - g(W_i, p_0, \xi_0(p_0)))\}$ ,  $\|p - p_0(q)\| \leq N^{-1/2}$ ,  $q \in \mathcal{S}^{d-1}$ ,  $p \in \mathcal{P}\}$ . The function class is obtained by the multiplication of a random element of class  $\mathcal{F}_{pp_0}$  by an integrable random variable  $e_i$ . Therefore,  $\mathcal{F}_{pp_0}^e$  is also  $P$ -Donsker and has bounded uniform covering entropy. The expectation of the random element is as decomposed as:

$$\mathbb{E}[(g(W_i, p, \xi_0(p)) - g(W_i, p_0, \xi_0(p_0)))] = (p - p_0) \cdot G(p_0) + R(p, p_0).$$

By Assumption 7, the remainder term  $R(p, p_0) = o(\|p - p_0(q)\|)$  uniformly over  $q \in \mathcal{S}^{d-1}$ . The

variance of each element of the class  $\mathcal{F}_{pp_0}^e$  is bounded as:

$$\mathbb{E}[(g(W_i, p, \xi_0(p)) - g(W_i, p_0, \xi_0(p_0)))^0 e_i^2]^2 \leq 2\mathbb{E}[g(W_i, p, \xi_0(p)) - g(W_i, p_0, \xi_0(p_0))]^2 \lesssim r'_n,$$

where the bound follows from the conditional independence of  $e_i$  from  $W_i$ ,  $\mathbb{E}(e_i)^2 = 2$  for  $e_i \sim \text{Exp}(1)$ , and Assumption 6. By Lemma 6

$$\mathbb{E}[\sup_{p \in \mathcal{P}} \mathbb{G}_{n,k}[(g(W_i, p, \xi_0(p)) - g(W_i, p_0, \xi_0(p_0)))^0] \frac{e_i(1 - \bar{e})}{\bar{e}} | \mathcal{E}_N] = o_P(1). \text{Q.E.D.}$$

□

*Proof of Theorem 2.* The difference between the bootstrap and the true support function as follows:

$$\begin{aligned} \sqrt{N}(\tilde{\sigma}(q, \mathcal{B}) - \sigma(q, \mathcal{B})) &= \underbrace{\frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{e_i}{\bar{e}} (g(W_i, (\tilde{\Sigma}^{-1})^\top q, \hat{\xi}((\tilde{\Sigma}^{-1})^\top q)) - g(W_i, (\tilde{\Sigma}^{-1})^\top q, \xi_0((\tilde{\Sigma}^{-1})^\top q)))}_{K_{\tilde{\xi}\tilde{\xi}_0}(q)} \\ &+ \underbrace{\frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{e_i}{\bar{e}} (g(W_i, (\tilde{\Sigma}^{-1})^\top q, \xi_0((\tilde{\Sigma}^{-1})^\top q)) - g(W_i, (\Sigma^{-1})^\top q, \xi_0((\Sigma^{-1})^\top q)))}_{K_{pp_0}(q)} \\ &+ \underbrace{\frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{e_i}{\bar{e}} (g(W_i, (\Sigma^{-1})^\top q, \xi_0((\Sigma^{-1})^\top q)) - \sigma(q, \mathcal{B}))}_{K_e} \end{aligned}$$

By Lemma 7  $\sup_{q \in \mathcal{S}^{d-1}} |K_{\tilde{\xi}\tilde{\xi}_0}| = o_P(1)$ . By Lemma 8  $K_{pp_0}(q) = -q^\top \Sigma^{-1} (\mathbb{E}_N e_i A(W_i, \eta_0) - \Sigma) \Sigma^{-1} G((\Sigma^{-1})^\top q) + o_P(1)$  uniformly in  $q$ . By Lemma 6,

$$\begin{aligned} K_e &= \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{e_i}{\bar{e}} (g(W_i, (\Sigma^{-1})^\top q, \xi_0((\Sigma^{-1})^\top q)) - \sigma(q, \mathcal{B})) \\ &=^i \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{e_i}{\bar{e}} (g(W_i, (\Sigma^{-1})^\top q, \xi_0((\Sigma^{-1})^\top q)))^0 \\ &=^{ii} \frac{1}{\sqrt{N}} \sum_{i=1}^N e_i (g(W_i, (\Sigma^{-1})^\top q, \xi_0((\Sigma^{-1})^\top q)))^0 + o_P(1), \end{aligned}$$

where  $i$  follows from (3.7) and  $ii$  from Lemma 6. A similar argument applies to the leading term

of  $K_{pp_0}$

$$q'\Sigma^{-1}(\mathbb{E}_N \frac{e_i}{e} A(W_i, \eta_0) - \Sigma)\Sigma^{-1}G((\Sigma^{-1})^\top q) = q'\Sigma^{-1}(\mathbb{E}_N e_i A(W_i, \eta_0) - \Sigma)\Sigma^{-1}G((\Sigma^{-1})^\top q) + o_P(1).$$

The first statement of Theorem 2 follows from:

$$\begin{aligned} \tilde{\mathcal{S}}_N(q) &=^i \sqrt{N}(\tilde{\sigma}(q, \mathcal{B}) - \hat{\sigma}(q, \mathcal{B})) \\ &= \sqrt{N}(\tilde{\sigma}(q, \mathcal{B}) - \sigma(q, \mathcal{B}) - (\hat{\sigma}(q, \mathcal{B}) - \sigma(q, \mathcal{B}))) \\ &=^{ii} \frac{1}{\sqrt{N}} \sum_{i=1}^N e_i h(W_i, q) + o_P(1) \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^N e_i^0 (h(W_i, q))^0, \end{aligned}$$

where  $i$  is by definition of the bootstrap support function process and  $ii$  is by definition of  $h(W, q)$  in Theorem 1. Once the asymptotic approximation conditional on the data has been established all further statements follows from Steps 2 and 3 of the Proof of Theorem 3 Chandrasekhar et al. (2011). □

### 7.1.2 Proof of Theorem 5

The following Lemma is useful for proving Theorem 5.

