12-29-2018

Bounding Average Returns to Schooling using Unconditional Moment Restrictions

Desire Kedagni
Iowa State University, dkedagni@iastate.edu

Lixiong Li
Penn State University

Ismael Mourifie
University of Toronto

Original Release Date: December 29, 2018

Bounding Average Returns to Schooling using Unconditional Moment Restrictions

Abstract

Abstract. In the last 20 years, the bounding approach for the average treatment effect (ATE) has been developing on the theoretical side, however, empirical work has lagged far behind theory in this area. One main reason is that, in practice, traditional bounding methods fall into two extreme cases: (i) On the one hand, the bounds are too wide to be informative and this happens, in general, when the instrumental variable (IV) has little variation; (ii) while on the other hand, the bounds cross, in which case the researcher learns nothing about the parameter of interest other than that the IV restrictions are rejected. This usually happens when the IV has a rich support and the IV restriction imposed in the model — full, quantile or mean independence— is too stringent, as illustrated in Ginther (2000). In this paper, we provide sharp bounds on the ATE using only a finite set of unconditional moment restrictions, which is a weaker version of mean independence. We revisit Ginther's (2000) return to schooling application using our bounding approach and derive informative bounds on the average returns to schooling in US.

Keywords

Instrumental Variable, Unconditional Moment Restrictions, Zero-Covariance Assumption, Heterogeneous treatment effects, Return to schooling, sheepskin effect

Disciplines

Econometrics | Economic Theory | Education | Statistical Methodology

BOUNDING AVERAGE RETURNS TO SCHOOLING USING UNCONDITIONAL MOMENT RESTRICTIONS

DÉSIRÉ KÉDAGNI, LIXIONG LI, AND ISMAËL MOURIFIÉ

Iowa State University, Penn State University and the University of Toronto

Abstract. In the last 20 years, the bounding approach for the average treatment effect (ATE) has been developing on the theoretical side, however, empirical work has lagged far behind theory in this area. One main reason is that, in practice, traditional bounding methods fall into two extreme cases: (i) On the one hand, the bounds are too wide to be informative and this happens, in general, when the instrumental variable (IV) has little variation; (ii) while on the other hand, the bounds cross, in which case the researcher learns nothing about the parameter of interest other than that the IV restrictions are rejected. This usually happens when the IV has a rich support and the IV restriction imposed in the model — full, quantile or mean independence— is too stringent, as illustrated in Ginther (2000). In this paper, we provide sharp bounds on the ATE using only a finite set of unconditional moment restrictions, which is a weaker version of mean independence. We revisit Ginther's (2000) return to schooling application using our bounding approach and derive informative bounds on the average returns to schooling in US.

Keywords: Instrumental Variable, Unconditional Moment Restrictions, Zero-Covariance Assumption, Heterogeneous treatment effects, Return to schooling, sheepskin effect.

JEL subject classification: C12, C21, C26.

Date: The present version is as of December 29, 2018. We thank Marc Henry, Thomas Russell and Paul Schrimpf for useful discussions and comments. We are grateful to Donna Ginther for sharing her dataset with us. We thank Sara Hossain for excellent research assistance. We thank participants at the 2018 CIREQ Econometrics Conference on Recent Advances in the Method of Moments. Mourifié thanks the support from Connaught and SSHRC Insight Grants # 435-2016-0045, 435-2018-1273. All errors are ours. Correspondence address: Department of Economics, University of Toronto, 150 St. George Street, Toronto ON M5S 3G7, Canada, ismael.mourifie@utoronto.ca.

1. Introduction

One of the most common problems in applied econometrics and statistics consists of estimating a causal effect of a discrete multi-valued endogenous explanatory variable D on an outcome Y. To illustrate, let us consider a simple potential outcome model:

$$Y = \sum_{d\in\mathcal{D}} Y_d \mathbb{1}[D=d] \tag{1.1}$$

$$= \sum_{d=1}^{T} \alpha_{d,0}\mathbb{1}[D=d] + Y_0 \tag{1.2}$$

where $\mathbb{1}[\cdot]$ denotes the indicator function, $\alpha_{d,d'} \equiv Y_d - Y_{d'}$ is a random causal effect (heterogeneous treatment effect), $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is the observed outcome, D is the observed discrete endogenous treatment/regressor, and $\big(Y_d, d \in \mathcal{D} = \{0,1,2,...,T\}\big)$ the unobserved potential outcomes. Denote $Z \in \mathcal{Z} \subseteq \mathbb{R}^p$ as the vector of observed instrumental variables (IVs). The most common estimation approach used by applied researchers to address the endogeneity of D is the two-stage least squares (2SLS). It is now well documented that in presence of heterogeneous treatment effect, the 2SLS may totally lose its causal interpretation without additional restrictions, and even when these stronger requirements hold, its may have a very complicated causal interpretation —an increasingly complicated weighted average of local average treatment effects (LATEs) between different D and Z realizations— which heavily relies on the instrument under use.[1]

A more intuitive and stable parameter to estimate is the ATE $\mathbb{E}[Y_d - Y_{d'}]$ and/or a distributional treatment effect (DTE) such as $\mathbb{P}(Y_d \leq y) - \mathbb{P}(Y_{d'} \leq y)$. However, when the treatment effect is heterogeneous, these causal parameters are not in general point-identified, even in the presence of valid instruments. Seminal works of Manski (1990, 1994) derived bounds on the ATE under the mean independence assumption ($\mathbb{E}[Y_d|Z] = \mathbb{E}[Y_d]$). Subsequently, several contributors to this literature investigate the identification power of stronger IV assumptions, like the quantile independence assumption and full independence assumptions, i.e., $Y_d \perp Z, d \in \mathcal{D}$.[2] Recently, Chesher and Rosen (2017) provides a unifying

[1] See Angrist and Imbens (1995), Angrist, Graddy, and Imbens (2000), and Heckman and Vytlacil (2005) among others; we also refer interested readers to Oreopoulos (2006), Deaton (2009), Heckman and Urzua (2010), Pearl (2011) and Aronow and Carnegie (2013) for additional discussions on whether the LATE is of genuine scientific interest.

[2] See Balke and Pearl (1997), Manski (2003), Kitagawa (2009, 2010), Beresteanu, Molchanov and Molinari (2011, 2012), Mourifié et al. (2018), Kédagni and Mourifié (2017), and Russell (2017), among many others.

framework that summarizes the sharp bounds of those quantities under mean, quantile and full independence.

Although the bounding approach is well-developed on the theoretical side, and also theoretically attractive given it relies on less stringent assumptions (no restrictions are imposed on the treatment equation, i.e., D is fully unrestricted), this approach is not often used in empirical works, especially in the empirical labor literature. One main reason is that, in practice, traditional bounding methods fall into two extreme cases: (i) On the one hand, the bounds are too wide to be informative and this happens, in general, when the IV has little variation; (ii) while on the other hand, the bounds cross, in which case the researcher learns nothing about the parameter of interests other than that the IV restrictions are rejected. This usually happens when the IV has a rich support and the IV restriction imposed in the model — full, quantile or mean independence— is too stringent. In other terms, the model is over-restricted.

This is illustrated in Ginther (2000) who studied the returns to schooling using a bounding approach. She was aiming to construct the Manski (1994) bounds on the ATE under mean independence assumption using family structure, distance to college, and school-quality characteristics (proxied by teacher-pupil ratio, percent of teachers with post-graduate degrees, beginning teacher salaries) as instruments. She argued based on references therein that these variables would be valid instruments. However, the bounds cross for all school-quality proxies which are continuous, and do not cross only for the instruments which are binary. As a consequence, Ginther (2000) no longer used the school-quality proxies, and constructed bounds only using the binary instruments, and found the resulting bounds uninformative. In doing so, Ginther (2000) went from an over-restricted model —$\mathbb{E}[Y_d|Z] = \mathbb{E}[Y_d]$— to a completely unrestricted model. However, the mean independence $\mathbb{E}[Y_d|Z] = \mathbb{E}[Y_d]$ is equivalent to $Cov(Y_d, h(Z)) = 0$ *for all integrable function h(.)*, so that even if $\mathbb{E}[Y_d|Z] \neq \mathbb{E}[Y_d]$, there may exist a large class of measurable functions of Z for which $Cov(Y_d, h(Z)) = \mathbb{E}[Y_d(h(Z) - \mathbb{E}h(Z))] = 0$.[3] While much weaker than the mean independence assumption, our set of unconditional IV moment restrictions may still

[3]For illustration, consider that there is a U-relationship between the potential earnings and the school-quality proxies that could be modelled as follows $Y_d = Z^2$, where Z is a symmetric variable with mean normalized to 0. We have $\mathbb{E}[Y_d|Z] = Z^2$. So, clearly both mean and full independences do not hold. However, $Cov(Y_d, h(Z)) = 0$ for all function $h(.)$ such that $h(z) = z^{2k+1}$ for each natural number k.

provide informative bounds.[4] Weakening over-restrictive identifying assumptions has been previously discussed in the literature. Manski and Pepper (2000) proposed to weaken the IV mean independence with its monotone version, namely the monotone IV (MIV) — $\mathbb{E}[Y_d|Z = z'] \geq \mathbb{E}[Y_d|Z = z]$ for $z' > z$. However, this monotonicity will not be realistic for some instruments, as we discuss in the empirical application. Manski and Pepper (2000, page 1009) have advocated the investigation of different ways to weaken the mean independence assumption.[5] To reach such a goal, this paper proposes a bounding approach of the ATE/DTE under unconditional IV moment restrictions —or equivalently a zero-covariance assumption— whenever the treatment effect is heterogeneous.

The first contribution of this paper is to derive sharp bounds on the ATE/DTE under unconditional IV moment restrictions. While the validity of the our proposed bounds will be quite intuitive, the proof of sharpness is much more involved and will require an appeal to the convex analysis literature. Using this bounding approach will allow a researcher to avoid going directly from an over-restricted model to an uninformative and/or unrestricted model, by giving her the possibility to visit various intermediate unconditional IV moment restrictions that are more likely to fit with the empirical framework under study. The choice of $h(.)$ should be firstly led by economic intuition. However, some inappropriate choices of $h(.)$ can been refuted by the data. In fact, our sharp bounds can cross in some circumstances, which shows that the unconditional IV moment restriction under use is violated or, in other words, is incompatible with the data generating process.

Second, we visit two statistical inference procedures that deliver uniformly valid confidence region for the sharp bounds of the ATE/DTE. The first one specializes Andrews and Shi (2017) to our framework by constructing a uniformly valid confidence region for parameters defined by a continuum of unconditional moment inequalities. The second one recasts our sharp bounds' characterization as a finite set of conditional moment inequalities and makes use of the Chernozhukov, Lee, Kim and Rosen (2015) or Andrews, Kim and Shi (2017) Stata packages. We show that both procedures have correct uniform asymptotic size and exclude parameter values outside the identified set with probability approaching one under relatively weak regularity conditions, especially the first approach. While the

[4]It allows point identification when α is random but uncorrelated with the treatment, i.e., $Cov(\alpha_{d,d'}, D) = 0$. Heckman, Urzua, and Vytlacil (2006) refer to it as a model without essential heterogeneity. This model contains the common linear IV model as special case.

[5]Manski and Pepper (2000, page 1009) said "Yet another variation on the MIV theme would be to weaken mean independence to some form of approximate mean independence."

second approach is more user-friendly, it is worth noting that, to be valid it requires more regularity assumptions than needed.

The last contribution is empirical. We revisit Ginther (2000)'s empirical application using our bounding approach. Notice that our bounding approach can be combined with additional prior information to sharpen the bounds, such as the Manski and Pepper (2000) monotone treatment response (MTR). In fact, in the empirical application we consider an assumption, namely the *ordered treatment response* (OTR) (Assumption 4) which does not restrict the selection mechanism into schooling, and also weakens the MTR assumption in the sense that it does not impose an ex-ante sign restriction on the treatment effect. Basically, the OTR assumption imposes that the potential outcomes are ordered but does not restrict the direction of the order. By definition, the linear IV model imposes the OTR assumption. However, a model with the OTR assumption is more general since it allows for heterogeneous treatment effects. Combining our bounding approach with the OTR assumption, we find that the causal effects of having more than 13 years of education versus 12 years on earnings is always non-negative, and it becomes strictly positive only after 16 years of education (which in general corresponds to a college degree). We cannot reject the hypothesis that dropping out of college confers no salary advantage over a high school diploma. Our findings are consistent the sheepskin effect in the returns to education literature, see Hungerford and Solon (1987), and Belman and Heywood (1991). The magnitude of our bounds is informative enough to rule out some of the returns to schooling point estimates present in the literature.

Other related literature. Chesher (2002, 2003, 2005) showed how to (partially) identify average structural functions in different nonparametric triangular systems when the full independence assumption is relaxed to its local version. Recently, Masten and Poirier (2017, 2018) brought up to date the breakdown point identification approach earlier discussed in Horowitz and Manski (1995) and references therein. This approach tries to determine the boundary between the set of assumptions which lead to a specific conclusion, and those which do not. However, the weakest combination of assumptions that lead to the desired conclusion would not be easily interpretable. While sharing some similarity, our goal here is different, in that we are interested in a set of unconditional IV moment restrictions for which the identified set of our parameter of interest is non-empty and potentially informative.

Outline. The remainder of the paper is organized as follows. Section 2 details the derivation of sharp bounds on the ATE/DTE under unconditional IV moment restrictions. Section 3 concerns the inference procedure. Section 4 presents our empirical application. The last section concludes. Proofs of the main results are collected in the appendix.

2. Identification power of the unconditional IV moment restrictions

Consider the potential outcome model (POM) described in Equation (1.1). In this section, we will provide sharp bounds on the following general form of treatment effect parameters:

$$\mathbb{E}[g(Y_d) - g(Y_{d'})],\ d \neq d' \in \mathcal{D} \tag{2.1}$$

for any integrable real function $g(.)$. In fact, the POM defined for Y implies the following POM:

$$g(Y) \quad = \quad \sum_{d \in \mathcal{D}} g(Y_d)\mathbb{1}[D = d]. \tag{2.2}$$

This latter formulation will allow us to provide a general characterization for the sharp bounds of a large class of treatment effect parameters. For instance, if $g(Y_d) = Y_d$ we recover the ATE, and if $g(Y_d) = 1\{Y_d \leq y\}$, we recover the DTE. Hereafter, we assume that all random variables $g(Y_d), d \in \mathcal{D}$ are integrable, i.e., $\mathbb{E}[|g(Y_d)|] < \infty$, $d \in \mathcal{D}$. We make the following assumption:

Assumption 1. *[Bounded supports] $g(.)$ is an integrable known real function of Y, and (i) Rectangular supports: $Supp\{g(Y_d)\} = Supp\{g(Y_d)|D = d\} = Supp\{g(Y_d)|D = d'\}, d, d' \in \mathcal{D}$; (ii) The supports of $g(Y_d), d \in \mathcal{D}$ are bounded, i.e., $\underline{g}_d \equiv \inf\{Supp[g(Y_d)]\}$ and $\overline{g}_d \equiv \sup\{Supp[g(Y_d)]\}$ are finite.*

This assumption is similar to the assumption usually considered by Manski. See for instance Manski (1990, 1994) among many others who assume that the potential outcomes Y_d, $d \in \mathcal{D}$ have bounded supports that are known by the researcher. While not formally stated in Manski (1990, 1994), assumption 1 (i) is always assumed in Manski's derivations. We impose the support condition on $g(Y_d)$ instead on Y_d, because even if Y_d is not itself bounded, we can derive bounds on all transformations of Y_d that have known bounded support; for instance, the DTE where $g(Y_d) = 1\{Y_d \leq y\}$.

Technically, we may relax Assumption 1 (i) and allow that $Supp(g(Y_d)|D = d) \neq Supp(g(Y_d)|D = d')$; $d, d' \in \mathcal{D}$, but in such a case Assumption 1 (ii) should be replaced by the knowledge of all $Supp(g(Y_d)|D = d')$; $d', d \in \mathcal{D}$.

Assumption 2. *[Zero-Covariance] There exists a set of instrumental variables $Z \in \mathcal{Z} \subseteq \mathbb{R}^p$, $p \geq 1$ and an integrable function $h(\cdot)$ mapping Z to $\mathbb{R}^m, 1 \leq m < \infty$, such that $Cov(g(Y_d), h(Z)) = 0$, i.e., $\mathbb{E}[g(Y_d)(h(Z) - \mathbb{E}h(Z))] = \mathbb{E}[(g(Y_d) - \mathbb{E}g(Y_d))h(Z)] = 0$.*

Now, we propose our approach to derive the bounds on the treatment effects under Assumptions 1 and 2. To simplify the analysis, we make the following normalization $\mathbb{E}[h(Z)] = 0$. Under Assumption 2, we have

$$Cov(g(Y_d), h(Z)) = 0, \tag{2.3}$$

which implies that

$$Cov(g(Y_d), \lambda' h(Z)) = 0, \tag{2.4}$$

for all $\lambda \in \mathbb{R}^m$, which in turn implies

$$\mathbb{E}[g(Y_d)] = \mathbb{E}[g(Y_d)(1 + \lambda' h(Z))], \tag{2.5}$$

given that $\mathbb{E}[h(Z)] = 0$. For $d = 0$, we have

$$\mathbb{E}[g(Y_0)] = \mathbb{E}\big[g(Y)(1 + \lambda' h(Z))\mathbb{1}[D = 0]\big] + \mathbb{E}\big[g(Y_0)(1 + \lambda' h(Z))\mathbb{1}[D \neq 0]\big], \tag{2.6}$$

where the first term of the right hand side (RHS) holds as $g(Y)1\{D = d\} = g(Y_d)1\{D = d\}$, $d \in \mathcal{D}$. Under Assumption 1, the last term of the RHS can be bounded as follows:

$$\begin{aligned} &\mathbb{E}\big[\mathbb{1}[D \neq 0] \min\{\overline{g}_0(1 + \lambda' h(Z)), \underline{g}_0(1 + \lambda' h(Z))\}\big] \\ &\quad \leq \mathbb{E}\big[g(Y_0)(1 + \lambda' h(Z))\mathbb{1}[D \neq 0]\big] \leq \\ &\mathbb{E}\big[\mathbb{1}[D \neq 0] \max\{\overline{g}_0(1 + \lambda' h(Z)), \underline{g}_0(1 + \lambda' h(Z))\}\big]. \end{aligned} \tag{2.7}$$

Therefore, we can derive the following bounds on $\mathbb{E}[g(Y_0)]$:

$$\begin{aligned} &\mathbb{E}\big[g(Y)(1 + \lambda' h(Z))\mathbb{1}[D = 0]\big] + \mathbb{E}\big[\mathbb{1}[D \neq 0] \min\{\overline{g}_0(1 + \lambda' h(Z)), \underline{g}_0(1 + \lambda' h(Z))\}\big] \\ &\qquad \leq \mathbb{E}[g(Y_0)] \leq \\ &\mathbb{E}\big[g(Y)(1 + \lambda' h(Z))\mathbb{1}[D = 0]\big] + \mathbb{E}\big[\mathbb{1}[D \neq 0] \max\{\overline{g}_0(1 + \lambda' h(Z)), \underline{g}_0(1 + \lambda' h(Z))\}\big]. \end{aligned} \tag{2.8}$$

for all $\lambda \in \mathbb{R}^m$. Similarly, we can derive the bounds for $\mathbb{E}[g(Y_d)], d \neq 0$. To ease the exposition, we use some additional notation:

$$\begin{aligned} \underline{f}_d(Y, D, \delta) \equiv &\mathbb{1}[D = d]\delta g(Y) + \mathbb{1}[D \neq d] \min\Big(\delta \underline{g}_d, \delta \overline{g}_d\Big), \\ \overline{f}_d(Y, D, \delta) \equiv &\mathbb{1}[D = d]\delta g(Y) + \mathbb{1}[D \neq d] \max\Big(\delta \underline{g}_d, \delta \overline{g}_d\Big), \end{aligned}$$

for $d \in \mathcal{D}$. Thus, under Assumption 1 and 2, from inequality (2.8) we obtain the following bounds on the average structural functions:

$$\sup_{\lambda \in \mathbb{R}^m} \mathbb{E}\Big[\underline{f}_d(Y, D, 1 + \lambda' h(Z))\Big] \leq \mathbb{E}[g(Y_d)] \leq \inf_{\lambda \in \mathbb{R}^m} \mathbb{E}\Big[\overline{f}_d(Y, D, 1 + \lambda' h(Z))\Big],\ d \in \mathcal{D}.$$

As can be seen, because the zero-covariance assumption still holds for any linear combination of the element of the vector $h(Z)$, our bounding approach consists of looking for, within the class of all linear combinations of $h(Z)$, the ones that are able to provide the tightest bounds for $\mathbb{E}[g(Y_d)], d \in \mathcal{D}$. While it clearly appears from our above derivations that the bounds are valid, the following theorem shows that the bounds are indeed sharp.

Theorem 1. *Assume, as a normalization, that $\mathbb{E}h(Z) = 0$. Under Assumptions 1 and 2, Θ_I the identified set of $(\theta_0, ..., \theta_T) \equiv \Big(\mathbb{E}[g(Y_0)], ..., \mathbb{E}[g(Y_T)]\Big)$ is given as follows:*

$$\Theta_I = \Big[\underline{\theta}_0, \overline{\theta}_0\Big] \times ... \times \Big[\underline{\theta}_T, \overline{\theta}_T\Big]$$

where

$$\underline{\theta}_d \equiv \sup_{\lambda \in \mathbb{R}^m} \mathbb{E}\Big[\underline{f}_d(Y, D, 1 + \lambda' h(Z))\Big], \quad \overline{\theta}_d \equiv \inf_{\lambda \in \mathbb{R}^m} \mathbb{E}\Big[\overline{f}_d(Y, D, 1 + \lambda' h(Z))\Big],\ d \in \mathcal{D}.$$

As a result, the interval $[\underline{\theta}_d, \overline{\theta}_d]$ is the sharp bound for $\mathbb{E}[g(Y_d)]$ for $d \in \mathcal{D}$, and interval $[\underline{\theta}_d - \overline{\theta}_{d'}, \overline{\theta}_d - \underline{\theta}_{d'}]$ is the sharp bound for $\mathbb{E}[g(Y_d) - g(Y_{d'})]$ $d, d' \in \mathcal{D}$.

To see the intuition behind Theorem 1, we provide a heuristic proof. The formal proof is relegated in Appendix C.1.2. Let us propose a vector of potential outcomes $(G_0, ..., G_T)$ to achieve the upper bounds $\overline{\theta}_0, ..., \overline{\theta}_T$ which depend on observables (Y, D, Z), the knowledge of $(\underline{g}_d, \overline{g}_d), d \in \mathcal{D}$, and rationalizes our model, (i.e., Equation (2.2) and Assumptions 1 and 2). Heuristically, assume for each λ, $\mathbb{P}(1 + \lambda' h(Z) = 0) = 0$. In other words, this means that the event where $\overline{f}_d(Y, D, 1 + \lambda' h(Z))$ is not differentiable in λ is of measure zero. Moreover, assume there exists a finite minimizer λ_d^* such that $\overline{\theta}_d = \mathbb{E}\big[\overline{f}_d(Y, D, 1 + \lambda_d^{*\prime} h(Z))\big]$. Then, the first order condition implies that,

$$\left.\frac{\partial \mathbb{E}\big[\overline{f}_d(Y, D, 1 + \lambda' h(Z))\big]}{\partial \lambda}\right|_{\lambda = \lambda^*} = 0, \tag{2.9}$$

where it can be shown that:

$$\left.\frac{\partial \mathbb{E}[\overline{f}_d(Y,D,1+\lambda' h(Z))]}{\partial \lambda}\right|_{\lambda=\lambda^*} = \mathbb{E}\Bigg[\Big\{\mathbb{1}[D=d]g(Y) + \mathbb{1}[D\neq d]\times \Big(\underline{g}_d \mathbb{1}(1+\lambda_d^{*\prime}h(Z)\leq 0) + \overline{g}_d \mathbb{1}(1+\lambda_d^{*\prime}h(Z)\geq 0)\Big)\Big\}\cdot h(Z)\Bigg],$$

$$\equiv \mathbb{E}[G_d \cdot h(Z)],$$

with

$$G_d \equiv \mathbb{1}[D=d]g(Y) + \mathbb{1}[D\neq d]\left(\underline{g}_d \mathbb{1}(1+\lambda_d^{*\prime}h(Z)\leq 0) + \overline{g}_d \mathbb{1}(1+\lambda_d^{*\prime}h(Z)\geq 0)\right). \tag{2.10}$$

As can be seen, by construction we have: (i) $g(Y) = \sum_{d\in\mathcal{D}} G_d \mathbb{1}[D=d]$, (ii) $\inf\{Supp(G_d|D\neq d)\} = \inf\{Supp(G_d|D=d)\} = \underline{g}_d$, $\sup\{Supp(G_d|D\neq d)\} = \sup\{Supp(G_d|D=d)\} = \overline{g}_d$ which verifies Assumption 1, (iii) by combining (2.9) and (2.10), we get $\mathbb{E}[G_d h(Z)] = 0$, which is equivalent to the zero-covariance assumption under the normalization $\mathbb{E}[h(Z)] = 0$, and finally (iv) it is easy to check that

$$\overline{f}_d(Y,D,1+\lambda_d^{*\prime}h(Z)) = G_d\cdot(1+\lambda_d^{*\prime}h(Z)).$$

Then, by taking the expectation of the latter equality together with $\mathbb{E}[G_d h(Z)] = 0$ we obtain that $\overline{\theta}_d = \mathbb{E}[G_d]$, $d\in\mathcal{D}$. This completes our heuristic argument. However, as can be seen this construction heavily relies on the fact that $\mathbb{E}\overline{f}_d(Y,D,1+\lambda' h(Z))$ was assumed to be differentiable in λ and that the minimizer $\lambda_d^*, d\in\mathcal{D}$ exists. Both of these conditions may not be true in general. We therefore appeal to some convex analysis results to show that our constructive approach can be carried over whenever both conditions fail to hold.

It is worth emphasizing that the current expression of our bounds in Theorem 1 relies on the normalization $\mathbb{E}h(Z) = 0$. This normalization —which is without loss of generality for the identification result— was very useful in having a clear and simple exposition of our identification strategy. However, this simplification may bring additional challenges for the inference procedure. More precisely, if $\mathbb{E}h(Z) = \beta \neq 0$, β is a nuisance parameter that needs to be estimated consistently. We can certainly construct the joint confidence region for (θ_d, β) and then project out β. However, such a projection could generate an overly conservative confidence region. We will therefore propose a more general characterization of the identified set Θ_I that does not rely on the normalization and for which the inference procedure will not require the estimation of β.

We develop a formal proof in the Appendix C.1.1 for this general setting, yet the intuition remains the same. Let $\mathcal{S}_{m+1}$ be the unit sphere in $\mathbb{R}^{m+1}$, i.e. $\mathcal{S}_{m+1} \equiv \{x\in\mathbb{R}^{m+1} : \|x\| = 1\}$.

Theorem 2. *Under Assumptions 1 and 2, for any* $(\theta_0, ..., \theta_T) \in \mathbb{R}^T$*, the following statements are equivalent,*

(1) $(\theta_0, ..., \theta_T) \in \Theta_I$,
(2) *for each* $d \in \mathcal{D}$, θ_d *satisfies the inequality*

$$0 \leq \inf_{(\gamma,\lambda)\in\mathcal{S}_{m+1}} \mathbb{E}\big[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(\gamma + \lambda' h(Z))\big] \tag{2.11}$$

where γ *is a scalar and* λ *is an* m*-dimensional vector.*

We have a few important remarks.

Remark 1 (Link with Theorem 1)**.** *When* $\mathbb{E}h(Z) = 0$*, inequality (2.11) delivers exactly the same sharp bounds proposed in Theorem 1. The proof of this equivalence is relegated to Appendix C.1.2.*

Remark 2 (Testable implications of the Zero-Covariance assumption)**.** *Inequality (2.11) can be violated; more precisely, there may not exist* $\theta_d \in \mathbb{R}$ *such that (2.11) holds. In fact, if* $\mathbb{E}h(Z) = 0$*, the sharp testable implications of Assumptions 1 and 2 are given by:*

$$\sup_{\lambda\in\mathbb{R}^m} \mathbb{E}\Big[\underline{f}_d(Y, D, 1 + \lambda' h(Z))\Big] < \inf_{\lambda\in\mathbb{R}^m} \mathbb{E}\Big[\overline{f}_d(Y, D, 1 + \lambda' h(Z))\Big], \;\; d \in \mathcal{D}. \tag{2.12}$$

In such a case, the violation of the latter inequality implies the violation of either Assumption 1 or Assumption 2. Interestingly, notice that when $g(Y_d) = 1\{y < Y_d \leq y'\}$*, for* $y, y' \in \mathcal{Y} \cup \{-\infty, \infty\}$*, Assumption 1 trivially holds. Then, such a violation implies directly a violation of the Zero-Covariance assumption as stated in Assumption 2. This contrasts with the prevalent and long-held idea in the applied economics discipline that the IV zero-covariance assumption is fundamentally non-testable, as can be found in Wooldridge (2010) econometrics textbook, p.92: "the covariance* $Cov(g(Y_d); h(Z))$ *involves the unobservable* $g(Y_d)$*, and therefore we cannot test anything about* $Cov(g(Y_d); h(Z))$*...".*

Remark 3. *Ekeland, Galichon and Henry (2010) use the optimal transport approach to provide a characterization of the identified set for parameters of interest associated with a large class of incomplete models (including the POM). They also consider models where constraints on the unobservables are characterized by set of unconditional moment restrictions. However, their sharpness proof rely on a uniform integrability and tightness restrictions that may not hold in presence of moments with unbounded supports, as we entertained here, i.e.,* $\mathbb{E}[g(Y_d)(h(Z) - \mathbb{E}h(Z))]$*. Recall that we avoid to restrict* $h(Z)$ *as much as possible in order to be able to visit a large class of instruments.*

Remark 4. *Schennach (2014) provides an alternative entropy based approach for models with unconditional moment restrictions on both observable and unobservables. The POM studied in this paper can be transformed to fit in the framework of Schennach (2014). However, Schennach(2014) studied an* extended *notion of identified set, while our notion of identified set is in line with the traditional definition used in the literature (e.g. Ekeland, Galichon and Henry (2010)). The two notions are different. Hence, the identification result in Schennach (2014) cannot ensure the sharpness of her approach in our context. In addition, the approach in Schennach(2014) requires a choice of reference distribution, the support of which could affect the set of parameters characterized by her method. Finally, as shown later in Proposition 1, our identification conditions can be further simplified into finite number of moment inequalities when the instrument has finite support, which makes the inference procedure much easier to implement.*

As can be seen, the identified set for θ_d is characterized by intersection bounds where the infima/suprema are taken over the unit sphere $\mathcal{S}_{m+1}$ of $\mathbb{R}^{m+1}$. In practice, the finite sample performance of the inferential method depends on the set over which we search, especially when many choices of λ provide redundant information. One solution is to follow the idea discussed in Galichon and Henry (2006, 2011) and Chesher and Rosen (2017) to find a low (or the lowest) subset $\Lambda \subseteq \mathcal{S}_{m+1}$ that preserves the sharpness of the characterization of the identified set, in the following sense:

Definition 1 (Core-determining class)**.** $\Lambda \subseteq \mathcal{S}^{m+1}$ *is a core-determining class, if inequality* (2.11) *is equivalent to*

$$0 \leq \inf_{(\gamma,\lambda)\in\Lambda} \mathbb{E}\big[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(\gamma + \lambda' h(Z))\big].$$

Assumption 3 (Non-Redundant Instruments)**.** *The support of* $h(Z)$ *contains* $m+1$ *vectors* $h_1, ..., h_{m+1}$*, such that the vectors* $(1, h_1), ..., (1, h_{m+1})$ *are linearly independent.*

If Assumption 3 is violated, then one component of $h(Z)$ can be represented as an affine transformation of others. As the zero-covariance condition in Assumption 2 is invariant to affine transformations, Assumption 3 essentially implies there is no redundant instruments in $h(Z)$.

Proposition 1. *Suppose Assumption 3 hold, and suppose one of the following conditions holds,*

(1) *Supp*$(h(Z))$ *is a finite set,*

(2) *Supp*$(h(Z))$ *is a convex hull of a finite set.*

Then, the following set Λ *is a core-determining class,*

$$\Lambda \equiv \Big\{(\gamma,\lambda)\in\mathcal{S}_{m+1}: \text{ there exists } m \text{ linearly independent vectors } h_1,...,h_m \text{ in Supp}(h(Z)) \text{ such that for each } i=1,...,m,\ \gamma+\lambda' h_i=0\Big\}.$$

Remark 5. *When* *Supp*$(h(Z))$ *is a finite set, then* Λ *is also a finite set. If* *Supp*$(h(Z))$ *contains* k *vectors, then* Λ *contains at most* $2\times k!/(m!(k-m)!)$ *points. As a special case, when* $m=1$ *and* *Supp*$(h(Z))=\{h_1,...,h_k\}$*, then* Λ *in Proposition 1 can be written as*

$$\Lambda=\left\{\left(\frac{-h_i}{\sqrt{1+h_i^2}},\frac{1}{\sqrt{1+h_i^2}}\right): i=1,...,k\right\}\bigcup\left\{\left(\frac{h_i}{\sqrt{1+h_i^2}},\frac{-1}{\sqrt{1+h_i^2}}\right): i=1,...,k\right\}.$$

Furthermore, when $m=1$ *and* *Supp*$(h(Z))=[\underline{h},\overline{h}]$*,* Λ *can be written as*

$$\Lambda=\left\{\left(\frac{-h}{\sqrt{1+h^2}},\frac{1}{\sqrt{1+h^2}}\right): h\in[\underline{h},\overline{h}]\right\}\bigcup\left\{\left(\frac{h}{\sqrt{1+h^2}},\frac{-1}{\sqrt{1+h^2}}\right): h\in[\underline{h},\overline{h}]\right\}.$$

Now that we have exposed our identification approach under unconditional IV moment restrictions, it is worth noting that our bounding approach can be combined with additional prior information to sharpen the bounds. In some applications, applied researchers have stronger prior information about the average effect of the treatment. For instance, in the return to education literature, the treatment which is the number of years of schooling is ordered, and Manski and Pepper (2000) considered the monotone treatment response (MTR) assumption, i.e., $(d>d' \Rightarrow Y_d \geq Y_{d'}$ a.s.), which means that an additional year of education does not decrease potential earnings. Notice that this restriction assumes an answer to part of the question under study since it imposes that the sign of the treatment effect is known and is positive. Therefore, it only allows the researcher to tighten the bounds on the magnitude of the treatment effect.

In our case, we will consider a generalization of this assumption, which does not ex-ante impose the sign of the treatment effect. We name it the *ordered treatment response*:

Assumption 4. *[OTR]* $\mathcal{D}$ *is an ordered set and for all* (d,d') *such that* $d>d'$*,*

$$\text{either} \quad Y_d \geq Y_{d'} \quad a.s. \quad \text{or} \quad Y_d \leq Y_{d'} \quad a.s. \tag{2.13}$$

Like the MTR, the OTR imposes that the potential outcome are all ordered in one direction, but unlike the MTR, it does not ex-ante impose a direction. So, under this assumption, the sign of the treatment effect remains ex-ante unknown, but could be identified by the

data. To have a better intuition of the assumption, consider the following random coefficient model: $Y = \alpha D + \varepsilon$, where α is a random coefficient. We have $Y_{d+s} - Y_d = \alpha \times s$, then the OTR holds if α is either a non-negative random variable or a non-positive random variable, more precisely $\mathbb{P}(\alpha < 0) = 0$ or $\mathbb{P}(\alpha > 0) = 0$, respectively. However, MTR imposes that α can only be a non-negative random variable, i.e., $\mathbb{P}(\alpha < 0) = 0$. In the linear IV model where α is constant and its sign is ex-ante unknown, OTR holds, but MTR does not. A special case of the OTR assumption has been recently introduced by Machado, Shaikh and Vytlacil (2018). They consider a special framework where the outcome, the treatment and the instrument are all binary. To ease the exposition of the main text, we relegate in Appendix A our derivations that show how to combine our unconditional IV moment restrictions with the OTR to provide tighter bounds on the treatment effect of interest and potentially identify the sign of the treatment effect.