**Lemma 9** (Indicator Function). *Suppose the following statements hold. (1) There exists a bound  $B_{UL} < \infty$  such that the interval width is absolutely surely bounded for all nuisance parameter values  $\eta \in \mathcal{T}_N$   $\sup_{\eta \in \mathcal{T}_N} Y_{U, \eta} - Y_{L, \eta} \leq B_{UL} < \infty$ . (2) A collection of distributions of  $\{p^\top V_{\eta_0}, p \in \mathcal{P}\}$  is uniformly continuous on  $\mathcal{P}$*

$$\sup_{p \in \mathcal{P}} \sup_{\eta \in \mathcal{T}_N} \mathbb{E} |p^\top V_\eta - p^\top V_{\eta_0}| 1_{\{0 < |p^\top V_{\eta_0}| \leq |p^\top V_\eta - p^\top V_{\eta_0}|\}} \lesssim \mathbb{E} \|V_\eta - V_{\eta_0}\|^2.$$

(3) *The following convergence bound applies*

$$\sup_{\eta \in \mathcal{T}_N} (\mathbb{E} \|V_\eta - V_{\eta_0}\|^2)^{1/2} \lesssim \sup_{\eta \in \mathcal{T}_N} (\mathbb{E} \|\eta - \eta_0\|^2)^{1/2}.$$

(4) The nuisance parameter  $\eta$  is estimated at least at  $o(N^{-1/4})$  rate

$$\sup_{\eta \in \mathcal{T}_N} (\mathbb{E} \|\eta - \eta_0\|^2)^{1/2} = o(N^{-1/4}).$$

Then replacing (3.6) by its smooth analog produces a bias of order  $o(N^{-1/2})$

$$\sup_{p \in \mathcal{P}} \sup_{\eta \in \mathcal{T}_N} |\mathbb{E}[m(W, p, \eta) - m_0(W, p, \eta)]| = o(N^{-1/2}).$$

Proof is left to Supplementary Appendix.

*Proof of Theorem 5.* Step 1. Verification of Assumption 5(1). Consider the original moment function:

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

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

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

Second, we apply Lemma 11 to the moment function  $m_0(W, p, \eta)$  and derive the bias correction term

$$\alpha_0(W, p, \eta) = -\gamma(p, X)[D - \eta(X)],$$

where for each  $p \in \mathcal{P}$  the function  $\gamma(p, \cdot)$  is a  $P$ -square integrable function. The true value of the nuisance parameter  $\xi(p, X) = \{\eta(X), \gamma(p, X)\}$  is:

$$\xi_0(p, X) = \{\mathbb{E}[D|X], \mathbb{E}[(Y_U - Y_L) \mathbf{1}_{\{p^\top (D - \eta_0(X)) > 0\}} | X]\}.$$

Therefore,

$$g(W, p, \xi(p)) = p^\top (D - \eta(X)) [\Gamma(Y_L, Y_U - Y_L, p^\top (D - \eta(X))) - \gamma(p, X)]$$

has zero Gateaux derivative with respect to  $\xi(p)$ . Since the function  $m_0(W, p, \eta)$  is linear in  $\eta$ , its second derivative w.r.t  $\eta$  is equal to zero. Therefore, Assumption 6 holds. Assumption 17 implies

that Assumption 5 is satisfied. Step 2. Verification of Lemma 9. Assumption 1 of Lemma 9 follows from Assumption 8 (3). Assumptions 3 and 4 follows from the Assumption 8 (6). Assumption 2 is verified below:

$$\begin{aligned}
& \sup_{p \in \mathcal{P}} \sup_{\eta \in \mathcal{T}_N} \mathbb{E} |p^\top V_\eta - p^\top V_{\eta_0}| \mathbf{1}_{0 < |p^\top V_{\eta_0}| < |p^\top V_\eta - p^\top V_{\eta_0}|} \\
&= \sup_{p \in \mathcal{P}} \sup_{\eta \in \mathcal{T}_N} \mathbb{E} |p^\top (\eta(X) - \eta_0(X))| \mathbf{1}_{0 < |p^\top (D - \eta_0(X))| < |p^\top (\eta(X) - \eta_0(X))|} \\
&\leq \mathbb{E}_X |p^\top (\eta(X) - \eta_0(X))| \int_0^{|p^\top (\eta(X) - \eta_0(X))|} \rho_{p^\top (D - \eta_0(X))|X}(t, X) dt && (\mathbb{E}[\cdot] = \mathbb{E}_X[\cdot] \mathbb{E}_X[\cdot|X]) \\
&\leq K_h \mathbb{E}_X (p^\top (\eta(X) - \eta_0(X)))^2 && (\text{Assumption 8 (4)}) \\
&\leq K_h \|p\|^2 \mathbb{E}_X \|\eta(X) - \eta_0(X)\|^2 && (\text{Cauchy Schwartz}) \\
&\lesssim o(N^{-1/2}). && (\text{Assumption 8 (6)})
\end{aligned}$$

Step 3. Verification of Assumption 5(2). The moment condition for  $\Sigma$  is insensitive to the biased estimation of  $\eta$ :

$$\partial_\eta \mathbb{E} A(W, \eta_0) = 2 \partial_\eta \mathbb{E} (D - \eta_0(X)) (\eta(X) - \eta_0(X))^\top = 0.$$

Step 4. Verification of Assumption 7. The moment function  $g(W, p, \xi(p))$  takes the form:

$$g(W, p, \xi(p)) = p^\top (D - \eta(X)) (\Gamma(Y_L, Y_U - Y_L, p^\top (D - \eta(X))) - \gamma(p, X)).$$

The expectation of  $g(W, p, \xi(p))$  evaluated at  $\xi_0(p) = \{\eta_0(X), \gamma_0(p, X)\}$  is equal to:

$$L(p) := \mathbb{E} g(W, p, \xi_0(p)) = p^\top \mathbb{E} (D - \eta_0(X)) (\Gamma(Y_L, Y_U - Y_L, p^\top (D - \eta_0(X))),$$

since  $\mathbb{E} \gamma_0(p, X) (D - \eta_0(X)) = 0$ . According to Lemma 3 of Chandrasekhar et al. (2011), the function  $L(p)$  is differentiable on  $\mathcal{P}$  with a uniformly continuous derivative:

$$\nabla L(p) := \mathbb{E} (D - \eta_0(X)) (\Gamma(Y_L, Y_U - Y_L, p^\top (D - \eta_0(X))) =: G(p).$$

Moreover, the gradient  $G(p)$  is uniformly continuous on  $\mathcal{P}$ . Verification of (3.13). The intermediate

Value Theorem implies:

$$L(p) - L(p_0) = \nabla L(p'_0)(p - p_0) = G(p'_0)(p - p_0),$$

where  $p'_0$  is a point on the interval  $[p_0, p]$ . Therefore,

$$\begin{aligned} G(p'_0)(p - p_0) &= G(p_0)(p - p_0) + (G(p'_0) - G(p_0))(p - p_0) \\ &= G(p_0)(p - p_0) + o(\|p - p_0\|), \end{aligned}$$

where the last equality follows from the uniform continuity of  $G(p)$  that is established in Lemma 3 of Chandrasekhar et al. (2011).