3. Inference Procedure

In the previous section, we show that the identified set of our parameter of interest can be characterized as follows:

$$\inf_{(\gamma,\lambda)\in\mathcal{S}_{m+1}} \mathbb{E}\big[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(\gamma + \lambda' h(Z))\big] \geq 0, \quad \forall\ d \in \mathcal{D}, \tag{3.1}$$

which can equivalently be rewritten as follows:

$$\mathbb{E}\big[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(\gamma + \lambda' h(Z))\big] \geq 0, \quad \forall\ (\gamma, \lambda, d) \in \mathcal{S}_{m+1} \times \mathcal{D}. \tag{3.2}$$

Therefore, constructing a valid confidence region for the identified set Θ_I is equivalent to constructing a valid confidence region for parameters defined by a continuum of unconditional moment inequalities. Inference for a continuum of unconditional moment inequalities has been recently discussed in Andrews and Shi (2017) and references therein.

Notice that the continuum of inequalities we are considering here can also equivalently be rewritten as a finite number of conditional moment inequalities. In fact, let $(V, W) \in \mathcal{V} \times \mathcal{W} = \Lambda$ be a vector of random variables that are statistically independent of (Y, Z, D), i.e., $(V, W) \perp (Y, Z, D)$, where V is a univariate random variable and W is an m-dimensional random vector. Λ is a core-determining class as stated in Definition 1. When it is possible, the researcher could use Λ as defined in Proposition 1 or she could just set $\Lambda = \mathcal{S}^{m+1}$. Under this independence assumption, it can easily be shown that inequality (3.2) is equivalent to:

$$\mathbb{E}\big[\overline{f}_d(Y, D, V + W' h(Z)) - \theta_d(V + W' h(Z)) | V, W\big] \geq 0, \quad \forall\ d \in \mathcal{D}, \text{ almost surely.} \tag{3.3}$$

Inference for finite number of conditional moment inequalities as defined in (3.3) has been entertained in Andrews and Shi (2013) and Chernozhukov, Lee, and Rosen (2013). The potential practical "advantage" of rewriting (3.2) as (3.3) is to provide a valid inference procedure that is more user-friendly for applied researchers. Indeed, Chernozhukov, Kim, Lee, and Rosen (2015, CKLR) and Andrews, Kim and Shi (2017, AKS) provide Stata packages to implement uniformly asymptotically valid testing and inference procedures for conditional moment inequalities that verify the set of conditions in Chernozhukov, Lee, and Rosen (2013) and Andrews and Shi (2013), respectively.

However, recasting (3.2) as in (3.3) requires imposing more regularity assumptions than needed. In the following, we propose two main testing procedures that allow to construct uniformly valid confidence region for θ_d and $\theta_d - \theta_{d'}$. The first procedure is the specialization of Andrews and Shi (2017)'s testing procedure to our context, and the second is a procedure that can be implemented using the Stata Packages proposed by AKS, or CKLR. We assume throughout this section that a sample consists of independent and identically distributed (i.i.d.) observations, $\{(Y_i, Z_i, D_i) : i \geq 1\}$.

Approach 1: Based on Andrews and Shi (2017).

Algorithm 1. (Implementation)

Step 1. Compute the test statistic:

$$T_n(\theta_d, d, \Lambda) \equiv \left[\min\left(\inf_{(\gamma,\lambda)\in\Lambda}\left[\frac{n^{1/2}\overline{m}_n(\theta_d, d, \gamma, \lambda)}{\sqrt{\epsilon + \hat{\sigma}_n^2(\theta_d, d, \gamma, \lambda))}}\right], 0\right)\right]^2,$$

where

$$\overline{m}_n(\theta_d, d, \gamma, \lambda) \equiv \frac{1}{n}\sum_{i=1}^{n}\left[\overline{f}_d(Y_i, D_i, \gamma + \lambda' h(Z_i)) - \theta_d(\gamma + \lambda' h(Z_i))\right],$$

$$\hat{\sigma}_n^2(\theta_d, d, \gamma, \lambda) \equiv \frac{1}{n}\sum_{i=1}^{n}\left(\overline{f}_d(Y_i, D_i, \gamma + \lambda' h(Z_i)) - \theta_d(\gamma + \lambda' h(Z_i)) - \overline{m}_n(\theta_d, d, \gamma, \lambda)\right)^2,$$

and fix $\epsilon = 0.05$.[6]

Step 2. Compute

$$\bar{\varphi}_n(\theta_d, d, \gamma, \lambda) \equiv \hat{\sigma}_n(\theta_d, d, \gamma, \lambda) B_n \mathbb{1}\left(\kappa_n^{-1} n^{1/2} \frac{\overline{m}_n(\theta_d, d, \gamma, \lambda)}{\epsilon + \hat{\sigma}_n(\theta_d, d, \gamma, \lambda))} > 1\right)$$

[6] In principle, ϵ can be any sufficiently small positive constant. Here, we follow Andrews and Shi (2017) to set $\epsilon = 0.05$.

where $B_n \equiv (0.4\ln(n)/\ln\ln(n))^{1/2}$ and $\kappa_n \equiv (0.3\ln(n))^{1/2}$.

Step 3. Generate B bootstrap samples $\{(Y^*_{i,s}, Z^*_{i,s}, D^*_{i,s}) : i = 1, ..., n\}$ for $s = 1, ..., B$ using the standard nonparametric i.i.d. bootstrap.

Step 4. Compute

$$\begin{aligned}
\overline{m}^*_n(\theta_d, d, \gamma, \lambda) &\equiv \frac{1}{n}\sum_{i=1}^n \left[\overline{f}_d(Y^*_i, D^*_i, \gamma + \lambda' h(Z^*_i)) - \theta_d(\gamma + \lambda' h(Z^*_i))\right] \\
\hat{\sigma}^{*2}_n(\theta_d, d, \gamma, \lambda) &\equiv \frac{1}{n}\sum_{i=1}^n \left(\overline{f}_d(Y^*_i, D^*_i, \gamma + \lambda' h(Z^*_i)) - \theta_d(\gamma + \lambda' h(Z^*_i)) - \overline{m}_n(\theta, d, \gamma, \lambda)\right)^2 \\
T^*_{n,s}(\theta_d, d, \Lambda) &\equiv \left[\min\left(\inf_{(\gamma,\lambda)\in\Lambda}\left[\frac{n^{1/2}(\overline{m}^*_n(\theta_d, d, \gamma, \lambda) - \overline{m}_n(\theta_d, d, \gamma, \lambda)) + \bar{\varphi}_n(\theta_d, d, \gamma, \lambda)}{\sqrt{\epsilon + \hat{\sigma}^{*2}_n(\theta_d, d, \gamma, \lambda))}}\right], 0\right)\right]^2
\end{aligned}$$

Step 5. Take the bootstrap critical value, namely $c_{n,1-\alpha}(\theta_d, d, \Lambda)$ to be the $1-\alpha$ sample quantile of the bootstrap statistics $\{T^*_{n,s}(\theta_d, d, \Lambda) : s = 1, ..., B\}$.

Step 6. Construct the level $1-\alpha$ confidence interval for θ_d as follows

$$\mathrm{CI}_{n,1-\alpha}(d) \equiv \{\theta_d \in [\underline{g}_d, \overline{g}_d] : T_n(\theta_d, d, \Lambda) \leq c_{n,1-\alpha}(\theta_d, d, \Lambda)\}. \tag{3.4}$$

A level $1-\alpha$ confidence interval for $\theta_d - \theta_{d'}$ can be constructed based on confidence interval for θ_d and $\theta_{d'}$ using a Bonferroni correction, i.e.,

$$\mathrm{CI}_{n,1-\alpha}(d, d') \equiv \left\{\theta_d - \theta_{d'} : \theta_d \in \mathrm{CI}_{n,1-\alpha/2}(d), \theta_{d'} \in \mathrm{CI}_{n,1-\alpha/2}(d')\right\} \tag{3.5}$$

Remark 6. *Calculation of $T_n(\theta_d, d, \Lambda)$ and $T^*_{n,s}(\theta_d, d, \Lambda)$ involves solving a nonlinear minimization problem, which can be numerically challenging when Λ is not a finite set. In practice, one could discretize Λ into a dense grid. The resulted confidence regions still have a correct level, but they have less power since the discretized Λ may not be core-determining.*

Remark 7. *The confidence interval constructed via Bonferroni correction is generally conservative. When Λ is finite, it's possible to get sharper confidence intervals for $\theta_d - \theta_{d'}$ using the procedures in Kaido, Molinari and Stoye (2017) or Bugni, Canay and Shi (2017).*

Approach 2: Based on Andrews and Shi (2013) and AKS. Let $F_{V,W}$ denote the pre-specified distribution of (V, W) whose support of (V, W) equals Λ.

Algorithm 2. (Implementation using **cmi-test**)

Step 1. Draw auxiliary i.i.d. samples $(V_i, W_i)_{i=1}^n$ from distribution $F_{V,W}$ independent of data samples $\{Y_i, Z_i, D_i\}_{i=1}^n$.

Step 2. Let $\phi_d(\theta_d, \alpha, \{Y_i, Z_i, D_i, V_i, W_i\}_{i=1}^n) \in \{0,1\}$ be the result of the Stata command **cmi-test** implemented to test the null hypothesis:

$$H_0: \ \theta_d \text{ satisfies (3.3)} \ \ vs \ \ H_a: \theta_d \text{ does not satisfy (3.3)}$$

For the joint parameter $(\theta_d, \theta_{d'})$, let $\phi(\theta_d, \theta_{d'}, \alpha, \{Y_i, Z_i, D_i, V_i, W_i\}_{i=1}^n) \in \{0,1\}$ be the result of the Stata command **cmi-test** implemented to test the null hypothesis:

$$\tilde{H}_0: \ (\theta_d, \theta_{d'}) \text{ satisfies (3.3)} \ \ vs \ \ \tilde{H}_a: (\theta_d, \theta_{d'}) \text{ does not satisfy (3.3)}$$

with nominal level $1-\alpha$ and data observation $\{Y_i, Z_i, D_i, V_i, W_i\}_{i=1}^n\}$.

Step 3. If the null hypothesis H_0 (resp. $\tilde{H}_0$) is rejected, let $\phi_d(\theta_d, \alpha, \{Y_i, Z_i, D_i, V_i, W_i\}_{i=1}^n) = 1$ (resp. $\phi(\theta_d, \theta_{d'}, \alpha, \{Y_i, Z_i, D_i, V_i, W_i\}_{i=1}^n) = 1$).

Step 4. Construct the level $1-\alpha$ confidence interval for θ_d as follows

$$\text{CI}_{n,1-\alpha}(d) \equiv \{\theta_d \in [\underline{g}_d, \overline{g}_d] : \phi_d(\theta_d, \alpha, \{Y_i, Z_i, D_i, V_i, W_i\}_{i=1}^n) = 0\}. \tag{3.6}$$

Step 5. Construct the level $1-\alpha$ confidence interval for $(\theta_d - \theta_{d'})$ as follows

$$\text{CI}_{n,1-\alpha}(d,d') \equiv \Big\{\theta_d - \theta_{d'} : (\theta_d, \theta_{d'}) \in [\underline{g}_d, \overline{g}_d] \times [\underline{g}_{d'}, \overline{g}_{d'}], \phi(\theta_d, \theta_{d'}, \alpha, \{Y_i, Z_i, D_i, V_i, W_i\}_{i=1}^n) = 0\Big\}. \tag{3.7}$$

Remark 8. *In the Step 2 of Algorithm 2, one could alternatively use the* **clrtest** *Stata command proposed by Chernozhukov, Kim, Lee, and Rosen (2015).*

We show in Appendix D that under Assumption 1, and additional regularity conditions the confidence regions (3.4, 3.5) and (3.6, 3.7) computed as described in Algorithm 1 and 2, respectively, have correct uniform asymptotic size and exclude parameter values outside the identified set with probability approaching one. For each of the procedures, we spell out the regularity conditions under which each procedure provides a uniformly valid confidence region. As we will see, these conditions are relatively weak in our context, especially the ones for the Andrews and Shi (2017) approach. The proofs mainly show that all the good statistical properties of the testing procedure developed in Andrews and Shi (2013, 2017) carry through in our context under these regularity conditions. To ease the discussion in the main text, we relegate all the technical derivations and proofs to Appendix D.

4. Bounding Average Returns to Schooling

Estimating the causal impact of college education on later earnings has always been troublesome for economists because of the endogeneity of the level of education. To evaluate the returns to schooling in the US, researchers have used various point estimators. The

most significant part of the applied literature has considered the 2SLS estimand, and most of the well known 2SLS point estimates (confidence interval) vary between 0.060 ($\pm$ 0.0001) (Angrist and Krueger 1991, 1930-39 cohort, IV: quarter of birth interacted with years of birth) to 0.167 ($\pm$ 0.0003) (Ashenfelter and Krueger, 1994, Table 3 and page 1169, IV: using sibling's report of the other sibling's education as instrument). See also Card (2001, table II) for a survey. However, those point estimates should be interpreted with caution. In most of these papers, it is interpreted as the return to **1 year** of schooling. However, such an interpretation can be maintained only if we consider a linear IV model (homogeneous treatment effect), i.e., $\Delta(d+1,d) \equiv \theta_{d+1} - \theta_d = \alpha$ for all $d = 0, ..., T-1$. In such a context, a return to a number s of years of schooling is $\Delta(d+s,d) = \alpha \times s$. As discussed earlier, when considering heterogeneous treatment effects, the 2SLS must be interpreted as weighted average of local average treatment effects (LATEs) between different D and Z realizations whenever the LATE assumptions hold. More precisely, in a simple case where Z is binary (single instrument), the LATE assumptions require (i) the potential treatment D_z to be monotone in z, and (ii) the instrument to be statistically independent with the potential variables, i.e., $(Y_0, ..., Y_T, D_1, D_0) \perp Z$.[7] Under these assumptions, Angrist and Imbens (1995) have shown that the 2SLS estimator is a consistent estimator of an estimand that they named the average causal response (ACR):

$$\tilde{\alpha} \equiv \sum_{d=1}^{T} \frac{\mathbb{P}(D_1 \geq d > D_0)}{\sum_{k=1}^{T} \mathbb{P}(D_1 \geq k > D_0)} \mathbb{E}[Y_d - Y_{d-1} | D_1 \geq d > D_0]. \tag{4.1}$$

As can be seen, in the heterogeneous context, the interpretation of the 2SLS changes considerably, as it must now be interpreted as a weighted average treatment effect of each additional year of education for those individuals whose education level has been affected by the instrument. Notice that it is not ensured that this weighted average is over the same group of compliers over each year. It is also worth noting that the assumptions under which 2SLS can consistently estimate $\tilde{\alpha}$ can be too stringent in some cases and their validity can be refuted in many empirical applications. See Kitagawa (2015), Huber and Mellace (2015), and Mourifié and Wan (2017).[8] Even when these assumptions hold, the ACR is still often criticized since in addition to being difficult to interpret, it is not a stable parameter in the sense that its interpretation is instrument-dependent. Also, it depends on who is treated. Heckman and Vytlacil (2005) have proposed to use the local instrumental variable (LIV)

[7] D_z is the counterfactual variable denoting whether the observation would have received treatment if Z had been externally set to z.

[8] All these papers reject the LATE assumptions for the distance to college instrument used by Card (1995).

to identify the marginal treatment effect (MTE) causal parameter, which is defined as the ATE for the subpopulation at the margin. Unlike the LATE, but like the ATE, the MTE is a stable parameter, and easily interpretable even with multiple or continuous instruments. However, as with 2SLS, the LIV can identify the MTE only under the LATE-type of assumptions and can allow to recover non-parametrically the ATE only when the propensity score has full support. Applying the LIV method to the National Longitudinal Survey of Youth (NLSY) dataset, Carneiro, Heckman, and Vytlacil (2011, Table 6) approximate the ATE of *having at least 1 year of college education*. Their estimates varies between 0.0626 ($\pm$ 0.002) and 0.1409 ($\pm$ 0.002).[9] Up to now, the MTE has been mainly developed only for a binary treatment and is not extended to the multivalued treatment case that we consider here yet. Thus, we cannot directly compare Carneiro, Heckman, and Vytlacil (2011) estimates with those using multivalued treatments.

On the other hand, while the bounding strategy can be attractive — since it does not require imposing non-credible or ad-hoc restrictions on the treatment equation, which may not be compatible with individuals' behavior, and also does not suffer from weak instruments issues that are very prevalent in the 2SLS estimation literature — we are not aware of many published papers that propose "reasonable" informative bounds on the average returns to schooling. This could be the result of the two extreme situations discussed earlier in the introduction. In fact, to be able to find informative results using the bounding approach, the applied researcher needs to find an instrument with enough variation, and hope at the same time that the bounds do not cross under the existing assumptions used in the literature so far, i.e., mean, quantile, or full independence. If opting for a stronger specification that leads to point-identification, for instance 2SLS or LIV, the researcher will always be able to report point-estimates even if the model she considers is mis-specified and incompatible with the data generating process. In fact, the latter happens because applied researchers rarely derive or provide specification tests for the general class of models under study. The validity of the model is entirely left to the appreciation of the readers. This context may generate a preference bias for more restricted models even if this may lead to less credible point-estimates.

Nevertheless, the most informative bounds on the average return to schooling using US data we are aware of are the ones proposed by Manksi and Pepper (2000). Using the NLSY 1979 dataset, Manski and Pepper (2000) derived bounds on the yearly average return to

[9]More precisely these estimates are the results of the integration of the MTE over the support [0.0324; 0.9775]. So it only gives an approximation of the ATE that should be an integration over [0,1].

schooling, i.e., $\Delta(d+1,d) = \theta_{d+1} - \theta_d$ under the MTR and the monotone treatment selection (MTS) assumption ($k > l \Rightarrow \mathbb{E}[Y_d|D = k] \geq \mathbb{E}[Y_d|D = l]$). The 95% confidence interval of their bounds on $\Delta(d+1,d)$ varies between $(0; 0.226)$ and $(0; 539)$. The lower bounds are just the result of the MTR assumption, the upper is the more important to analyze in their context. It is worth-noting that while the MTR would be a reasonable assumption to maintain in the returns to schooling literature, the MTS is debatable. As recognized by Manski and Pepper (2000), the MTS assumption fails in cases where ability and taste for schooling are potentially negatively associated, as discussed in Card (1994). A testable implication of the MTR-MTS assumption is that the mean observed wage must be non-decreasing in the number of years of schooling, more precisely $\mathbb{E}[Y|D = d]$ is monotone in d. This testable implication can be tested using existing inference methods in Chetverikov (2013) or Hsu, Liu, and Shi (2018). As can be seen in Figure 1, in our case we observe some decreases. We implement the testing approach proposed by Hsu, Liu and Shi (2018) and reject the MTR-MTS testable implication. In this application, we will not assume either the MTS or the MTR.

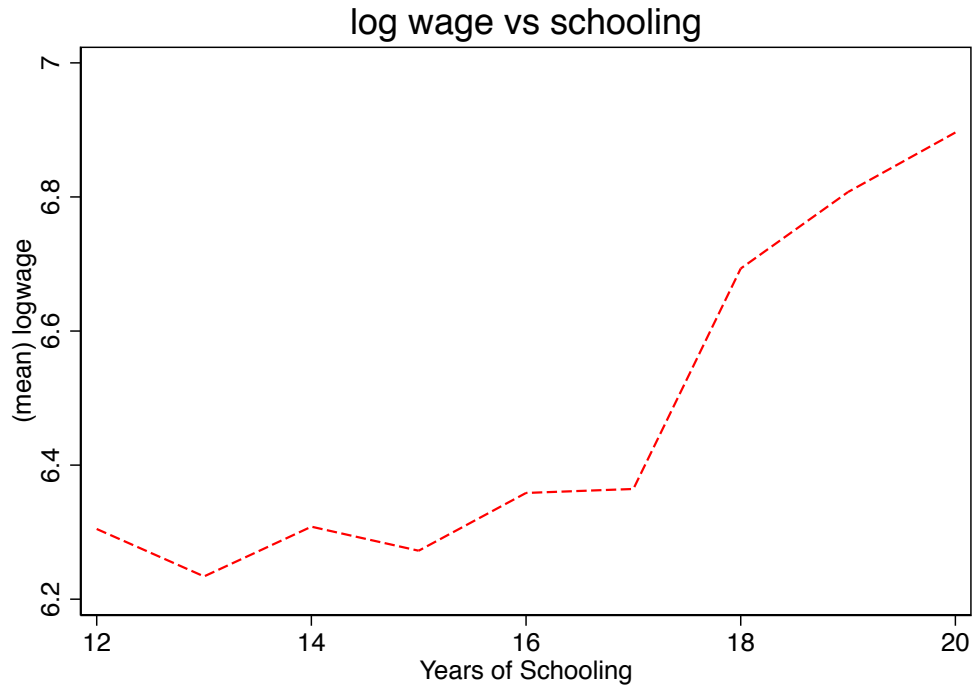


FIGURE 1. Log weekly earning vs highest level of schooling completed.

We recognize that the bounds proposed by Manski and Pepper (2000) may appear not that encouraging for empirical researchers, but this was a very early attempt. Since then we have improved our knowledge on the bounding approach and we will show that we can identify the sign of the treatment effect and obtain relatively reasonably informative bounds on the magnitude of the average returns to schooling using only the OTR assumption and our unconditional moment restrictions.

4.1. **Data and Empirical context.** We consider the data used by Ginther (2000). The data consists of a sample of white, employed males from the 1994 National Longitudinal Survey of Youth Geographic Micro-Data (NLSY). Our main outcome and endogenous variables of interest are the log weekly wage and the highest level of schooling completed, respectively. For a full description of the dataset and the sample under use, please refer to Ginther (2000).

We consider a school-quality characteristic proxied by *teacher-pupil* ratio as a potential instrument, in the sense of the zero-covariance assumption for only the first moment, i.e. $h(Z) = Z$ then $Cov(Y_d, Z) = 0$. This variable has been initially considered as an instrument (respecting the mean independence assumption) by Ginther (2000), motivated by some references therein. When using this variable, she showed that the Manski (1990, 1994) bounds cross, revealing that the mean independence assumption was too stringent for the teacher-pupil ratio instrument. We redo the test using more recent inferential methods like Chernozhukov, Kim, Lee, and Rosen (2015) and find the same result as Ginther (2000). However, notice that the mean independence can be rejected only because of some specific tail dependence between the instrument and the potential outcomes. For instance, as discussed in Dearden et al. (2002), the dependence between the school-quality proxies and the potential outcomes can be explained by at least two main facts: (i) parents with greater interest in their child's education and with higher earnings may choose to live or move to districts with better observed school-quality proxies. These children with such concerned and active parents may benefit from family environments that may enhance their talents, and thus their potential future earnings. We may therefore have people in the upper tail of the potential earnings distribution that are more likely to have been enrolled in schools with high values of school-quality proxies; (ii) Economically disadvantaged populations or regions often receive higher government grants that are often invested to improve school-quality proxies. However, the deprived neighborhoods and environments may negatively affect kids' future potential earnings. Therefore, we may also have people in the lower tail of the potential earnings distribution likely to have been enrolled in schools that have high

values of school-quality proxies. These two facts are consistent with the idea that the invalidity of the mean independence is mainly driven by the observations from the upper or lower tail of the potential earnings distributions and there may not exist clear dependence patterns between middle class potential earnings and school-quality proxies.[10] Usually tail dependence is captured by higher order moments that we will not use here. Interestingly, surveying the large applied literature that studies the relationship between school quality and earnings, Betts (2010) states:

"*The entire body of work appears to agree that the relation between school resources and earnings of adults ranges between none and small but positive. Even the most positive results, based on studies that measure spending per pupil based on each worker's state of birth, suggest an internal rate of return far below the rate of return to an extra year of high school or university, and below the real rate of interest.*" In addition, the few works that find small correlations as in Dearden et al. (2002) did so only for females with low ability.

We therefore think that using only the first moment of the instrument is a reasonable assumption to maintain. In any case, if this assumption is too stringent, our bounds will cross, otherwise we could not reject the validity of this assumption based on the data in hand.

4.1.1. *Methodology.* In our sample under study, the highest level of schooling completed varies between 12 and 20. See Tables 1 and 2 in Appendix B for the summary statistics. We construct the confidence region for the ATEs, $\Delta(12+s,12)$, $s \in \{1,...,8\}$ under the OTR assumption. Because we maintain the OTR assumption, we implement Algorithm 3 in Appendix A with the "clrtest" command and the "local linear" method. We use the default choices of bandwidth and kernel function recommended by Chernozhukov, Kim, Lee, and Rosen (2015, page 31). Since the theoretical support of the log wage is indeed unbounded, and also to avoid the bounding results being too sensitive to outliers, Ginther (2000) proposed to map the observed wage to a trimmed wage in the following way:

$$\tilde{Y}_i = \begin{cases} Q^Y_\tau \text{ if } Y_i \leq Q^Y_\tau, \\ Y_i \text{ if } Q^Y_\tau < Y_i < Q^Y_{1-\tau}, \\ Q^Y_{1-\tau} \text{ if } Y_i > Q^Y_{1-\tau}, \end{cases}$$

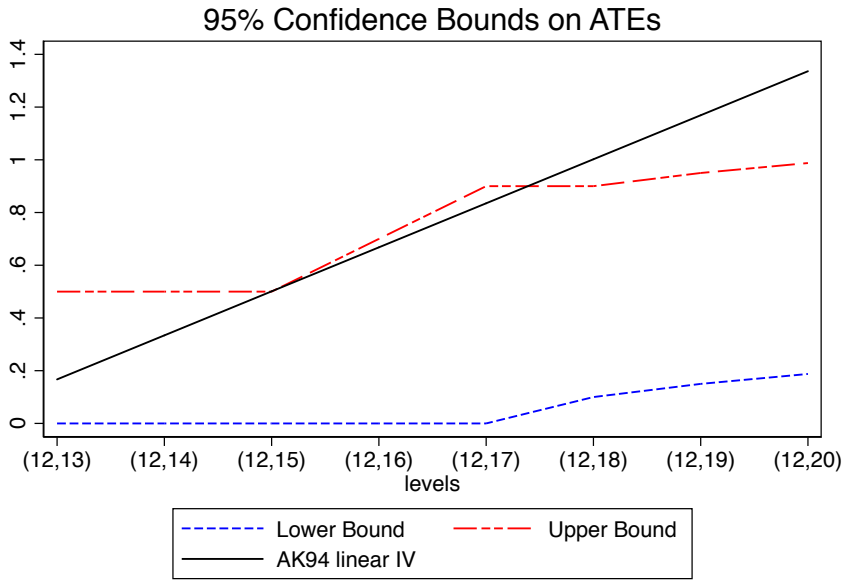
[10]Notice that (i) and (ii) could suggest a U-relationship between the school-quality proxies and the potential earnings. In such a context, we could not pretend that the IV has a monotone effect on the potential earnings. Then we could not use the monotone IV approach developed by Manski and Pepper (2000).

where Q_τ^Y is the τ-quantile of the observed wage.[11] We follow Ginther (2000) and use this same transformation. Figure 2 depicts the 95% confidence regions for the ATEs: $\Delta(12+s,12)$, $s \in \{1,...,8\}$ and this for different levels of trimming: $\tau = 5\%, 10\%$ using the teacher-pupil ratio instrument. The dashed lines represent the lower and upper confidence bounds. The solid line represents the point estimates of the ATEs, $\Delta(12+s,12)$, $s \in \{1,...,8\}$ according the linear IV model used in Ashenfelter and Krueger (1994) discussed earlier.[12] Tables 3 and 4 relegated to Appendix B present the exact values of our confidence regions.

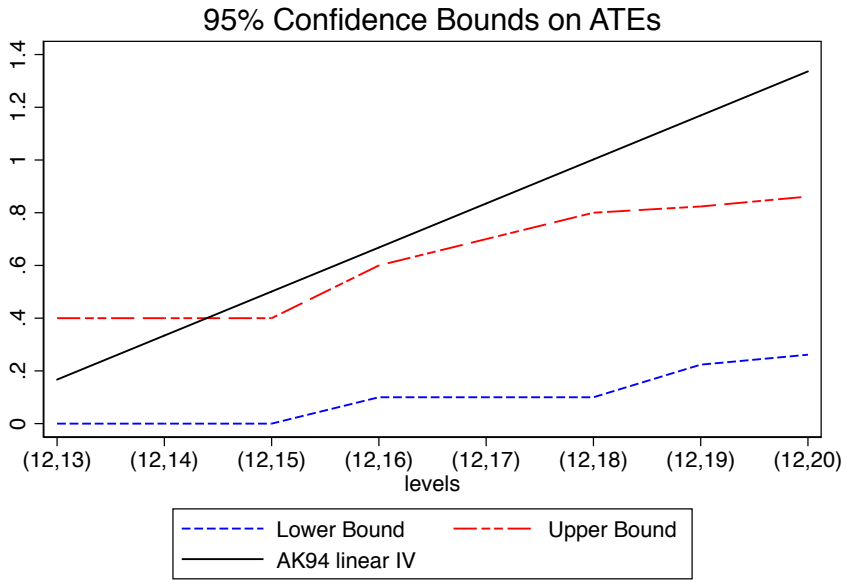
4.1.2. *Results and Interpretation.* The results could be summarized as follows: (i) First, in all cases, the signs of all the ATEs, $\Delta(12+s,12)$, $s \in \{1,...,8\}$ are always identified and non-negative. (ii) Second, while we reject the hypothesis that the ATEs: $\Delta(13,12)$, $\Delta(14,12)$, and $\Delta(15,12)$ are negative, we cannot reject the hypothesis that they are zero. It is worth noting that these treatment effects represent the treatment effect of dropping out of college versus being a high school graduate. These results are not surprising given the shape of the mean observed wage in Figure 1. The mean observed wage does not show any wage advantage for college dropouts even in presence of *a priori* positive selection bias for any additional year of education. In contrast, we observe some small decreases at 13 and 15 years of education compared to 12 years of education. Furthermore, because both the upper and lower bounds for $\Delta(13,12)$, $\Delta(14,12)$, and $\Delta(15,12)$ are almost the same, we have no evidence that dropping out from college at 15 instead of 14 or 13 confers any additional wage returns. However, we observe a clear change in the upper bound at 16 years of education, which corresponds to the college graduates (those who have completed their college diploma) which appears in both Figures 2(a) and 2(b). In the latter figure, the lower bound for $\Delta(16,12)$ even becomes strictly positive. Starting from 16 years of education, both the lower and upper bounds tend to increase at each year, especially in Figure 2(b). These results are consistent with the "*sheepskin effects*" in the returns to education. Based on the screening theory of education, the sheepskin effects in the returns to education suggest that individuals with more schooling tend to earn more not because (or not only because) schooling makes them more productive, but rather because it signals them as more productive. This theory therefore predicts that *potential wages* should rise faster with extra years of education when the extra year also conveys a certificate. See Hungerford and Solon

[11] Lee (2009) also considers a transformation of the observed outcome for his application.

[12] We do not plot the confidence regions for this latter since their estimates are very precise, i.e., $\Delta(d+1,d)$=0.167 ($\pm$ 0.0003) so that the confidence intervals are very close to the line already depicted in the graph.



(a) Trimming Level $\tau = 5\%$



(b) Trimming Level $\tau = 10\%$

FIGURE 2. Bounds on the ATEs.

(1987), Belman and Heywood (1991), and Jaeger and Page (1996) for a detailed discussion on the evidence of sheepskin effects in the US. Notice that, in our case, any additional year of education starting from 16 would confer a certificate to some individuals in the sample depending on the length of a masters degree —that varies between 1 and 2 years— and a Ph.D which would be conferred at least 3 years or more after college graduation, depending on the fields. This could explain why we observe some changes in either the upper or lower bounds at each additional year of schooling after college graduation. (iii) While some evidence of the presence of sheepskin effects in returns to education in US has been discussed in the above cited papers, for some unclear reasons, the empirical literature has mainly focused on the linear IV model which by construction assumes away any potential sheepskin effect. Indeed, unlike the potential outcome model we entertain here that is consistent with all potential non-linearity in the wage equation, the linear IV model by construction does not allow for the possibility of a discontinuity in the functional form at the year of schooling that confers a diploma. This linearity imposes a very strong restriction on the data and may often lead to misleading interpretations. For instance, according to the linear IV estimates of the yearly returns to education we surveyed earlier, the returns to schooling of college dropouts versus high school graduates are estimated to be strictly positive, varying between $\Delta(d+s,12) = 0.060 \times s$ and $\Delta(d+s,12) = 0.167 \times s$ for $s = 1, 2, 3$. However, this may be due to a misspecification issue in which the researcher tries to fit a discontinuous piecewise function where positive jumps appear only on the schooling year that delivers a certificate, as suggested by the screening theory of education. We also see that the magnitude of our bounds are informative enough to reject various linear IV point estimates found in the literature; for instance, the Ashenfelter and Krueger (1994) point estimates fall outside our confidence regions. More precisely, for the data under use, Figure 2(a) (resp. 2(b)) suggests that any linear estimates of the returns to education must be lower than 0.123 (resp. 0.1075). (iv) Finally, it is worth noting that in the presence of sheepskin effects, the ACR estimand must be interpreted with a lot of caution to avoid misleading predictions. Indeed, since it is a weighted average, it could overvalue the causal effect of dropping out of college and undervalue the impact of graduation. Moreover, the weight should be analyzed more carefully than it is commonly done, because if we have more college droppouts than college graduates, the ACR could be less informative (if not non-informative at all) about the causal effect of college graduation and vice-versa. Overall, we think that over-simplified models (like the linear IV model) may rule out by construction some potential relevant economic models, like sheepskin effects, and some over-simplified estimators like 2SLS may not necessarily capture the causal parameter of interest. We think that the

partial identification approach that we entertain here could be general enough to avoid (by construction) ruling-out some relevant theories that are potentially more compatible with the data and at the same time could provide informative bounds to discriminate between various potential specifications.

5. Conclusion

In this paper, we propose a novel approach to derive sharp bounds on various treatment effect parameters using only a finite set of unconditional moment restrictions. To show the sharpness of our bounds, we appeal to the convex analysis literature. From a practical point of view, our method is useful in empirical applications where the commonly invoked IV mean independence restriction is too stringent of a requirement and incompatible with the data. We revisited Ginther's (2000) returns to schooling application using our bounding approach and derive informative bounds on average returns to schooling. Our results are consistent with sheepskin effects in the returns to education literature. However, our sample is not large enough and does not contain enough disaggregated information on the exact year each individual obtained a certificate, which would be required for a deeper analysis of the sheepskin effects using our bounding approach. We therefore leave the full exploration of this question for future research.

Appendix A. Additional Results: Zero-Covariance and OTR.