Step 5. Verification of Assumption 6. Consider the setting of Lemma 19. Consider the function class  $\mathcal{A}_\xi = \{p^\top(D - \eta(X))\gamma(p, X), p \in \mathcal{P}\}$ . Consider an envelope function  $F_g(X) = \|p\| \|D - \eta(X)\| B_{UL}$ . By Assumption 10, this function is integrable. Since the class is obtained by multiplying a linear function  $p \rightarrow p^\top(D - \eta(X))$  by a Lipschitz function  $\gamma(p, X)$ , its uniform covering entropy is bounded. The application of Lemma 19 with the function class  $\mathcal{A}_\xi$  implies that Assumptions 6(4) hold.

Step 6. Verification of Assumption 6. We use the following notation:  $Y_{p,\eta} := \Gamma(Y_L, Y_U - Y_L, p^\top(D - \eta(X)))$ ,  $p \in \mathcal{P}, \eta \in \mathcal{T}_N$ . Let us decompose the difference of  $g(W, p, \xi(p))$  and  $g(W, p, \xi_0(p))$

$$\begin{aligned} g(W, p, \xi(p)) - g(W, p, \xi_0(p)) &:= p^\top(D - \eta(X))(Y_{p,\eta} - \gamma(p, X)) - p^\top(D - \eta_0(X))(Y_{p,\eta_0} - \gamma_0(p, X)) \\ &= \underbrace{p^\top(\eta(X) - \eta_0(X))\gamma(p, X)}_{I_1} + \underbrace{p^\top(D - \eta_0(X))(\gamma_0(p, X) - \gamma(p, X))}_{I_2} \\ &\quad + \underbrace{p^\top(D - \eta_0(X))(Y_{p,\eta} - Y_{p,\eta_0})}_{I_3} + \underbrace{p^\top(\eta_0(X) - \eta(X))(Y_{p,\eta} - Y_{p,\eta_0})}_{I_4}. \end{aligned}$$

Under Assumptions 8, the terms  $I_k, k \in \{1, 2, 3, 4\}$  exhibit mean square convergence with the rate

$o(N^{-1/4})$ :

$$\begin{aligned}
(\mathbb{E}I_1^2)^{1/2} &\lesssim \lambda_{\min}^{-1} Y_{UL} \|\eta - \eta_0\|_{L_{P,2}} = o(N^{-1/4}), \\
(\mathbb{E}I_2^2)^{1/2} &\lesssim \lambda_{\min}^{-1} D \|\xi(p) - \xi_0(p)\|_{L_{P,2}} = o(N^{-1/4}), \\
(\mathbb{E}I_3^2)^{1/2} &\lesssim \mathbb{E}(p^\top (D - \eta_0(X)))^2 \mathbf{1}_{0 < |p^\top (D - \eta_0(X))| < |p^\top (\eta(X) - \eta_0(X))|} B_{UL} \\
&\leq (\mathbb{E}|p^\top (\eta(X) - \eta_0(X))|^2)^{1/2} = o(N^{-1/4}), \\
(\mathbb{E}I_4^2)^{1/2} &\lesssim \lambda_{\min}^{-1} 2Y_{UL} \|\eta - \eta_0\|_{L_{P,2}}.
\end{aligned}$$

Under Assumption 8, the terms  $S_k, k \in \{1, 2, 3, 4\}$  exhibit mean square convergence:

$$\begin{aligned}
g(W, p, \xi_0(p)) - g(W, p_0, \xi_0(p_0)) &:= p^\top (D - \eta_0(X))(Y_{p, \eta_0} - \gamma_0(p, X)) - p_0^\top (D - \eta_0(X))(Y_{p_0, \eta_0} - \gamma_0(p_0, X)) \\
&= - \underbrace{p_0^\top (D - \eta_0(X))(\gamma_0(p, X) - \gamma_0(p_0, X))}_{S_1} - \underbrace{(p - p_0)^\top (D - \eta_0(X))\gamma_0(p, X)}_{S_2} \\
&\quad + \underbrace{p_0^\top (D - \eta_0(X))(Y_{p, \eta_0} - Y_{p_0, \eta_0})}_{S_3} + \underbrace{(p - p_0)^\top (D - \eta_0(X))Y_{p, \eta_0}}_{S_4}.
\end{aligned}$$

By Assumption 8(9),

$$(\mathbb{E}S_1^2)^{1/2} \lesssim g'_N : \quad g'_N \log(1/g'_N) = o(1)$$

holds. The bound on the mean square convergence of the other terms is as follows:

$$\begin{aligned}
(\mathbb{E}S_2^2)^{1/2} &\lesssim \|p - p_0\| D Y_{UL} = O(N^{-1/2}), \\
(\mathbb{E}S_3^2)^{1/2} &\leq (\mathbb{E}[p_0^\top (D - \eta_0(X))]^2 \mathbf{1}_{0 < |p^\top (D - \eta_0(X))| < |(p - p_0)^\top (D - \eta_0(X))|})^{1/2} \\
&\leq \|p - p_0\| D = O(N^{-1/2}), \\
(\mathbb{E}S_4^2)^{1/2} &\lesssim \|p - p_0\| D Y_{UL} = O(N^{-1/2}).
\end{aligned}$$

□

## References

- Abdulkadiroglu, A., Pathak, P., and Walters, C. (2018). Free to choose: Can school choice reduce student achievement? *American Economic Journal: Applied Economics*, 10(1):175–206.
- Angrist, J., Bettinger, E., , and Kremer, M. (2006). Long-term educational consequences of secondary school vouchers: Evidence from administrative records in colombia. *The American Economic Review*, 96(3):847–862.
- Angrist, J., Bettinger, E., Bloom, E., King, E., and Kremer, M. (2002). Vouchers for private schooling in colombia: Evidence from a randomized natural experiment. *The American Economic Review*, 92(5):1535–1558.
- Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. (2013). Program evaluation and causal inference with high-dimensional data. *arXiv preprint arXiv:1311.2645*.
- Belloni, A., Chernozhukov, V., and Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4):606–619.
- Beresteanu, A. and Molinari, F. (2008). Asymptotic properties for a class of partially identified models. *Econometrica*, 76(4):763–814.
- Beresteanu, A., Molinari, F., and Molchanov, I. (2011). Sharp identification regions in models with convex predictions. *Econometrica*, 79(6).
- Bontemps, C., Magnac, T., and Maurin, E. (2012). Set identified linear models. *Econometrica*.
- Chandrasekhar, A., Chernozhukov, V., Molinari, F., and Schrimpf, P. (2011). Inference for best linear approximations to set identified functions. *Discussion Paper, University of British Columbia*.
- Chen, X., Tamer, E., and Torgovitsky, A. (2011). Sensitivity analysis in semiparametric likelihood models. *COWLES FOUNDATION DISCUSSION PAPER*.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2017a). Double/debiased machine learning for treatment and causal parameters.

- Chernozhukov, V., Escanciano, J. C., Ichimura, H., and Newey, W. (2017b). Locally robust semi-parametric estimation.
- Chernozhukov, V., Hansen, C., and Spindler, M. (2018). High-dimensional metrics.
- Chernozhukov, V., Hong, H., and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75(5):1243–1284.
- Ciliberto, F. and Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6):1791–1828.
- Engberg, J., Epple, D., Imbrogno, J., Sieg, H., and Zimmer, R. (2014). Evaluating education programs that have lotteried admission and selective attrition. *Journal of Labor Economics*, 32(1).
- Hardle, W. and Stoker, T. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of American Statistical Association*, 84(408):986–995.
- Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.
- Huber, M., Laffers, L., and Mellace, G. (2017). Sharp iv bounds on average treatment effects on the treated and other populations under endogeneity and noncompliance. *Journal of Applied Econometrics*, 32:56–79.
- Ichimura, H. and Newey, W. (2017). The influence function of semiparametric estimators. <https://economics.mit.edu/files/10669>.
- Imbens, G. and Manski, C. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.
- Kaido, H. (2016). A dual approach to inference for partially identified econometric models. *Journal of Econometrics*, 192(1):269–290.
- Kaido, H. (2017). Asymptotically efficient estimation of weighted average derivatives with an interval censored variable.

- Kaido, H. and Santos, A. (2014). Asymptotically efficient estimation of models defined by convex moment inequalities. *Econometrica*, 82(1):387–413.
- Kaido, H. and White, H. (2014). A two-stage procedure for partially identified models. *Journal of Econometrics*, 1(182):5–13.
- Lee, D. (2008). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102.
- Manski, C. (1989). The anatomy of the selection problem. *Journal of Human Resources*, 24(3):343–360.
- Manski, C. and Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica*, 70(2):519–546.
- Molinari, F. and Molchanov, I. (2018). *Random Sets in Econometrics*. Cambridge University Press.
- Newey, W. (1994). The asymptotic variance of semiparametric estimators. 62:245–271.
- Newey, W. and Stoker, T. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica*, 61(5):1199–1223.
- Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics*, 213(57).
- Neyman, J. (1979).  $c(\alpha)$  tests and their use. *Sankhya*, pages 1–21.
- Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25:303–325.
- Robins, J. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of American Statistical Association*, 90(429):122–129.
- Robins, J., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association*, 89(427):846–866.

Robinson, P. M. (1988). Root- $n$  consistent semiparametric regression.

Semenova, V. (2018). Machine learning for dynamic discrete choice and other moment inequalities.

Sieg, H. and Wang, Y. (2018). The impact of student debt on education, career, and marriage choices of female lawyers. *European Economic Review*.

Taddy, M. (2011). One-step estimator paths for concave regularization. <https://arxiv.org/abs/1308.5623>.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, (2):167–95.

van der Vaart, A. (1998). Asymptotic statistics.

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 11 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  $\beta_U$  as a solution to a semiparametric moment equation. Inspecting (2.11), one can see that the identified set (2.11) 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  $\beta_U$ , or, equivalently, the largest possible least squares coefficient, I construct a random variable  $Y^{\text{UBG}}$  as

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

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<sup>4</sup>. As a result, the upper bound is characterized by the semiparametric moment equation

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

(see, e.g. Beresteanu and Molinari (2008) or Bontemps et al. (2012)). The major difficulty when estimating  $\beta_U$  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  $\beta_U$  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  $n = \lfloor N/2 \rfloor$  such that  $J_1 \cup J_2 = \{1, 2, \dots, N\}$ . 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  $D_i - \hat{\eta}(X_i)$

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

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

$$\sqrt{N} |\hat{\beta}_U^{\text{NAIVE}} - \beta_U| \rightarrow \infty \quad (8.3)$$

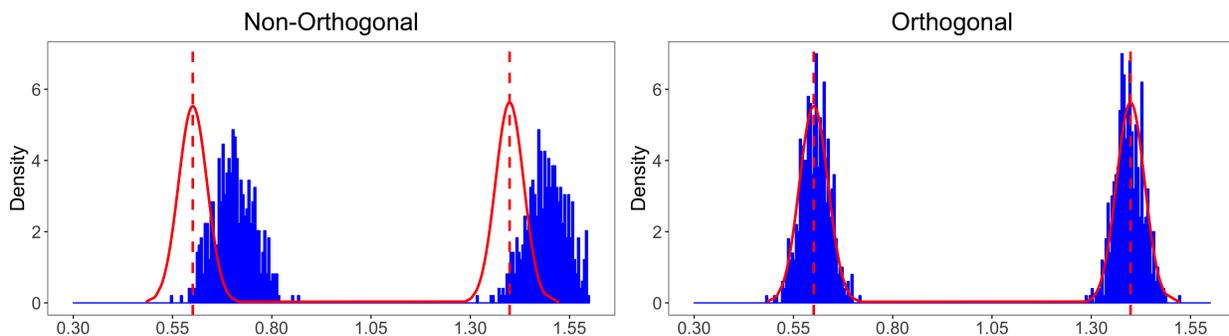
and cannot be used to conduct inference about  $\beta_U$  using standard Gaussian approximation. The behavior of the naive estimator is shown in Figure 8(a).