Consider that the OTR assumption holds and assume that $\mathbb{E}[h(Z)] = 0$. OTR can be viewed as the union of MTR^+ ($d > d' \Rightarrow Y_d \geq Y_{d'}$ a.s) and $MTR^-(d > d' \Rightarrow Y_d \leq Y_{d'}$ a.s). Notice that the MTR^+ implies that for all *non-decreasing* integrable function g(.), we have

$$d > d' \Rightarrow g(Y_d) \geq g(Y_{d'}) \quad a.s. \tag{A.1}$$

For instance, we can consider $g(Y_d) = Y_d$ or $g(Y_d) = 1\{Y_d > y\}$. Imposing MTR^+ mainly affects the previous bounds we derived on the unobserved counterfactuals, i.e., $\mathbb{E}\big[g(Y_d)(1 + \lambda' h(Z))\mathbb{1}[D \neq d]\big]$. To ease the exposition, we use the shorthand notation $\delta \equiv 1 + \lambda' h(Z)$ and then under the MTR^+ we have the following bounds on the unobserved counterfactuals:

$$\begin{aligned} &\mathbb{E}\Big[\mathbb{1}[D > d] \min\{\delta g(Y), \delta \underline{g}_d\} + \mathbb{1}[D < d] \min\{\delta g(Y), \delta \overline{g}_d\}\Big] \\ &\quad \leq \mathbb{E}\big[g(Y_d)\delta \mathbb{1}[D \neq d]\big] \leq \\ &\mathbb{E}\Big[\mathbb{1}[D > d] \max\{\delta g(Y), \delta \underline{g}_d\} + \mathbb{1}[D < d] \max\{\delta g(Y), \delta \overline{g}_d\}\Big]. \end{aligned} \tag{A.2}$$

Then, we have

$$\begin{aligned} &\mathbb{E}\Big[\delta g(Y)\mathbb{1}[D = d] + \mathbb{1}[D > d] \min\{\delta g(Y), \delta \underline{g}_d\} + \mathbb{1}[D < d] \min\{\delta g(Y), \delta \overline{g}_d\}\Big] \\ &\quad \leq \mathbb{E}\big[g(Y_d)\big] \leq \\ &\mathbb{E}\Big[\delta g(Y)\mathbb{1}[D = d] + \mathbb{1}[D > d] \max\{\delta g(Y), \delta \underline{g}_d\} + \mathbb{1}[D < d] \max\{\delta g(Y), \delta \overline{g}_d\}\Big]. \end{aligned} \tag{A.3}$$

for all $\lambda \in \mathbb{R}^m$. Similarly, under MTR^- we can show that

$$\begin{aligned}&\mathbb{E}\Big[\delta g(Y)\mathbb{1}[D=d]+\mathbb{1}[D<d]\min\{\delta g(Y),\delta \underline{g}_d\}+\mathbb{1}[D>d]\min\{\delta g(Y),\delta \overline{g}_d\}\Big]\\ &\qquad\leq \mathbb{E}\big[g(Y_d)\big]\leq \\ &\mathbb{E}\Big[\delta g(Y)\mathbb{1}[D=d]+\mathbb{1}[D<d]\max\{\delta g(Y),\delta \underline{g}_d\}+\mathbb{1}[D>d]\max\{\delta g(Y),\delta \overline{g}_d\}\Big].\end{aligned} \tag{A.4}$$

for all $\lambda \in \mathbb{R}^m$.

Denote by $\tilde{\Theta}^+_{d,d'}$ (resp. $\tilde{\Theta}^-_{d,d'}$) the *outer set* of the joint parameters $(\theta_d, \theta_{d'})$ under Assumptions 1, 2, and MTR^+ (resp. MTR^-). We call them the outer set instead of the identified set because we will not show here their sharpness even if we conjecture that they are sharp. By generalizing the above bounds when $\mathbb{E}[h(Z)] \neq 0$, we can derive the following characterization of the outer sets: For $d > d'$, we have

$$\begin{aligned}\tilde{\Theta}^+_{d,d'} \equiv \Bigg\{&(\theta_d,\theta_{d'})\in[\underline{g}_d,\overline{g}_d]\times[\underline{g}_{d'},\overline{g}_{d'}] \text{ such that } \theta_d\geq\theta_{d'} \text{ and}\\ (*)0\leq &\inf_{(\gamma,\lambda)\in\mathcal{S}_{m+1}}\mathbb{E}\big[(\gamma+\lambda'h(Z))g(Y)\mathbb{1}[D=l]+\mathbb{1}[D>l]\max\{(\gamma+\lambda'h(Z))g(Y),(\gamma+\lambda'h(Z))\underline{g}_l\}\\ &+\mathbb{1}[D<l]\max\{(\gamma+\lambda'h(Z))g(Y),(\gamma+\lambda'h(Z))\overline{g}_l\}-\theta_l(\gamma+\lambda'h(Z))\big],\ \text{for l }\in\{d,d'\}\Bigg\}\end{aligned} \tag{A.5}$$

and

$$\begin{aligned}\tilde{\Theta}^-_{d,d'} \equiv \Bigg\{&(\theta_d,\theta_{d'})\in[\underline{g}_d,\overline{g}_d]\times[\underline{g}_{d'},\overline{g}_{d'}] \text{ such that } \theta_d\leq\theta_{d'} \text{ and}\\ (**)0\leq &\inf_{(\gamma,\lambda)\in\mathcal{S}_{m+1}}\mathbb{E}\big[(\gamma+\lambda'h(Z))g(Y)\mathbb{1}[D=l]+\mathbb{1}[D<l]\max\{(\gamma+\lambda'h(Z))g(Y),(\gamma+\lambda'h(Z))\underline{g}_l\}\\ &+\mathbb{1}[D>l]\max\{(\gamma+\lambda'h(Z))g(Y),(\gamma+\lambda'h(Z))\overline{g}_l\}-\theta_l(\gamma+\lambda'h(Z))\big],\ \text{for l }\in\{d,d'\}\Bigg\}.\end{aligned} \tag{A.6}$$

Therefore, the outer set for $(\theta_d, \theta_{d'})$ under the OTR assumption is

$$\tilde{\Theta}_{d,d} = \Theta^+_{d,d'} \cup \Theta^-_{d,d'}.$$

Notice that we are in presence of what is called an intersection-union test (Berger, 1982) widely used in Bioequivalence hypotheses. Theorem 1 of Berger and Hsu (1996) showed that the construction of $1-\alpha$ confidence regions $CI^+_{n,1-\alpha}(d,d')$, and $CI^-_{n,1-\alpha}(d,d')$ such that $\mathbb{P}(CI^+_{n,1-\alpha}(d,d') \supseteq \Theta^+_{d,d'}) \geq 1-\alpha$ and $\mathbb{P}(CI^-_{n,1-\alpha}(d,d') \supseteq \Theta^-_{d,d'}) \geq 1-\alpha$ ensures to have $\mathbb{P}(CI^+_{n,1-\alpha}(d,d') \cup CI^-_{n,1-\alpha}(d,d') \supseteq \Theta_{d,d'}) \geq 1-\alpha$. In other words, the Theorem 1 of Berger and Hsu (1996) means that if each of the individual tests is performed at level α, then the overall test also has the same level. There is no need for multiplicity adjustment for performing multiple tests. We therefore propose the following algorithm to construct a valid confidence region for $\theta_d - \theta_{d'}$ for $d > d'$.

Algorithm 3. (Implementation using **cmi-test/clrtest**)

Step 1. Draw auxiliary i.i.d. samples $(V_i, W_i)_{i=1}^n$ from distribution $F_{V,W}$ independent of data samples $\{Y_i, Z_i, D_i\}_{i=1}^n$.

Step 2. Let $\phi^+(\theta_d, \theta_{d'}, \alpha, \{Y_i, Z_i, D_i, V_i, W_i\}_{i=1}^n) \in \{0,1\}$ (resp. $\phi^-(\theta_d, \theta_{d'}, \alpha, \{Y_i, Z_i, D_i, V_i, W_i\}_{i=1}^n) \in \{0,1\}$) be the result of the Stata command **cmi-test** or **clrtest** implemented to test the null hypothesis:
$H_0^+ :\ \theta_d \geq \theta_{d'}$ satisfies $(*)$ vs $H_a^+ : \theta_d \geq \theta_{d'}$ doesn't satisfy $(*)$
$\Big($resp. $H_0^- :\ \theta_d \leq \theta_{d'}$ satisfies $(*)$ vs $H_a^- : \theta_d \leq \theta_{d'}$ doesn't satisfy $(**)$ $\Big)$
with nominal level $1-\alpha$ and data observation $\{Y_i, Z_i, D_i, V_i, W_i\}_{i=1}^n\}$.

Step 3. If the null hypothesis H_0 (resp. $\tilde{H}_0$) is rejected, let $\phi^+(\theta_d, \theta_{d'}, \alpha, \{Y_i, Z_i, D_i, V_i, W_i\}_{i=1}^n) = 1$ (resp. $\phi^+(\theta_d, \theta_{d'}, \alpha, \{Y_i, Z_i, D_i, V_i, W_i\}_{i=1}^n) = 1$).

Step 4. Construct the level $1-\alpha$ confidence interval for $(\theta_d - \theta_{d'})$ as follows

$$\mathrm{CI}_{n,1-\alpha}(d,d') \equiv \Big\{\theta_d - \theta_{d'} : (\theta_d, \theta_{d'}) \in [\underline{g}_d, \overline{g}_d] \times [\underline{g}_{d'}, \overline{g}_{d'}],$$
$$\phi^+(\theta_d, \theta_{d'}, \alpha, \{Y_i, Z_i, D_i, V_i, W_i\}_{i=1}^n) = 0 \text{ and } \phi^-(\theta_d, \theta_{d'}, \alpha, \{Y_i, Z_i, D_i, V_i, W_i\}_{i=1}^n) = 0\Big\}.$$

Appendix B. Summary Statistics and Results

Table 1. Summary Statistics

	Total
Observations	874
log wage	6.3886 (0.4649)
years of schooling	14.5709 (2.5851)
Teacher-to-pupil ratio	0.0543 (0.0128)

Average and standard deviation (in the parentheses)

Table 2. Empirical distribution of years of schooling

Years of schooling (D)	Observations	$\mathbb{P}(D=d)$
12	332	0.3799
13	66	0.0755
14	61	0.0698
15	47	0.0538
16	194	0.2220
17	51	0.0584
18	41	0.0469
19	15	0.0172
20	67	0.0767
Total	874	1

Table 3. Confidence bounds on ATEs for $\tau = 5\%$ trimming

ATEs	95% Conf. LB	95% Conf. UB
(12,13)	0	0.5
(12,14)	0	0.5
(12,15)	0	0.5
(12,16)	0	0.7
(12,17)	0	0.9
(12,18)	0.1	0.9
(12,19)	0.15	0.95
(12,20)	0.19	0.99

conf. LB (UB) stands for confidence lower (upper) bound.

TABLE 4. Confidence bounds on ATEs for $\tau = 10\%$ trimming

ATEs	95% Conf. LB	95% Conf. UB
(12,13)	0	0.4
(12,14)	0	0.4
(12,15)	0	0.4
(12,16)	0.1	0.6
(12,17)	0.1	0.7
(12,18)	0.1	0.8
(12,19)	0.22	0.82
(12,20)	0.26	0.86

conf. LB (UB) stands for confidence lower (upper) bound.

APPENDIX C. PROOF OF THE MAIN RESULTS

C.1. **Proof of Theorems 1 and 2.** We will start by proving Theorem 2, and then we will show in C.1.2 that Theorem 1 is a special case of Theorem 2 under the normalization $\mathbb{E}h(Z) = 0$. Before doing so, let's provide a formal definition of the identified set in our context.

Definition 2 (Identified Set)**.** *Given Assumption 1 and 2, define the identified set* Θ_I *of* $(\mathbb{E}[g(Y_0)], ..., \mathbb{E}[g(Y_T)])$ *as the set of all* $(\theta_0, ..., \theta_T)$ *for which there exists random variable* $(G_0, ..., G_T)$ *such that the joint distribution of* $(Y, Z, D, G_0, ..., G_T)$ *satisfy the following conditions:*

(1) $\mathbb{E}G_d = \theta_d$, $\forall d \in \mathcal{D}$,
(2) $\mathbb{E}[(G_d - \theta_d) \cdot h(Z)] = 0$ $\forall d \in \mathcal{D}$,
(3) $\mathbb{P}(G_d = g(Y)|D = d) = 1$ *and* $\mathbb{P}(G_d \in \Gamma_d | D \neq d) = 1$, *where* $\Gamma_d \equiv \text{Supp}(g(Y_d)|D = d)$, *and* $\underline{g}_d \equiv \inf \Gamma_d, \overline{g}_d \equiv \sup \Gamma_d$, $\forall d \in \mathcal{D}$.

C.1.1. *Proof of Theorem 2.* We first show $(\theta_0, ..., \theta_T) \in \Theta_I$ implies inequality (2.11) for each $d \in \mathcal{D}$. Let $(\theta_0, ..., \theta_T)$ be any point in Θ_I, and let $(G_0, ..., G_T)$ be random variables satisfying Condition (1)-(3) in Definition 2. Condition (1) and (2) imply, for any $d \in \mathcal{D}$, any $(\gamma, \lambda) \in \mathcal{S}_{m+1}$

$$0 = \mathbb{E}(G_d - \theta_d)(\gamma + \lambda' h(Z)).$$

Combining the above equality with Condition (3), we know that:

$$0 \leq \inf_{(\gamma,\lambda)\in\mathcal{S}_{m+1}} \mathbb{E}\Big[\mathbb{1}(D = d)(g(Y) - \theta_d)(\gamma + \lambda' h(Z)) + \mathbb{1}(D \neq d) \sup_{y\in\Gamma_d} \Big\{(y - \theta_d)(\gamma + \lambda' h(Z))\Big\}\Big]$$

By the definition of $\underline{g}_d$ and $\overline{g}_d$, we can rewrite the above inequality as

$$0 \leq \inf_{(\gamma,\lambda)\in\mathcal{S}_{m+1}} \mathbb{E}\left[\mathbb{1}(D = d)(g(Y) - \theta_d)(\gamma + \lambda' h(Z)) + \mathbb{1}(D \neq d) \max_{y\in\{\underline{g}_d, \overline{g}_d\}} \Big\{(y - \theta_d)(\gamma + \lambda' h(Z))\Big\}\right].$$

which is equivalent to inequality (2.11).

Next, we show when $(\theta_0, ..., \theta_T)$ satisfies inequality (2.11), then $(\theta_0, ..., \theta_T) \in \Theta_I$. Let $(\theta_0, ..., \theta_T)$ be a point satisfying inequality (2.11) for $d \in \mathcal{D}$. Then, we want to construct random variables $(G_0, ..., G_T)$ which satisfy all conditions in Definition 2. To do so, define

$$\begin{aligned} \varphi_d(Y, D, Z; \gamma, \lambda) &\equiv \overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(\gamma + \lambda' h(Z)) \\ \varphi_d(\gamma, \lambda) &\equiv \mathbb{E}\varphi_d(Y, D, Z; \gamma, \lambda). \end{aligned}$$

Recall that

$$\overline{f}_d(Y,D,\gamma+\lambda'h(Z)) \equiv \mathbb{1}[D=d](\gamma+\lambda'h(Z))g(Y) + \mathbb{1}[D\neq d]\max\Big((\gamma+\lambda'h(Z))\underline{g}_d, (\gamma+\lambda'h(Z))\overline{g}_d\Big).$$

Using these notation, inequality (2.11) can be written as $0 \leq \inf\{\varphi_d(\gamma,\lambda) : (\gamma,\lambda) \in \mathcal{S}_{m+1}\}$. Note that $\varphi_d(Y,D,Z;\gamma,\lambda)$ is linearly homogeneous in (γ,λ). That is, for any $\alpha > 0$, we have

$$\varphi_d(Y,D,Z;\alpha\gamma,\alpha\lambda) = \alpha\cdot\varphi_d(Y,D,Z;\gamma,\lambda).$$

Therefore, for any $(\gamma,\lambda)\in\mathbb{R}^{m+1}$ with $0<\|(\gamma,\lambda)\|$, we have

$$\begin{aligned}\varphi_d(\gamma,\lambda)\geq 0 \quad &\Leftrightarrow \quad \|(\gamma,\lambda)\|\,\varphi_d\left(\frac{\gamma}{\|(\gamma,\lambda)\|},\frac{\lambda}{\|(\gamma,\lambda)\|}\right)\geq 0\\ &\Leftrightarrow \quad \varphi_d\left(\frac{\gamma}{\|(\gamma,\lambda)\|},\frac{\lambda}{\|(\gamma,\lambda)\|}\right)\geq 0,\end{aligned}$$

Since $\left(\frac{\gamma}{\|(\gamma,\lambda)\|},\frac{\lambda}{\|(\gamma,\lambda)\|}\right)\in\mathcal{S}_{m+1}$, we know that $0\leq\inf\{\varphi_d(\gamma,\lambda):(\gamma,\lambda)\in\mathcal{S}_{m+1}\}$ is equivalent to $0\leq\inf\{\varphi_d(\gamma,\lambda):(\gamma,\lambda)\in\mathbb{R}^{m+1}\setminus\{0\}\}$. Since $\varphi_d(0,0)=0$ for any θ_d, we also know that $0\leq\inf\{\varphi_d(\gamma,\lambda):(\gamma,\lambda)\in\mathcal{S}_{m+1}\}$ is equivalent to $0\leq\inf\{\varphi_d(\gamma,\lambda):(\gamma,\lambda)\in\mathbb{R}^{m+1}\}$.

Now, notice that $\varphi_d(\gamma,\lambda)$ is a convex function of (γ,λ) mapping from $\mathbb{R}^{m+1}$ to $\mathbb{R}$, so that its subgradient always exists. Also, since inequality (2.11) is equivalent to $0\leq\inf\{\varphi_d(\gamma,\lambda):(\gamma,\lambda)\in\mathbb{R}^{m+1}\}$ and $\varphi_d(0,0)=0$, the infimum is always achieved at $\gamma=0$ and $\lambda=0$. This implies $0\in\partial\varphi_d(0,0)$, where $\partial\varphi_d(0,0)$ stands for the partial subgradient of φ_d at point $\gamma=0$ and $\lambda=0$. Let $\partial\varphi_d(Y,D,Z;0,0)$ denotes the subgradient of $\varphi_d(Y,D,Z;\gamma,\lambda)$ at point $\gamma=0$ and $\lambda=0$. Then, Proposition 2.2 in Bertsekas (1973) and $0\in\partial\varphi_d(0,0)$ implies, there exists a measurable function $\psi_d(\cdot)$ such that $\mathbb{E}\psi_d(Y,D,Z)=0$ and $\psi_d(Y,D,Z)\in\partial\varphi_d(Y,D,Z;0,0)$ almost surely. We can show that

$$\partial\varphi_d(Y,D,Z;0,0) = \big\{[\mathbb{1}(D=d)g(Y)+\mathbb{1}(D\neq d)x-\theta_d]\cdot(1,h(Z))' : x\in[\underline{g}_d,\overline{g}_d]\big\}$$

where $(1,h(Z))$ stands for a vector in $\mathbb{R}^{m+1}$ whose first element is 1 and the rest element is $h(Z)$.

Let $\psi_d^{(1)}(Y,D,Z)$ denote the first dimension in $\psi_d(Y,D,Z)$. Let Q be a random variable, distributed as uniform distribution in $[0,1]$ and is independent of (Y,D,Z). Construct function $G_d(Y,D,Z,Q)$ as the following

$$G_d(Y,D,Z,Q) = \begin{cases} g(Y) & \text{if } D=d,\\ \overline{g}_d & \text{if } D\neq d, \text{ and } Q\leq(\overline{g}_d-\underline{g}_d)^{-1}\big(\psi_d^{(1)}(Y,D,Z)+\theta_d-\underline{g}_d\big)\\ \underline{g}_d & \text{if otherwise}\end{cases} \tag{C.1}$$

and then define $G_d(Y,D,Z)\equiv\mathbb{E}[G_d(Y,D,Z,Q)|Y,D,Z]$.

By the construction of $G_d(Y,D,Z)$, we have Condition (3) in Definition 2 satisfied and it can be shown that

$$(G_d(Y,D,Z)-\theta_d)\cdot(1,h(Z))' = \psi_d(Y,D,Z) \quad \text{almost surely.}$$

Notice that since $\psi_d(Y,D,Z)=g(Y)$ when $D=d$, the above equality obviously holds when $D=d$. To see that it also holds when $D\neq d$, remark that the independence between Q and (Y,D,Z) implies that $(G_d(Y,D,Z)-\theta_d)=\psi_d^{(1)}(Y,D,Z)$ when $D\neq d$. Moreover, since $\psi_d(Y,D,Z)\in\partial\varphi_d(Y,D,Z;0,0)$ we know that $\psi_d(Y,D,Z)=\psi_d^{(1)}(Y,D,Z)\cdot(1,h(Z))'$ when $D\neq d$.

Therefore since $\mathbb{E}\psi_d(Y,D,Z)=0$, we have $\mathbb{E}G_d(Y,D,Z)=\theta_d$ and $\mathbb{E}(G_d(Y,D,Z)-\theta_d)h(Z)=0$, which then implies Condition (1) and (2) in Definition 2. As a result, we know $(\theta_0,...,\theta_T)\in\Theta_I$. This completes the proof.

C.1.2. *Proof of Theorem 1.* In this section, we prove Theorem 1 using results in Theorem 2. To do so, we show inequality (2.11) is equivalent to $\underline{\theta}_d \leq \theta_d \leq \overline{\theta}_d$, when $\mathbb{E}h(Z) = 0$.

When $\mathbb{E}h(Z) = 0$, inequality (2.11) can be rewritten as

$$0 \leq \inf_{(\gamma,\lambda)\in\mathcal{S}_{m+1}} \mathbb{E}\big[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \gamma\theta_d\big] \tag{C.2}$$

Since $\mathbb{E}\big[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \gamma\theta_d\big]$ is a continuous function of (γ, λ), inequality (C.2) holds if and only the following two inequalities hold

$$0 \leq \inf_{\gamma>0,(\gamma,\lambda)\in\mathcal{S}_{m+1}} \mathbb{E}\big[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \gamma\theta_d\big] \tag{C.3}$$

$$0 \leq \inf_{\gamma<0,(\gamma,\lambda)\in\mathcal{S}_{m+1}} \mathbb{E}\big[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \gamma\theta_d\big] \tag{C.4}$$

Also note that, for any $\gamma > 0$,

$$\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \gamma\theta_d = \gamma\left[\overline{f}_d(Y, D, 1 + \gamma^{-1}\lambda' h(Z)) - \theta_d\right] \quad \text{almost surely.}$$

Similarly, for any $\gamma < 0$,

$$\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \gamma\theta_d = \gamma\left[\underline{f}_d(Y, D, 1 + \gamma^{-1}\lambda' h(Z)) - \theta_d\right] \quad \text{almost surely.}$$

Let $\tilde{\lambda} = \gamma^{-1}\lambda$, inequality (C.3) is equivalent to

$$0 \leq \inf_{\tilde{\lambda}} \mathbb{E}\left[\overline{f}_d(Y, D, 1 + \tilde{\lambda}' h(Z)) - \theta_d\right]$$

which is equivalent to $\theta_d \leq \overline{\theta}_d$. Similarly, inequality (C.4) is equivalent to

$$0 \geq \sup_{\tilde{\lambda}} \mathbb{E}\left[\underline{f}_d(Y, D, 1 + \tilde{\lambda}' h(Z)) - \theta_d\right]$$

which is equivalent to $\theta_d \geq \underline{\theta}_d$.

C.2. **Proof for Proposition 1.** First of all, note that inequality (2.11) is equivalent to

$$0 \leq \inf_{\gamma\in\mathbb{R},\lambda\in\mathbb{R}^m} \mathbb{E}\big[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(\gamma + \lambda' h(Z)).\big] \tag{C.5}$$

Hence, to show Λ is core-determining, we only need to show

$$0 \leq \inf_{(\gamma,\lambda)\in\Lambda} \mathbb{E}\big[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(\gamma + \lambda' h(Z))\big]$$

is equivalent to inequality (C.5).

C.3. **When Supp$(h(Z))$ is finite.** Let k be the number of points in Supp$(h(Z))$, and Supp$(h(Z)) = \{h_1, ..., h_k\}$. Fix any $d \in \{0, 1\}$. For $\alpha \in \{-1, 1\}^k$, define $LP(\alpha)$ as

$$\begin{aligned} LP(\alpha) = \quad & \inf_{\gamma,\lambda} \quad \mathbb{E}\big[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(\gamma + \lambda' h(Z))\big] \\ & \text{s.t.} \qquad \alpha_i(\gamma + \lambda' h_i) \geq 0, \quad \forall i = 1, ..., k \end{aligned}$$

where α_k is k-th element in vector α.

Since

$$\inf_{\gamma\in\mathbb{R},\lambda\in\mathbb{R}^m} \mathbb{E}\big[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(\gamma + \lambda' h(Z))\big] = \min\{LP(\alpha) : \alpha \in \{-1, 1\}^k\}, \tag{C.6}$$

inequality (C.5) is equivalent to for any $\alpha \in \{-1,1\}^k$, $LP(\alpha) \geq 0$. Therefore, we only need to show, for any $\alpha \in \{-1,1\}^k$,

$$\begin{aligned} 0 > LP(\alpha) \Leftrightarrow 0 > \quad \inf_{\gamma,\lambda} \quad & \mathbb{E}\big[\overline{f}_d(Y,D,\gamma+\lambda' h(Z)) - \theta_d(\gamma+\lambda' h(Z))\big] \\ \text{s.t.} \quad & \alpha_k(\gamma+\lambda' h_i) \geq 0 \quad i = 1,...,k \\ & (\gamma,\lambda) \in \Lambda \end{aligned}$$

Let $p_i = \mathbb{P}(h(Z) = h_i | D \neq d)$. Then, one can rewrite $LP(\alpha)$ as the following linear programming problem,

$$\begin{aligned} LP(\alpha) = \quad \inf_{\gamma,\lambda} \quad & c_{\gamma,\alpha}\gamma + c'_{\lambda,\alpha}\lambda \\ \text{s.t.} \quad & \alpha_i(\gamma+\lambda' h_i) \geq 0, \quad \forall i = 1,...,k \end{aligned}$$

where

$$\begin{aligned} c_{\gamma,\alpha} \quad &\equiv \quad \mathbb{E}\big[\mathbb{1}(D=d)(g(Y)-\theta_d)\big] + \sum_{i=1}^k p_i\big(\mathbb{1}(\alpha_i=1)\overline{g}_d + \mathbb{1}(\alpha_i=-1)\underline{g}_d - \theta_d\big) \\ c_{\lambda,\alpha} \quad &\equiv \quad \mathbb{E}\big[\mathbb{1}(D=d)(g(Y)-\theta_d)h(Z)\big] + \sum_{i=1}^k p_i\big(\mathbb{1}(\alpha_i=1)\overline{g}_d + \mathbb{1}(\alpha_i=-1)\underline{g}_d - \theta_d\big)h_i \end{aligned}$$

Therefore, $LP(\alpha) < 0$ if and only if $LP(\alpha) = -\infty$.

By Assumption 3 and Theorem 4.12 in Bertsimas and Tsitsiklis (1997), the polyhedral cone $\{(\gamma,\lambda) : \alpha_i(\gamma + \lambda' h_i) \geq 0, \forall i = 1,...,k\}$ is pointed (i.e. the zero vector is an extreme point of the polyhedral cone). By Theorem 4.13 in Bertsimas and Tsitsiklis (1997), $LP(\alpha) < 0$ if and only if there exists a nonzero (γ^*,λ^*), such that (i) $c_{\gamma,\alpha}\gamma^* + c'_{\lambda,\alpha}\lambda^* < 0$, (ii) $\alpha_i(\gamma^* + \lambda^{*\prime} h_i) \geq 0$, $\forall i = 1,...,m$ and (iii) there exists m linearly independent vectors h_{τ_1},..., h_{τ_m} in $\{h_1,...,h_k\}$ which satisfies $\gamma + \lambda' h_{\tau_j} = 0$ for all $j = 1,...,m$. Now, Let $(\tilde{\gamma},\tilde{\lambda}) = (\gamma,\lambda)/\left\|(\gamma,\lambda)\right\|$. Then, we have $c_{\gamma,\alpha}\tilde{\gamma} + c'_{\gamma,\alpha}\tilde{\lambda} < 0$ and $(\tilde{\gamma},\tilde{\lambda}) \in \Lambda$.

This implies, for an arbitrary $\alpha \in \{-1,1\}^k$, $LP(\alpha) < 0$ if and only if there exists some $(\gamma,\lambda) \in \Lambda$ such that $c_{\gamma,\alpha}\gamma + c_{\lambda,\alpha}\lambda < 0$, or equivalently, $\mathbb{E}[\overline{f}_d(Y,D,\gamma - \lambda' h(Z)) - \theta_d(\gamma + \lambda' h(Z))] < 0$. This result together with (C.6) proves Λ is a core-determining class.

C.4. **When Supp$(h(Z))$ is a convex hull of a finite set.** Fix $d \in \mathcal{D}$ and θ_d. Define $\Psi(\gamma,\lambda) \equiv \mathbb{E}[\overline{f}_d(Y,D,\gamma + \lambda' h(Z)) - \theta_d(\gamma + \lambda' h(Z))]$. Then, inequality (C.5) is equivalent to $0 \leq \inf_{\gamma,\lambda} \Psi(\gamma,\lambda)$.

To show Λ is core-determining, we only need to show that if $\Psi(\gamma,\lambda) < 0$ for some $(\gamma,\lambda) \in \mathcal{S}_{m+1}\backslash\Lambda$, then there exists some $(\tilde{\gamma},\tilde{\lambda}) \in \Lambda$ such that $\Psi(\tilde{\gamma},\tilde{\lambda}) < 0$.

For any set A, let $|A|$ be the cardinality of set A. We first show that for any $(\gamma,\lambda) \in \mathcal{S}_{m+1}\backslash\Lambda$, one of the two conditions must hold: (i) for all $h \in \text{Supp}(h(Z))$, $\gamma + \lambda' h \geq 0$; (ii) for all $h \in \text{Supp}(h(Z))$, $\gamma + \lambda' h \leq 0$. To see why it is so, note that Assumption 3 implies, there exists $m+1$ vectors $h_1,...,h_{m+1}$ in Supp$(h(Z))$ such that $(1,h_1),...,(1,h_{m+1})$ are linearly independent. Let $H^* \equiv \{h_1,...,h_{m+1}\}$. For any $(\gamma,\lambda) \in \mathcal{S}_{m+1}\backslash\Lambda$, let $H^+ \equiv \{h \in H^* : \gamma + \lambda' h > 0\}$, let $H^- \equiv \{h \in H^* : \gamma + \lambda' h < 0\}$ and let $H^0 \equiv \{h \in H^* : \gamma + \lambda' h_i = 0\}$. Now, suppose the claim is not true, i.e. $|H^+| \geq 1$ and $|H^-| \geq 1$. Pick any $h_+ \in H^+$ and $h_- \in H^-$, and construct $\tilde{H}^-$ and $\tilde{H}^+$ as

$$\begin{aligned} \tilde{H}^+ \quad &\equiv \quad \left\{ \frac{\gamma + \lambda' h}{\lambda'(h - h_-)} h_- + \frac{-\gamma - \lambda' h_-}{\lambda'(h - h_-)} h : h \in H^+ \backslash \{h_+\} \right\} \\ \tilde{H}^+ \quad &\equiv \quad \left\{ \frac{\gamma + \lambda' h_+}{\lambda'(-h + h_+)} h + \frac{-\gamma - \lambda' h}{\lambda'(-h + h_+)} h_+ : h \in H^- \right\}. \end{aligned}$$

By construction, for any $h \in \tilde{H}^+ \cup \tilde{H}^-$, we have $\gamma + \lambda' h = 0$. Moreover, $|\tilde{H}^+| + |\tilde{H}^-| + |H^0| = m$. Since $\text{Supp}(h(Z))$ is convex, we have $\tilde{H}^+ \cup \tilde{H}^- \subseteq \text{Supp}(h(Z))$. Finally, that $(1, h_1), ..., (1, h_{m+1})$ are linearly independent implies $\{(1, h) : h \in \tilde{H}^+ \cup \tilde{H}^- \cup H^0\}$ is also linearly independent. This contradicts to $(\gamma, \lambda) \in \mathcal{S}_{m+1} \backslash \Lambda$.

Now, suppose $\Psi(\gamma^*, \lambda^*) < 0$ for some $(\gamma^*, \lambda^*) \in \mathcal{S}_{m+1} \backslash \Lambda$. By the previous result, we know either (a) for all $h \in \text{Supp}(h(Z))$, $\gamma^* + \lambda^{*\prime} h \geq 0$, or (b) for all $h \in \text{Supp}(h(Z))$, $\gamma^* + \lambda^{*\prime} h \leq 0$. Let's assume, without loss of generality, that (a) is the case. Then, $\Psi(\gamma^*, \lambda^*) < 0$ implies

$$\begin{aligned} 0 > \quad & \inf \quad \Psi(\gamma, \lambda) \\ & \text{s.t.} \quad \gamma + \lambda' h \geq 0, \forall h \in \text{Supp}(h(Z)) \end{aligned}$$

Moreover, one can show that when $\gamma + \lambda' h \geq 0, \forall h \in \text{Supp}(h(Z))$,

$$\Psi(\gamma, \lambda) = \gamma \mathbb{E}[\mathbb{1}(D = d) g(Y) + \mathbb{1}(D \neq d) \overline{g}_d - \theta_d] + \lambda' \mathbb{E}[(\mathbb{1}(D = d) g(Y) + \mathbb{1}(D \neq d) \overline{g}_d - \theta_d) h(Z)].$$

Therefore, we have

$$\begin{aligned} 0 > \quad & \inf \quad \gamma \mathbb{E}[\mathbb{1}(D = d) g(Y) + \mathbb{1}(D \neq d) \overline{g}_d - \theta_d] + \lambda' \mathbb{E}[(\mathbb{1}(D = d) g(Y) + \mathbb{1}(D \neq d) \overline{g}_d - \theta_d) h(Z)] \\ & \text{s.t.} \quad \gamma + \lambda' h \geq 0, \forall h \in \text{Supp}(h(Z)). \end{aligned}$$

Since $\text{Supp}(h(Z))$ is a convex hull of a finite set, we can write $\text{Supp}(h(Z))$ as $\text{Supp}(h(Z)) = \text{co}\{h_1^*, ..., h_k^*\}$, where co denote the convex hull. Then,

$$\gamma + \lambda' h \geq 0, \forall h \in \text{Supp}(h(Z)) \Leftrightarrow \gamma + \lambda' h_i^* \geq 0, \forall i = 1, ..., k.$$

Therefore, we have

$$\begin{aligned} 0 > \quad & \inf \quad \gamma \mathbb{E}[\mathbb{1}(D = d) g(Y) + \mathbb{1}(D \neq d) \overline{g}_d - \theta_d] + \lambda' \mathbb{E}[(\mathbb{1}(D = d) g(Y) + \mathbb{1}(D \neq d) \overline{g}_d - \theta_d) h(Z)] \\ & \text{s.t.} \quad \gamma + \lambda' h_i^* \geq 0, \forall i \in \{1, ..., k\}. \end{aligned}$$

Theorem 4.13 in Bertsimas and Tsitsiklis (1997), this implies there exists a nonzero $(\tilde{\lambda}, \tilde{\gamma})$ such that $|\{i \in \{1, ..., k\} : \tilde{\gamma} + \tilde{\lambda}' h_i^* = 0\}| = m$ and

$$0 > \tilde{\gamma} \mathbb{E}[\mathbb{1}(D = d) g(Y) + \mathbb{1}(D \neq d) \overline{g}_d - \theta_d] + \tilde{\lambda}' \mathbb{E}[(\mathbb{1}(D = d) g(Y) + \mathbb{1}(D \neq d) \overline{g}_d - \theta_d) h(Z)].$$

This proves that $\inf\{\Psi(\gamma, \lambda) : (\gamma, \lambda) \in \Lambda\} < 0$. Hence, Λ is a core determining class.