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-

---

<sup>4</sup>In what follows, I assume that the residual  $V$  has a continuous distribution and is equal to zero with probability zero.

Figure 8: Finite-sample distribution of non-orthogonal (naive) and orthogonal estimates of the bounds



**Figure 8** shows the finite-sample distribution (blue histogram) of naive (left panel) and orthogonal (right panel) estimates of the lower ( $\beta_L$ ) and the upper ( $\beta_U$ ) 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 be 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  $\eta_0$ . Assume that I can somehow generate the true value of the upper bound generator  $\gamma_0 = Y^{\text{UBG}}(\eta_0)$ . Consider a smooth moment function

$$m_0(W, \beta_U, \eta_0) = (\gamma_0 - (D - \eta_0(X))\beta_U) \cdot (D - \eta_0(X)) \quad (8.4)$$

Then the difference between the infeasible moment equation  $m_0(W, \beta_U, \eta_0)$ , based on the true value of the nuisance parameter  $\eta_0$ , 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  $Y^{\text{UBG}}$ , 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 8(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

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

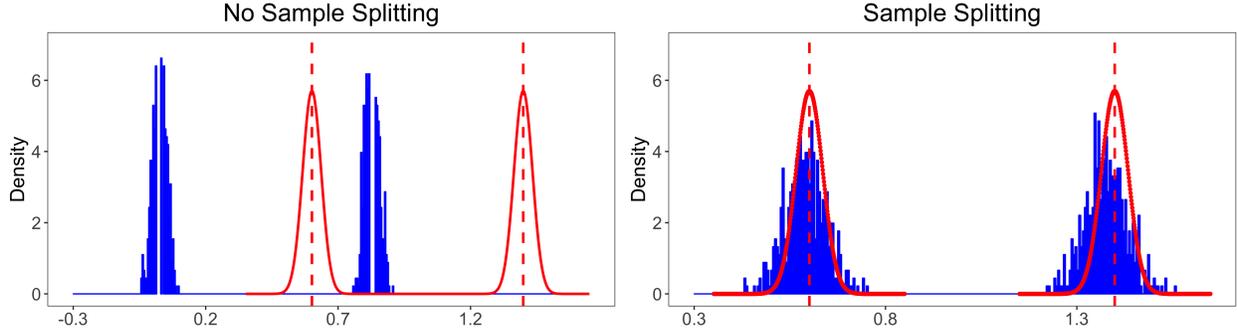
Return:  $\hat{\beta}_U$ .

---

**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 9 (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 9: Finite-sample distribution of the orthogonal estimator without and with sample splitting



**Figure 9** 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 be 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  $\Omega$  is a covariance matrix. The confidence region of level  $\alpha \in (0, 1)$  for the identified set  $[\beta_L, \beta_U]$  takes the

form

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

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  $\Omega$  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 (Härdle 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  $\eta_0$ , 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

$$g(W, p, \xi(p)) = m(W, p, \eta) + \alpha(W, p, \xi(p)) \quad (9.1)$$

obeys the Neyman-orthogonality condition (Assumption 5).

Lemma 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). It generalizes the result of ? from the point to set-identified case. 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  $\eta$  satisfies the generalized conditional exogeneity restriction*

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

where  $R(W, \eta) : \mathcal{W} \times \mathcal{T} \rightarrow \mathcal{R}^L$  is a known measurable map that maps a data vector  $W$  and a square-integrable vector-function  $\eta$  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

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

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  $\eta$  conditionally on  $X$ . Furthermore, the function  $g(W, p, \xi(p))$  in

(11.10) 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  $\xi$  at  $\xi_0$  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  $\xi$  at  $\xi_0$  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 (11.9).

**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  $l \in \{1, 2, \dots, L\}$  is defined by a separate exclusion restriction:  $\mathbb{E}[R_l(W, \eta_{l,0}(X)) | X] = 0, l \in \{1, 2, \dots, L\}$ . Then, the bias correction*

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

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

where each term  $\alpha_l(W, p, \xi_l(p))$  corrects for the estimation of  $\eta_l, l \in \{1, 2, \dots, L\}$  holding the other components  $\eta_{-l}$  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  $\eta$  be an  $L$ -dimensional vector. First, for each  $l \in \{1, 2, \dots, L\}$  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  $\eta$  appears inside the weighting variable  $V$  defined in (2.2). As a result, the moment equation (3.4) depends on  $\eta$  in a non-smooth way. In particular,  $V = V_\eta$  appears inside a function  $x \rightarrow x1_{x>0}$  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 Empirical Application to Angrist et al. (2006)

In this section, I re-examine the effectiveness of Colombia PACES program, a voucher initiative established in 1991 to subsidize private school education in low-income population, studied in Angrist et al. (2002) and in Angrist et al. (2006). After being admitted to a private school, a student participates in a lottery to win a voucher that partially covers his tuition fee. Each year, a student can renew an existing voucher if he passes to the next grade. After high school graduation, some students take a centralized test to enter a college. Following Angrist et al. (2006), I am interested in the average effect of winning the private school voucher today on the college admission test scores several years later.

I use the notation of Example 1 to define the voucher's effect. The variable  $D = 1$  is a dummy for whether a student has won a voucher,  $S_0 = 1$  is a dummy for whether a student would have participated in a test after losing the voucher,  $S_1 = 1$  is a dummy for whether a student would have participated in a test after winning the voucher. Similarly, the potential test scores  $Y_0$  and  $Y_1$  are the scores a student would have had after losing and winning the lottery, respectively. I am interested in the average voucher's effect on the group of students who would have taken the test regardless of receiving the voucher

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

or, briefly, the always-takers. The data contain the voucher status  $D$ , observed test participation<sup>5</sup>  $S$ , test score  $S \cdot Y$  observed only if a student takes a test, and the covariates. The covariates  $X$  include age, phone access, gender, and four indicators of having an invalid or inaccurately recorded ID constructed by Angrist et al. (2006) by matching PACES records to administrative data.

Because test participation may be endogenous, the average voucher effect is not point-identified. To bound the effect, Angrist et al. (2006) make two assumptions: receiving a voucher can neither

---

<sup>5</sup>The test participation  $S$  is not explicitly recorded in the data. I conclude that a student comes to a test if and only if his test score is positive  $S = 1_{\{Y>0\}}$ . My conclusion is based on two facts. For a given subject, Angrist et al. (2006) interprets the subset of voucher losers with positive test scores as the always-takers (page 14). To arrive at this interpretation, one needs to assume that  $S = 1_{\{Y>0\}}$ . Second, the test scores have a 66% point mass at zero value for both subjects.

deter the test participation  $S_1 \geq S_0$  a.s. , nor hurt the test scores  $Y_1 \geq Y_0$  a.s. . Angrist et al. (2006) state that the monotonicity assumption about the test scores may not hold if private school applicants anticipated educational gains that did not materialize. To relax this assumption, I use Lee (2008)'s bounds that are based only on the first assumption. I describe the construction of Lee (2008)' bounds in Example 1.

I estimate Lee (2008)'s bounds with all covariates in two stages. In the first stage, I estimate the probability of receiving the voucher given covariates  $X$  (i.e,  $\Pr(D = 1)$ ), the probability of test participation given the voucher status  $D$  and covariates  $X$  (i.e,  $s(D, X) = \mathbb{E}[S = 1|D, X]$ ), and the quantile function of the winners' test scores given the covariates. I estimate the first two functions using logistic lasso algorithm of Belloni et al. (2016) with the penalty choice described in Chernozhukov et al. (2018) package. Assuming that winners' test scores are determined by age and gender only, I estimate the quantile function by taking an empirical quantile<sup>6</sup> in the relevant group. In the second stage, I plug the estimates into Neyman-orthogonal moment equations for the bounds given in (11.12).

My empirical findings are as follows. Lee (2008)'s bounds based on the choice of covariates, made by Angrist et al. (2006), have opposite signs and cannot determine the direction of the effect (Table 1, Column 2). An attempt to include all 7 covariates into Lee (2008)'s bounds runs into perfect multicollinearity problem. In particular, for certain cells determined by a subset of binary covariates, there is no variation in treatment assignment. Logistic lasso algorithm finds the covariate that predicts selection best among the rest. As a result, the bounds produced by my method (Table 1, Column 3) are positive, four times narrower than the bounds in Column 2, and significant in case of Language.

Bounds become narrow because Angrist et al. (2006) data set contains a very informative covariate whose power to predict test-taking decision have been previously overlooked. This covariate, referred to as ID validity, is a binary indicator for whether ID number recorded at the moment of lottery registration, can be matched with an administrative database (i.e., corresponds to a valid ID), and, therefore, is pre-determined. This covariate explains 96% of the total variance in test participation, while age and gender, that have been previously used, explain only 35%. Once ID

---

<sup>6</sup>Because the test scores' distribution had multiple point masses, I added a small amount of  $N(0, 0.01)$  distributed noise in order to compute the exact quantiles.

validity is taken into account, voucher has little effect on the test-taking decision, resulting in tighter bounds.

Table 1: Lee (2008)'s bounds on Voucher Effect on Test Scores using the data in Angrist et al. (2006)

Covariates	None (1)	{ Age, gender } (2)	<b>My result, all 7 covs</b> (3)
<i>A. Mathematics</i>			
Estimate	[-1.304, 2.073]	[-1.100, 1.827]	<b>[0.160, 0.904]</b>
95% CR	(-2.131, 2.886)	(-1.875, 2.599)	<b>(-0.168, 1.570)</b>
<i>B. Language</i>			
Estimate	[-1.192, 2.640]	[-0.946, 2.341]	<b>[0.473, 1.112]</b>
95% CR	(-2.086, 3.542)	(-1.8007, 3.211)	<b>(0.144, 1.847)</b>

Table 1 reports estimated bounds for the voucher effect (Estimates) and a 95% confidence region (95% CR) for the identified set for the voucher effect for test scores in Mathematics (Panel A) and Language (Panel B). I report the results for 3 specifications: without covariates (Column 1), with age and gender covariates (Column 2), and my result based on all 7 covariates (Column 3).

## 11 Proofs

### 11.1 Proof of Section 2

**Lemma 14** (Derivation of Equation (2.6)). *Suppose Assumption 1 holds. Then, the bounds given in Equation (2.6) coincide with the bounds given in Proposition 1b of Lee (2008).*

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

$$\begin{aligned}
& \int_{x \in \mathcal{X}} f(x|D=0, S=1) \mathbb{E}[Y|D=1, S=1, Y \leq y_{\{p_0(x), x\}}, X=x] \\
&= \int_{x \in \mathcal{X}} \frac{f(x|D=0, S=1)}{f(x|D=1, S=1)} \mathbb{E}[Y|D=1, S=1, Y \leq y_{\{p_0(x), x\}}, X=x] f(x|D=1, S=1) \\
&= \int_{x \in \mathcal{X}} \frac{f(x|D=0, S=1)}{f(x|D=1, S=1)} \frac{1}{p_0(x)} \mathbb{E}[Y 1_{\{Y \leq y_{\{p_0(x), x\}}\}} | D=1, S=1, X=x] f(x|D=1, S=1), \quad (11.1)
\end{aligned}$$

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

$$\frac{f(x|D=0, S=1)}{f(x|D=1, S=1)} = \frac{\Pr(X=x, D=0, S=1) \Pr(D=1, S=1)}{\Pr(D=0, S=1) \Pr(X=x, D=1, S=1)} \quad (11.2)$$

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

$$\begin{aligned} p_0(x) &= \frac{\Pr(S=1|D=0, X=x)}{\Pr(S=1|D=1, X=x)} \\ &= \frac{\Pr(S=1, D=0, X=x) \Pr(D=1|X=x)}{\Pr(S=1, D=1, X=x) \Pr(D=0|X=x)} = \frac{\Pr(S=1, D=0, X=x) \Pr(D=1)}{\Pr(S=1, D=1, X=x) \Pr(D=0)} \\ &\quad \text{(Independence } D \perp X) \end{aligned} \quad (11.3)$$