Appendix D. Asymptotic Properties of the Confidence Regions

D.1. **Proofs of the Asymptotic validity of the Algorithm 1.** We are going to show the inference intervals proposed in (3.4) and (3.5) have correct asymptotic converge probability and are consistent against all fixed alternatives. Define $\mathcal{F}_d$ as the set of all (θ_d, F) pairs in which θ_d satisfies inequality (2.11), and define $\mathcal{F}_d^\dagger$ as the set of all (θ_d, F) pairs in which inequality (2.11) does not hold. Formally, for any given constants $\delta > 0$ and $C_1 < \infty$, we define $\mathcal{F}_d$ and $\mathcal{F}_d^\dagger$ as follows.

Definition 3. *Define $\mathcal{F}_d$ as the set of all (θ_d, F) pairs such that (i) $\{(Y_i, Z_i, D_i) : i \geq 1\}$ are i.i.d. under F; (ii) $\mathbb{E}_F \|h(Z)\|^{2+\delta} < C_1$; (iii) $\theta_d \in [\underline{g}_d, \overline{g}_d]$ and θ_d satisfies inequality* (2.11) *where the expectation is taken with respect to F, i.e.*

$$0 \leq \inf_{(\gamma, \lambda) \in \mathcal{S}_{m+1}} \mathbb{E}_F \Big[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(\gamma + \lambda' h(Z)) \Big]. \tag{D.1}$$

Define $\mathcal{F}_d^\dagger$ as the set of all (θ_d, F) pairs such that (i) $\{(Y_i, Z_i, D_i) : i \geq 1\}$ are i.i.d. under F; (ii) $\mathbb{E}_F |h(Z)|^{2+\delta} < C_1$; (iii) $\theta_d \in [\underline{g}_d, \overline{g}_d]$ and inequality (D.1) *does not hold.*

Proposition 2. *Suppose Assumption 1 hold. For any $d \in \mathcal{D}$ and any $(\theta_d, F) \in \mathcal{F}_d$*

$$\liminf_{n\to\infty} \mathbb{P}_F\Big(\theta_d \in CI_{n,1-\alpha}(d)\Big) \geq 1-\alpha. \tag{D.2}$$

Moreover, for any $(\theta_d, F) \in \mathcal{F}_d^\dagger$,

$$\lim_{n\to\infty} \mathbb{P}_F\Big(\theta_d \in CI_{n,1-\alpha}(d)\Big) = 0.$$

Definition 4. *Define $\mathcal{F}_{d_1,d_2}$ as the set of all $(\theta_{d_1}, \theta_{d_2}, F)$ pairs such that (i) $\{(Y_i, Z_i, D_i) : i \geq 1\}$ are i.i.d. under F; (ii) $\mathbb{E}_F \|h(Z)\|^{2+\delta} < C_1$; (iii) for each $d \in \{d_1, d_2\}$, $\theta_d \in [\underline{g}_d, \overline{g}_d]$ and*

$$\forall d \in \{d_1, d_2\}, \quad 0 \leq \inf_{(\gamma,\lambda)\in\mathcal{S}_{m+1}} \mathbb{E}_F\Big[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(\gamma + \lambda' h(Z))\Big]. \tag{D.3}$$

Proposition 3. *Suppose Assumption 1 hold. For any $d, d' \in \mathcal{D}$ with $d \neq d'$ and any $(\theta_d, \theta_{d'}, F) \in \mathcal{F}_{d,d'}$, we have*

$$\liminf_{n\to\infty} \mathbb{P}_F(\theta_d - \theta_{d'} \in CI_{n,1-\alpha}(d, d')) \geq 1-\alpha. \tag{D.4}$$

Inequality (D.2) and (D.4) mean that $\mathrm{CI}_{n,1-\alpha}(d)$ and $\mathrm{CI}_{n,1-\alpha}(d, d')$ have correct asymptotic level for any given DGP. Now, we will show that $\mathrm{CI}_{n,1-\alpha}(d)$ and $\mathrm{CI}_{n,1-\alpha}(d, d')$ also have correct uniform asymptotic level for a family of DGPs.

Recall set Λ is the pre-specified subset of $\mathcal{S}_{m+1}$ in the definition of $CI_{n,1-\alpha}(d)$. Define $h_{d,F}(\theta_d) \equiv \Big\{h_{d,F}(\theta_d, \gamma, \lambda, \gamma^\dagger, \lambda^\dagger) : (\gamma, \lambda), (\gamma^\dagger, \lambda^\dagger) \in \Lambda\Big\}$ as the asymptotic variance-covariance kernel of $n^{1/2}\overline{m}_n(\theta_d, d, \gamma, \lambda)$. That is, for any (γ, λ), $(\gamma^\dagger, \lambda^\dagger) \in \Lambda$,

$$h_{d,F}(\theta_d, \gamma, \lambda, \gamma^\dagger, \lambda^\dagger) = \mathrm{cov}_F\Big(\overline{f}_d(Y_i, D_i, \gamma + \lambda' h(Z_i)) - \theta_d(\gamma + \lambda' h(Z_i)),$$
$$\overline{f}_d(Y_i, D_i, \gamma^\dagger + \lambda^{\dagger\prime} h(Z_i)) - \theta_d(\gamma^\dagger + \lambda^{\dagger\prime} h(Z_i))\Big)$$

Define the set of variance-covariance kernels:

$$\mathcal{H}_d \equiv \{h_{d,F}(\theta_d) : (\theta_d, F) \in \mathcal{F}_d\}.$$

On the space of all continuous functions defined on $\Lambda \times \Lambda$, which is a superset of $\mathcal{H}_d$, we consider the topology generated by the uniform metric:

$$\rho(h, h^\dagger) \equiv \sup_{(\gamma,\lambda),(\gamma^\dagger,\lambda^\dagger)\in\Lambda} |h(\gamma, \lambda, \gamma^\dagger, \lambda^\dagger) - h^\dagger(\gamma, \lambda, \gamma^\dagger, \lambda^\dagger)|,$$

where h and $h^\dagger$ are two arbitrary continuous functions on $\Lambda \times \Lambda$. The following proposition is implied by Theorem 5.1 and Theorem 6.1 in Andrews and Shi (2017). Moreover, Proposition 2 is an immediate corollary of the following proposition.

Proposition 4. *For any $d \in \mathcal{D}$ and any compact subset $\mathcal{H}_{cpt}$ of $\mathcal{H}_d$, we have*

$$\liminf_{n\to\infty} \inf_{\substack{(\theta_d,F)\in\mathcal{F}_d: \\ h_{d,F}(\theta_d)\in\mathcal{H}_{cpt}}} \mathbb{P}_F\Big(\theta_d \in CI_{n,1-\alpha}(d)\Big) \geq 1-\alpha.$$

Moreover, for any $(\theta_d, F) \in \mathcal{F}_d^\dagger$,

$$\lim_{n\to\infty} \mathbb{P}_F\Big(\theta_d \in CI_{n,1-\alpha}(d)\Big) = 0.$$

Proof. Fix an arbitrary $d \in \mathcal{D}$ and let Λ be the subset of $\mathcal{S}_{m+1}$ in (3.4). Define $m(Y, D, Z, \theta_d, \gamma, \lambda) \equiv \overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(\gamma + \lambda' h(Z))$, and define $\overline{m}(Y, D, Z)$ as

$$\overline{m}(Y, D, Z) \equiv (|\underline{g}_d| + |\overline{g}_d|) \cdot \|(1, h(Z))\|$$

where $\|\cdot\|$ stands for the Euclidean norm in $\mathbb{R}^{m+1}$ and $(1, h(Z))$ stands for the vector in $\mathbb{R}^{m+1}$ whose first coordinate is 1 and the rest equals $h(Z)$. The Cauchy-Schwarz inequality implies

$$\forall(\gamma, \lambda) \in \Lambda,\ \forall \theta_d \in [\underline{g}_d, \overline{g}_d], \quad |m(Y, D, Z, \theta_d, \gamma, \lambda)| \leq \overline{m}(Y, D, Z).$$

Define $\mathcal{F}_+ \equiv \mathcal{F}_d \cup \mathcal{F}_d^\dagger$. By the definition of $\mathcal{F}_d$ and $\mathcal{F}_d^\dagger$, there exists some $C_2 < \infty$ (which only depends on δ and C_1 in the definition of $\mathcal{F}_d$ and $\mathcal{F}_d^\dagger$), such that for any $(\theta_d, F) \in \mathcal{F}_+$, $\mathbb{E}_F[\overline{m}(Y, D, Z)]^{2+\delta} \leq C_2$. Therefore, $\mathcal{F}_+$ satisfies Assumption PS1 in Andrews and Shi (2017), with notation correspondence in Table 5.

TABLE 5. Notation Correspondence Table

Notations in Andrews and Shi (2017)	Our Notations
τ	(γ, λ)
$\mathcal{T}$	Λ
W	(Y, D, Z)
k	1
p	1
θ	θ_d
Θ	$[\underline{g}_d, \overline{g}_d]$
$\sigma_{F,1}(\theta)$	1
$m_1(W, \theta, \tau)$	$m(Y, D, Z, \theta_d, \gamma, \lambda)$
$M(W)$	$\overline{m}(Y, D, Z)$

Next, we show that for any $\theta_d \in [\underline{g}_d, \overline{g}_d]$, and any two $(\gamma, \lambda), (\gamma^\dagger, \lambda^\dagger) \in \Lambda$,

$$|m(Y, D, Z, \theta_d, \gamma, \lambda) - m(Y, D, Z, \theta_d, \gamma^\dagger, \lambda^\dagger)| \leq \overline{m}(Y, D, Z) \left\|(\gamma - \gamma^\dagger, \lambda - \lambda^\dagger)\right\|. \tag{D.5}$$

To see this, note that when $D = d$,

$$\begin{aligned} |m(Y, D, Z, \theta_d, \gamma, \lambda) - m(Y, D, Z, \theta_d, \gamma^\dagger, \lambda^\dagger)| &\leq |g(Y)| \cdot |\gamma - \gamma^\dagger + (\lambda - \lambda^\dagger)' h(Z)| \\ &\leq (|\underline{g}_d| + |\overline{g}_d|) \cdot \|(1, h(Z))\| \cdot \left\|(\gamma - \gamma^\dagger, \lambda - \lambda^\dagger)\right\| \end{aligned}$$

where the inequality follows from Cauchy-Schwarz inequality. When $D \neq d$, it's easy to see that[13]

$$\begin{aligned}|m(Y,D,Z,\theta_d,\gamma,\lambda) - m(Y,D,Z,\theta_d,\gamma^\dagger,\lambda^\dagger)| \leq & \max\Big\{|\underline{g}_d| \cdot |\gamma - \gamma^\dagger + (\lambda-\lambda^\dagger)'h(Z)|, \\ & |\overline{g}_d| \cdot |\gamma - \gamma^\dagger + (\lambda-\lambda^\dagger)'h(Z)|\Big\} \\ \leq & (|\underline{g}_d| + |\overline{g}_d|) \cdot \|(1,h(Z))\| \cdot \Big\|(\gamma-\gamma^\dagger, \lambda-\lambda^\dagger)\Big\|\end{aligned}$$

Hence, inequality (D.5) must hold.

Therefore, whenever $\big\|(\gamma-\gamma^\dagger,\lambda-\lambda^\dagger)\big\| \leq \epsilon$, for any $\{\alpha_i \in [0,\infty) : i = 1,...,n\}$, for any $\theta_d \in [\underline{g}_d, \overline{g}_d]$,

$$\sum_{i=1}^n (\alpha_i m(Y_i,D_i,Z_i,\theta_d,\gamma,\lambda) - \alpha_i m(Y_i,D_i,Z_i,\theta_d,\gamma^\dagger,\lambda^\dagger))^2 \leq \sum_{i=1}^n \big(\alpha_i \overline{m}(Y_i,D_i,Z_i)\big)^2 \cdot \epsilon^2. \tag{D.6}$$

Recall for any $\epsilon > 0$ and $V \subseteq \mathbb{R}^m$, the packing number $D(\epsilon, V)$ is defined as the biggest integer K such that there exist $v_1,...,v_K \in V$ and $\|v_k - v_{k'}\| \geq \epsilon$ for each $1 \leq k < k' \leq K$. Define, for any $\alpha \in \mathbb{R}^n_+$,

$$V_n(\alpha) \equiv \big\{(\alpha_1 m(Y_1,D_1,Z_1,\theta_d,\gamma,\lambda),...,\alpha_n m(Y_n,D_n,Z_n,\theta_d,\gamma,\lambda)) \in \mathbb{R}^n : (\gamma,\lambda) \in \Lambda\big\}.$$

Then, (D.6) implies

$$D\left(\epsilon\sqrt{\sum_{i=1}^n \big(\alpha_i \overline{m}(Y_i,D_i,Z_i)\big)^2}, V_n\right) \leq D(\epsilon,\Lambda).$$

Since $\Lambda \subseteq \mathcal{S}_{m+1}$, $D(\epsilon,\Lambda) \leq D(\epsilon,\mathcal{S}_{m+1})$. Lemma 4.1 in Pollard (1990) implies that there exists some $C_3 < \infty$ such that $D(\epsilon,\mathcal{S}_{m+1}) \leq C_3/\epsilon^{m+1}$. Since $\int_0^1 \sqrt{\log(C_3/\epsilon^{m+1})} < \infty$, we conclude that $\mathcal{F}_+$ satisfies Assumption PS2 in Andrews and Shi (2017) with notation correspondence in Table 5.

Finally, note that our statistics $T_n(\theta_d,d,\Lambda)$ equals to the CvM statistics in (3.7) in Andrews and Shi (2017) with $S = S_1$, and our $c_{n,1-\alpha}(\theta_d,d,\Lambda)$ equals the critical value constructed in Section 4 in Andrews and Shi (2017), except that we omit the step which transforms conditional moment inequalities into unconditional moment inequalities with instruments since our condition (2.11) only consists of unconditional moment inequalities. Hence, the results of the propositions follows from Theorem 5.1 and Theorem 6.1 in Andrews and Shi (2017). □

Let's now discuss the asymptotic size of $\text{CI}_{n,1-\alpha}(d_1,d_2)$. As shorthand notations, we write $\tau_{d_1} \equiv (\gamma_{d_1},\lambda_{d_1}) \in \mathbb{R}^{m+1}$, $\tau_{d_2} \equiv (\gamma_{d_2},\lambda_{d_2}) \in \mathbb{R}^{m+1}$. Define $h_{d_1,d_2,F}(\theta_{d_1},\theta_{d_2})$ as the asymptotic variance-covariance kernel of

[13]For example, suppose $|m(Y,D,Z,\theta_d,\gamma,\lambda) - m(Y,D,Z,\theta_d,\gamma^\dagger,\lambda^\dagger)| = |\overline{g}_d(\gamma+\lambda'h(Z)) - \underline{g}_d(\gamma^\dagger+\lambda^{\dagger\prime}h(Z))|$. When $\overline{g}_d(\gamma+\lambda'h(Z)) - \underline{g}_d(\gamma^\dagger+\lambda^{\dagger\prime}h(Z)) \geq 0$, we have

$$\begin{aligned}|\overline{g}_d(\gamma+\lambda'h(Z)) - \underline{g}_d(\gamma^\dagger+\lambda^{\dagger\prime}h(Z))| = & \overline{g}_d(\gamma+\lambda'h(Z)) - \underline{g}_d(\gamma^\dagger+\lambda^{\dagger\prime}h(Z)) \\ \leq & \overline{g}_d(\gamma+\lambda'h(Z)) - \overline{g}_d(\gamma^\dagger+\lambda^{\dagger\prime}h(Z)) \\ = & |\overline{g}_d| \cdot |\gamma-\gamma^\dagger + (\lambda-\lambda^\dagger)'h(Z)|.\end{aligned}$$

When $\overline{g}_d(\gamma+\lambda'h(Z)) - \underline{g}_d(\gamma^\dagger+\lambda^{\dagger\prime}h(Z)) \leq 0$, we have

$$\begin{aligned}|\overline{g}_d(\gamma+\lambda'h(Z)) - \underline{g}_d(\gamma^\dagger+\lambda^{\dagger\prime}h(Z))| = & -\overline{g}_d(\gamma+\lambda'h(Z)) + \underline{g}_d(\gamma^\dagger+\lambda^{\dagger\prime}h(Z)) \\ \leq & -\underline{g}_d(\gamma+\lambda'h(Z)) + \underline{g}_d(\gamma^\dagger+\lambda^{\dagger\prime}h(Z)) \\ = & |\underline{g}_d| \cdot |\gamma-\gamma^\dagger + (\lambda-\lambda^\dagger)'h(Z)|.\end{aligned}$$

$(n^{1/2}\overline{m}_n(\theta_{d_1}, d_1, \tau_{d_1}), n^{1/2}\overline{m}_n(\theta_{d_2}, d_2, \tau_{d_2}))$, i.e.

$$\begin{aligned} & h_{d_1,d_2,F}(\theta_{d_1}, \theta_{d_2}, \tau_{d_1}, \tau_{d_2}, \tau_{d_1}^\dagger, \tau_{d_2}^\dagger) \\ = \quad & \mathrm{cov}_F \left(\begin{bmatrix} \overline{f}_{d_1}(Y_i, D_i, \gamma_{d_1} + \lambda'_{d_1} h(Z_i)) - \theta_{d_1}(\gamma_{d_1} + \lambda'_{d_1} h(Z_i)) \\ \overline{f}_{d_2}(Y_i, D_i, \gamma_{d_2} + \lambda'_{d_2} h(Z_i)) - \theta_{d_2}(\gamma_{d_2} + \lambda'_{d_2} h(Z_i)) \end{bmatrix}, \right. \\ & \left. \begin{bmatrix} \overline{f}_{d_1}(Y_i, D_i, \gamma_{d_1}^\dagger + \lambda_{d_1}^{\dagger\prime} h(Z_i)) - \theta_{d_1}(\gamma_{d_1}^\dagger + \lambda_{d_1}^{\dagger\prime} h(Z_i)) \\ \overline{f}_{d_2}(Y_i, D_i, \gamma_{d_2}^\dagger + \lambda_{d_2}^{\dagger\prime} h(Z_i)) - \theta_{d_2}(\gamma_{d_2}^\dagger + \lambda_{d_2}^{\dagger\prime} h(Z_i)) \end{bmatrix} \right) \end{aligned}$$

Define the set of variance-covariance kernels:

$$\mathcal{H}_{d_1,d_2} \equiv \{h_{d_1,d_2,F}(\theta_{d_1}, \theta_{d_2}) : (\theta_{d_1}, \theta_{d_2}, F) \in \mathcal{F}_{d_1,d_2}\}.$$

On the space of all continuous 2×2 matrix valued functions defined on $\Lambda^2 \times \Lambda^2$, which is a superset of $\mathcal{H}_{d_1,d_2}$, we consider the topology generated by the uniform metric,

$$\rho(h, h^\dagger) \equiv \sup_{(\tau_{d_1}, \tau_{d_2}), (\tau_{d_1}^\dagger, \tau_{d_2}^\dagger) \in \Lambda^2 \times \Lambda^2} \left\| h(\tau_{d_1}, \tau_{d_2}, \tau_{d_1}^\dagger, \tau_{d_2}^\dagger) - h^\dagger(\tau_{d_1}, \tau_{d_2}, \tau_{d_1}^\dagger, \tau_{d_2}^\dagger) \right\|.$$

where h and $h^\dagger$ are two arbitrary 2×2 matrix valued continuous functions on $\Lambda^2 \times \Lambda^2$. The following proposition is implied by Proposition 4 and Bonferroni correction.