Plugging (11.2) and (11.3) into (11.1) gives (2.8)

$$\int_{x \in \mathcal{X}} \frac{f(x|D=0, S=1)}{f(x|D=1, S=1)} \frac{1}{p_0(x)} \mathbb{E}[Y 1_{\{Y \leq y_{\{p_0(x), x\}}\}} | D=1, S=1, X=x] f(x|D=1, S=1) \quad (11.4)$$

$$= \mathbb{E} \frac{D \cdot S \cdot Y \cdot 1_{\{Y \leq y_{\{p_0(x), x\}}\}} \Pr(D=0)}{\Pr(D=0, S=1) \Pr(D=1)} \quad (11.5)$$

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:*

$$\mathcal{B} = \{\beta = \arg \min_{b \in \mathcal{R}^d, f \in \mathcal{M}} \mathbb{E}(Y - D^\top b - f(X))^2, \quad Y_L \leq Y \leq Y_U\} \quad (11.6)$$

*coincides with the identified set given by (2.11).*

*Proof.* Fix a random variable  $Y$  in a random interval  $[Y_L, Y_U]$ . Let us show that the minimizer  $\beta_0$  of (11.6) coincides with the minimizer  $\beta_0^s$  defined as:

$$\beta_0^s = \arg \min_{b \in \mathcal{R}^d, f \in \mathcal{M}} \mathbb{E}(Y - (D - \eta_0(X))'b)^2, \quad (11.7)$$

where  $\eta_0(X) = \mathbb{E}[D|X]$ . For each  $b$  in (11.6) 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 (11.6) gives:

$$\beta = \arg \min_{b \in \mathcal{R}^d} \mathbb{E}(Y - \mathbb{E}[Y|X] - (D - \eta_0(X))'b)^2. \quad (11.8)$$

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

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

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  $\beta_0^s$  in (2.11). According to Bontemps et al. (2012) (Proposition 2), the set (2.11) is a sharp identified set for  $\beta_0$ .

□

## 11.2 Proofs of Section 9

Below we present examples of bias correction terms  $\alpha(W, p, \xi(p))$  for various types of functional nuisance parameter  $\eta$ . 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:

$$M(p) - \mathbb{E}m(W, p, \eta_0) = 0, \quad (11.9)$$

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

$$g(W, p, \xi(p)) = m(W, p, \eta) + \alpha(W, p, \xi(p)) \quad (11.10)$$

obeys orthogonality condition with respect to  $\xi$  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  $\eta$  at  $\eta_0$  is:

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

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 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

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

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 (11.10) 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  $\eta_l(X), l \in \{1, 2, \dots, L\}$  is defined by a separate exclusion restriction, the matrix  $I_0(X)$  is a diagonal matrix whose  $l$ 'th ele-

ment on the diagonal is equal to  $I_{l,0}(X) := \partial_{\eta_{l,0}} \mathbb{E}[R_l(W, \eta_{l,0}(X))|X]$ . 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  $\xi$  at  $\xi_0$  is equal to:

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

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  $\eta$  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  $\eta$  in the notation:

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

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 (11.9) on a compact set  $\mathcal{P}$ . Suppose Assumption 6 holds on  $\mathcal{P}$ . Moreover, the moment condition (11.10) has zero Gateaux derivative for all vectors  $p$  in  $\mathcal{P}$ . Then, the moment*

condition (11.10) satisfies Assumption 6 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  $r \rightarrow \mathbb{E}[g(W, p, r(\xi(p) - \xi_0(p)) + \xi_0(p))]$ ,  $r \in (0, 1)$ .

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

□

**Lemma 18** (Achieving Small Bias Assumption for Support Function). *Suppose the conditions of Lemma 9 hold. Let  $m_0(W, p, \eta)$  be a smoothed 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  $\xi$  at  $\xi_0$  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 5.*

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

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

where the bound on the first summand is by Lemma 9 and on the second one by Lemma 17.  $\square$

**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  $\mathcal{F}_\eta = \{p^\top V(\eta)\Gamma(Y_{L,\eta}, Y_{U,\eta} - Y_{L,\eta}, p^\top V(\eta)), \|p - p_0(q)\| \leq RN^{-1/2}, q \in \mathcal{S}^{d-1}\}$  satisfies Assumption 6(1,2,3(a)). If, in addition, the function class*

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

satisfies Assumption 6, so does the class  $\mathcal{R}_g = \{g(W, p, \xi(p)), \|p - p_0(q)\| \leq RN^{-1/2}, q \in \mathcal{S}^{d-1}\}$ .

*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

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

and

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

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

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

have bounded uniform covering entropy (UCE), respectively:

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

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

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  $\mathcal{L}_\eta \cdot \mathcal{J}_\eta$

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.  $\square$

### 11.3 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

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

The notation  $\eta_{-k,0}, k \in \{1, 2, 3\}$  stands for the true value of  $\eta_{-k}$  obtained from  $\eta$  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 (11.12) inside the quantile  $y_{\{1-s(0,X)/s(1,X), X\}}$  and in the denominator of (11.12). The Gateaux derivative of the function  $\mathbb{E}[m_U(W, \eta)|X]$  with respect to  $\eta_1$  at  $\eta_{1,0} = s(0, X)$  is equal to:

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

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 (11.12). The Gateaux derivative of the function  $\mathbb{E}[m_U(W, \eta)|X]$  with respect

to  $\eta_2$  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  $\eta_3(u, x) = \mathcal{Q}_{Y|D=1, S=1, X=x}(u, x)$  appears in the numerator of (11.12). 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:

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

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:

$$m_L(W, \eta) := \frac{D \cdot S \cdot Y 1_{\{Y \leq \eta_3(\eta_1/\eta_2, X)\}} \Pr(D=0)}{\Pr(D=0, S=1) \Pr(D=1)}. \quad (11.11)$$

The bias correction term  $\alpha_L(W, \eta)$  for  $\beta_L$  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

$$g_U(W, \xi_U) = m_U(W, \eta) + \alpha_U(W, \xi_U), \quad (11.12)$$

$$g_L(W, \xi_L) = m_L(W, \eta) + \alpha_L(W, \xi_L), \quad (11.13)$$

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

Step 5. To apply Lemma 17 and conclude that Assumption 5 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 6(3(b)). The first derivative  $\partial_\eta \mathbb{E}[m_U(W, \eta_0)|X]$  is a composition of the functions  $(t_1, t_2) \rightarrow t_1/t_2$ ,  $t \rightarrow tf(t)$ , where  $f(\cdot)$  is the conditional density of  $Y$  given  $D=1, S=1, X=x$ , and the functional elements of  $\Xi_N^U$ :  $u \rightarrow \eta_3(u, x)$ . By Assumption 8, each of these functions is bounded and has a bounded first derivative. Therefore, Assumption 6 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  $\theta_L, \theta_U$  on  $\theta$  are given by the following moment conditions:

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

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

**Remark 3** (Asymptotic Theory for the Average Treatment Effect in Endogenous Sample Selection of Lee (2008)). *Suppose Assumption 8 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:*

$$\sqrt{N} \begin{pmatrix} \hat{\theta}_L - \theta_L \\ \hat{\theta}_U - \theta_U \end{pmatrix} \Rightarrow N(0, \Omega), \quad (11.14)$$

where  $\Omega$  is a positive-definite covariance matrix.

*Proof of Theorem 4.* Step 1. Verification of Assumption 5(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 smoothed 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  $\xi_0$  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  $\xi$  at  $\xi_0$ . 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 6 holds. Assumption 17 implies that Assumption 5(1) is satisfied.

Step 2. Verification of the conditions of Lemma 9. Assumptions 1-4 of Lemma 9 follows from Assumptions 9 (1)-(3). Step 3. Verification of Assumption 6. 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  $\mathcal{J}_\eta$  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  $\mathcal{J}_\eta \cdot (\mu_U(D, X) - \mu_L(D, X))$  are summed into  $\mathcal{L}_\eta \cdot \mu_L(D, X) + \mathcal{J}_\eta \cdot (\mu_U(D, X) - \mu_L(D, X))$ .
3. The class  $\mathcal{L}_\eta \cdot \mu_L(D, X) + \mathcal{J}_\eta \cdot (\mu_U(D, X) - \mu_L(D, X))$  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 6.

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

By Assumption 9, 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:

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

This step completes the proof of Theorem 4. □

**ASSUMPTION 12** (Identification in Endogenous Sample Selection). *The following assumptions hold. (1) The program admission  $D$  is independent of the potential employment and wage outcomes, as well as the subject covariates:  $D \perp (S_1, S_0, Y_1, Y_0, X)$ . (2) The program admission  $D$  cannot help selection  $S_0 \geq S_1$  a.s.*

Under Assumption 12, the observed treated group is a random sample of always-takers.

Let  $s(D, X) = \Pr(S = 1 | D, X)$ ,  $y_{u, X}$  be the conditional quantile of  $Y$  in  $D = 0, S = 1$  group:

$$y_{\{u, X\}} : \Pr(Y \leq y_{\{u, X\}} | X, D = 1, S = 1) = u, \quad u \in [0, 1],$$

and  $p_0(X) = \frac{s(1, X)}{s(0, X)}$  be the trimming proportion. Suppose Assumption 12 holds. Then, the sharp bounds on ATE are given by

$$\begin{aligned}
\mathbb{E}[Y | D = 1, S = 1] - \frac{(1 - D) \cdot S \cdot Y \mathbf{1}_{Y \geq y_{\{1 - p_0(X), X\}}} \Pr(D = 1)}{\Pr(D = 0) \Pr(D = 1, S = 1)} &:= \mathbb{E}[Y | D = 1, S = 1] - \beta_L \\
\mathbb{E}[Y | D = 1, S = 1] - \frac{(1 - D) \cdot S \cdot Y \mathbf{1}_{Y \leq y_{\{p_0(X), X\}}} \Pr(D = 1)}{\Pr(D = 0) \Pr(D = 1, S = 1)} &:= \mathbb{E}[Y | D = 1, S = 1] - \beta_U
\end{aligned}$$

The bounds follow from Lemma 14 with the roles of  $D = 0$  and  $D = 1$  flipped.

Orthogonal moment equation for the upper bound  $\beta_L$  is

$$\mathbb{E}[\beta_L - \frac{(1-D) \cdot S \cdot Y 1_{\{Y \geq Q_{\{Y|D=0, S=1, X\}}(1-p_0(X), X)\}} \Pr(D=1)}{\Pr(D=1, S=1) \Pr(D=0)} + \sum_{j=1}^3 \alpha_j(W, \eta)] = 0, \quad (11.15)$$

where  $W = (D, S, SY, X)$  is the vector of data and  $\alpha_j(W, \eta)$ ,  $j \in \{1, 2, 3\}$  is the function that corrects bias of  $j$ 'th component of  $\eta$ . The true values of each bias correction term are

$$\begin{aligned} \alpha_1(W, \eta) &= -y_{\{1-p_0(X), X\}} \frac{\Pr(D=1) s(1, X)}{\Pr(D=1, S=1) s(0, X)} \left( \frac{(1-D)S}{\Pr(D=0)} - \eta_1(X) \right), \\ \alpha_2(W, \eta) &= y_{\{1-p_0(X), X\}} \frac{\Pr(D=1)}{\Pr(D=1, S=1)} \left( \frac{DS}{\Pr(D=1)} - \eta_2(X) \right), \\ \alpha_3(W, \eta) &= -y_{\{1-p_0(X), X\}} \frac{\Pr(D=1) s(0, X)}{\Pr(D=1, S=1)} \left( 1_{\{Y \leq Q_{\{Y|D=1, S=1, X\}}(1-p_0(X), X)\}} - 1 + p_0(X) \right). \end{aligned}$$

Orthogonal moment equation for the lower bound  $\beta_L$  is

$$\mathbb{E}[\beta_U - \frac{(1-D) \cdot S \cdot Y 1_{\{Y \leq Q_{\{Y|D=0, S=1, X\}}(p_0(X), X)\}} \Pr(D=1)}{\Pr(D=1, S=1) \Pr(D=0)} + \sum_{j=4}^6 \alpha_j(W, \eta)] = 0, \quad (11.16)$$

where  $\alpha_j(W, \eta)$ ,  $j \in \{4, 5, 6\}$  is the function that corrects bias of  $j$ 'th component of  $\eta$ . The true values of each bias correction term 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:

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

$$\gamma_{5,0}(X) = y_{\{s(1,X)/s(0,X), X\}} \frac{\Pr(D=1)}{\Pr(S=1, D=1)} \text{ and } \gamma_{6,0}(X) = -y_{p_0(X)} \frac{s(0, X) \Pr(D=1)}{\Pr(D=1, S=1)}.$$