Proposition 5. *For any $d_1, d_2 \in \mathcal{D}$ and any compact subset $\mathcal{H}_{cpt}$ of $\mathcal{H}_{d_1,d_2}$, we have*

$$\liminf_{n\to\infty} \inf_{\substack{(\theta_{d_1}, \theta_{d_2}, F) \in \mathcal{F}_{d_1,d_2}: \\ h_{d_1,d_2,F}(\theta_{d_1}, \theta_{d_2}) \in \mathcal{H}_{cpt}}} \mathbb{P}_F\Big(\theta_{d_1} - \theta_{d_2} \in CI_{n,1-\alpha}(d_1, d_2)\Big) \geq 1 - \alpha.$$

Proof. By the construction of $\mathrm{CI}_{n,1-\alpha}(d_1, d_2)$, we know, for any θ_{d_1}, θ_{d_2} and F,

$$\begin{aligned} & \mathbb{P}_F\Big(\theta_{d_1} - \theta_{d_2} \in \mathrm{CI}_{n,1-\alpha}(d_1, d_2)\Big) \\ \geq \quad & \mathbb{P}_F\Big(\theta_{d_1} \in \mathrm{CI}_{n,1-\alpha/2}(d_1) \text{ and } \theta_{d_2} \in \mathrm{CI}_{n,1-\alpha/2}(d_2)\Big) \\ \geq \quad & \mathbb{P}_F\Big(\theta_{d_1} \in \mathrm{CI}_{n,1-\alpha/2}(d_1)\Big) + \mathbb{P}_F\Big(\theta_{d_2} \in \mathrm{CI}_{n,1-\alpha/2}(d_2)\Big) - 1, \end{aligned}$$

where the last inequality follows from the Fréchet inequality.

Hence,

$$\begin{aligned} & \inf_{\substack{(\theta_{d_1}, \theta_{d_2}, F) \in \mathcal{F}_{d_1,d_2}: \\ h_{d_1,d_2,F}(\theta_{d_1}, \theta_{d_2}) \in \mathcal{H}_{cpt}}} \mathbb{P}_F\Big(\theta_{d_1} - \theta_{d_2} \in \mathrm{CI}_{n,1-\alpha}(d_1, d_2)\Big) \\ \geq \quad & \inf_{\substack{(\theta_{d_1}, \theta_{d_2}, F) \in \mathcal{F}_{d_1,d_2}: \\ h_{d_1,d_2,F}(\theta_{d_1}, \theta_{d_2}) \in \mathcal{H}_{cpt}}} \mathbb{P}_F\Big(\theta_{d_1} \in \mathrm{CI}_{n,1-\alpha/2}(d_1)\Big) + \inf_{\substack{(\theta_{d_1}, \theta_{d_2}, F) \in \mathcal{F}_{d_1,d_2}: \\ h_{d_1,d_2,F}(\theta_{d_1}, \theta_{d_2}) \in \mathcal{H}_{cpt}}} \mathbb{P}_F\Big(\theta_{d_2} \in \mathrm{CI}_{n,1-\alpha/2}(d_2)\Big) - 1. \end{aligned}$$

For any $h \in \mathcal{H}_{cpt}$ and any $(\tau_{d_1}, \tau_{d_2}, \tau_{d_1}^\dagger, \tau_{d_2}^\dagger) \in \Lambda^2 \times \Lambda^2$, $h(\tau_{d_1}, \tau_{d_2}, \tau_{d_1}^\dagger, \tau_{d_2}^\dagger)$ is a 2×2 matrix in which element (1,1) only depends on θ_{d_1}, τ_{d_1} and $\tau_{d_1}^\dagger$. Let $\mathcal{H}_{cpt,d_1}$ be the subset in the space of continuous function on $\Lambda \times \Lambda$, in which each function $h_{d_1}(\tau_{d_1}, \tau_{d_1}^\dagger)$ equals the element (1,1) of some $h(\tau_{d_1}, \tau_{d_2}, \tau_{d_1}^\dagger, \tau_{d_2}^\dagger)$ in $\mathcal{H}_{cpt}$. Similarly, define $\mathcal{H}_{cpt,d_2}$ be the subset in the space of continuous functions on $\Lambda \times \Lambda$, in which each function $h_{d_2}(\tau_{d_2}, \tau_{d_2}^\dagger)$ equals the element (2,2) of some $h(\tau_{d_1}, \tau_{d_2}, \tau_{d_1}^\dagger, \tau_{d_2}^\dagger)$ in $\mathcal{H}_{cpt}$. Then, for each $d \in \{d_1, d_2\}$,

$$\liminf_{n\to\infty} \inf_{\substack{(\theta_{d1}, \theta_{d_2}, F) \in \mathcal{F}_{d_1,d_2}: \\ h_{d_1,d_2,F}(\theta_{d_1}, \theta_{d_2}) \in \mathcal{H}_{cpt}}} \mathbb{P}_F\Big(\theta_d \in \mathrm{CI}_{n,1-\alpha/2}(d)\Big) \geq \liminf_{n\to\infty} \inf_{\substack{(\theta_d, F) \in \mathcal{F}_d: \\ h_{d,F}(\theta_d) \in \mathcal{H}_{cpt,d}}} \mathbb{P}_F\Big(\theta_d \in \mathrm{CI}_{n,1-\alpha/2}(d)\Big)$$

Since $\mathcal{H}_{cpt}$ is compact, $\mathcal{H}_{cpt,d_1}$ and $\mathcal{H}_{cpt,d_2}$ are also compact. Proposition 4 implies

$$\liminf_{n\to\infty} \inf_{\substack{(\theta_d,F)\in\mathcal{F}_d:\\ h_{d,F}(\theta_d)\in\mathcal{H}_{cpt,d}}} \mathbb{P}_F\Big(\theta_d \in \mathrm{CI}_{n,1-\alpha/2}(d)\Big) \geq 1-\alpha/2.$$

Therefore, we conclude

$$\liminf_{n\to\infty} \inf_{\substack{(\theta_{d_1},\theta_{d_2},F)\in\mathcal{F}_{d_1,d_2}:\\ h_{d_1,d_2,F}(\theta_{d_1},\theta_{d_2})\in\mathcal{H}_{cpt}}} \mathbb{P}_F\Big(\theta_{d_1}-\theta_{d_2} \in \mathrm{CI}_{n,1-\alpha}(d_1,d_2)\Big) \geq 2(1-\alpha/2)-1 = 1-\alpha.$$

□

D.2. **Proofs of the Asymptotic validity of the Algorithm 2.** In this subsection, we are going to show the inference intervals proposed in (3.6) and (3.7) have correct asymptotic converge probability and are consistent against all fixed alternatives.

Proposition 6. *Suppose Assumption 1 hold. In addition, suppose (i) $\{Y_i, Z_i, D_i\}_{i=1}^n$ are i.i.d. samples; (ii) $\mathbb{E}\|h(Z)\|^{2+\delta} < \infty$ for some $\delta > 0$; (iii) for any $\theta_d \in [\underline{g}_d, \overline{g}_d]$, $Var(\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(V + W' h(Z))) > 0$; (iv) The support of (V, W) is core-determining. Then, for any θ_d satisfying inequality* (2.11)*, we have*

$$\liminf_{n\to\infty} \mathbb{P}(\theta_d \in CI_{n,1-\alpha}(d)) \geq 1-\alpha \tag{D.7}$$

Moreover, for any $\theta_d \in [\underline{g}_d, \overline{g}_d]$ not satisfying inequality (2.11)*, we have*

$$\lim_{n\to\infty} \mathbb{P}(\theta_d \in CI_{n,1-\alpha}(d)) = 0$$

Proposition 7. *Suppose Assumption 1 hold. In addition, suppose (i) $\{Y_i, Z_i, D_i\}_{i=1}^n$ are i.i.d. samples; (ii) $\mathbb{E}\|h(Z)\|^{2+\delta} < \infty$ for some $\delta > 0$; (iii) For each $d \in d_1, d_2$ and any $\theta_d \in [\underline{g}_d, \overline{g}_d]$, $Var(\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(V + W' h(Z))) > 0$; (iv) The support of (V, W) is core-determining; (v) For each $d \in \{d_1, d_2\}$, θ_d satisfying inequality* (2.11)*. Then, we have*

$$\liminf_{n\to\infty} \mathbb{P}(\theta_{d_1} - \theta_{d_2} \in CI_{n,1-\alpha}(d_1, d_2)) \geq 1-\alpha \tag{D.8}$$

Proof of Proposition 6. We first prove the first part the proposition. Suppose hypotheses (i)-(v) hold and θ_d satisfies inequality (2.11), we want to show

$$\liminf_{n\to\infty} \mathbb{P}(\theta_d \in \mathrm{CI}_{n,1-\alpha}(d)) \geq 1-\alpha.$$

This result follows directly from Theorem 2 in Andrews and Shi (2013). We only need to check that

$$\mathbb{E}[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(V + W' h(Z))]^{2+\delta} < +\infty.$$

To see why this is true, note that under Assumption 1,

$$\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(V + W' h(Z)) \leq (|\underline{g}_d| + |\overline{g}_d|) \cdot |(V + W' h(Z))| \text{ almost surely.}$$

Since $\|(V, M)\| = 1$, Cauchy-Schwarz inequality implies

$$\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(V + W' h(Z)) \leq (|\underline{g}_d| + |\overline{g}_d|) \cdot \|h(Z)\|.$$

Therefore,

$$\mathbb{E}[\overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(V + W' h(Z))]^{2+\delta} < (|\underline{g}_d| + |\overline{g}_d|)^{2+\delta} \mathbb{E}\|h(Z)\|^{2+\delta} < +\infty.$$

□

Uniform Asymptotic Properties of (3.6) and (3.7).

As a shorthand notation, define $VW \equiv (V, W)$ and define $m_d(Y, Z, D, VW)$ as

$$m_d(Y, Z, D, VW, \theta_d) \equiv \overline{f}_d(Y, D, \gamma + \lambda' h(Z)) - \theta_d(V + W' h(Z)).$$

Let F denote the joint distribution of (Y, Z, D, VW). Define $\mathcal{F}_d$ as the set of all (θ_d, F) pairs such that (i) $\theta_d \in [\underline{g}_d, \overline{g}_d]$; (ii) $\{(Y_i, Z_i, D_i, VW_i)\}_{i=1}^n$ are i.i.d. under F; (iii) θ_d satisfies (3.3) where the expectation is taken with respect to F, i.e.

$$\mathbb{E}_F\big[m_d(Y, Z, D, VW, \theta_d)\big|VW\big] \geq 0, \text{ almost surely,}$$

(iv) $0 < \text{Var}(m_d(Y, Z, D, VW, \theta_d)) < \infty$, and (v) $\mathbb{E}_F|m_d(Y, Z, D, VW, \theta_d)/\sigma_{F,d}(\theta_d)|^{2+\delta} \leq C_1$ for some constant $\delta > 0$ and $C_1 < \infty$, where $\sigma^2_{F,d}(\theta_d) \equiv \text{Var}(m_d(Y, Z, D, VW, \theta_d))$.

In order to state the uniform asymptotic properties of $\text{CI}_{n,1-\alpha}(d)$, we also need to introduce the instrumental functions used in Andrews, Kim and Shi (2017). Let VW_i^o be the standardized VW_i, i.e.

$$VW_i^o \equiv \Phi\Big(\widehat{\Sigma}_{VW,n}^{-1/2}\big(VW_i - \overline{VW}_n\big)\Big)$$

where $\overline{VW}_n = n^{-1}\sum_{i=1}^n VW_i \in \mathbb{R}^{m+1}$, $\widehat{\Sigma}_{VW,n} = n^{-1}\sum_{i=1}^n (VW_i - \overline{VW}_n)(VW_i - \overline{VW}_n)'$, and $\Phi(x) = (\Phi(x_1), ..., \Phi(x_{m+1}))' \in \mathbb{R}^{m+1}$ where Φ is the standard normal cumulative distribution function and $x = (x_1, ..., x_{m+1})'$.

In Andrews, Kim and Shi (2017), the following set $\mathcal{G}_{\text{c-cube}}$ of instrumental functions has been used,

$$\mathcal{G}_{\text{c-cube}} \equiv \Bigg\{ g_{a,r}(VW_i^0) = \mathbb{1}\Big(VW_i^o \in \times_{u=1}^{m+1}((a_u - 1)/(2r), a_u/(2r))\Big) : \\ a = (a_1, ..., a_{m+1})' \in \{1, 2, ..., 2r\}^{m+1} \text{ and } r = 1, 2, 3, ... \Bigg\}.$$

For any $g, g^\dagger \in \mathcal{G}_{\text{c-cube}}$, define

$$h_{d,F}(\theta_d, g, g^\dagger) \equiv \text{cov}_F\left(g(VW^o)\frac{m_d(Y, Z, D, VW, \theta_d)}{\sigma_{F,d}(\theta_d)}, g^\dagger(VW^o)\frac{m_d(Y, Z, D, VW, \theta_d)}{\sigma_{F,d}(\theta_d)}\right).$$

Let $\mathcal{H}_d \equiv \{h_{d,F}(\theta_d, \cdot, \cdot) : (\theta_d, F) \in \mathcal{F}_d\}$. On the space of all bounded real functions on $\mathcal{G}_{\text{c-cube}} \times \mathcal{G}_{\text{c-cube}}$, which is a superset of $\mathcal{H}_d$, we consider the uniform metric ρ as

$$\rho(h, \tilde{h}) \equiv \sup_{g, g^\dagger \in \mathcal{G}_{\text{c-cube}}} |h(g, g^\dagger) - \tilde{h}(g, g^\dagger)|.$$

where h and $\tilde{h}$ are two arbitrary bounded real functions on $\mathcal{G}_{\text{c-cube}} \times \mathcal{G}_{\text{c-cube}}$.

Proposition 8. *For any compact subset $\mathcal{H}_{cpt}$ of $\mathcal{H}_d$, we have*

$$\liminf_{n\to\infty} \inf_{\substack{(\theta_d, F) \in \mathcal{F}_d: \\ h_{d,F}(\theta_d) \in \mathcal{H}_{cpt}}} \mathbb{P}_F\left(\theta_d \in CI_{n,1-\alpha}(d)\right) \geq 1 - \alpha.$$

Proof. Because the choice of test statistics, the set of instrumental functions and all the tuning parameters in Andrews, Kim and Shi (2017) satisfy Assumption M, S1, S2 and Assumption GMS1 in Andrews and Shi (2013), this result follows directly from Theorem 2 in Andrews and Shi (2013). □

Let's now discuss the asymptotic properties of $\text{CI}_{n,1-\alpha}(d_1, d_2)$. For any $d_1, d_2 \in \mathcal{D}$ with $d_1 \neq d_2$, define $\mathcal{F}_{d_1,d_2}$ as the set of all $(\theta_{d_1}, \theta_{d_2}, F)$ pairs such that (i) for each $d \in \{d_1, d_2\}$, $\theta_d \in [\underline{g}_d, \overline{g}_d]$; (ii) $\{(Y_i, Z_i, D_i, VW_i)\}_{i=1}^n$ are i.i.d. under F; (iii) for each $d \in \{d_1, d_2\}$, θ_d satisfies (3.3) where the expectation is taken with respect to F, i.e.

$$\mathbb{E}_F\big[m_d(Y, Z, D, VW, \theta_d)\big|VW\big] \geq 0, \text{ almost surely,}$$

(iv) for each $d \in \{d_1, d_2\}$, $0 < \text{Var}(m_d(Y, Z, D, VW, \theta_d)) < \infty$, and (v) for each $d \in \{d_1, d_2\}$, $\mathbb{E}_F|m_d(Y, Z, D, VW, \theta_d)/\sigma_{F,d}(\theta_d)|^{2+\delta} \leq C_1$ for some constant $\delta > 0$ and $C_1 < \infty$, where $\sigma^2_{F,d}(\theta_d) \equiv \text{Var}(m_d(Y, Z, D, VW, \theta_d))$. For any $g, g^\dagger \in \mathcal{G}_{\text{c-cube}}$, define

$$\begin{aligned} m^*_{d_1,d_2}(Y, Z, D, V, W, \theta_{d_1}, \theta_{d_2}) &\equiv \left(\frac{m_{d_1}(Y, Z, D, VW, \theta_d)}{\sigma_{F,d_1}(\theta_{d_1})}, \frac{m_{d_2}(Y, Z, D, VW, \theta_d)}{\sigma_{F,d_2}(\theta_{d_2})} \right)' \\ h_{d_1,d_2,F}(\theta_{d_1}, \theta_{d_2}, g, g^\dagger) &\equiv \text{cov}_F\Big(g(VW^o) m^*_{d_1,d_2}(Y, Z, D, V, W, \theta_{d_1}, \theta_{d_2}), \\ &\qquad g^\dagger(VW^o) m^*_{d_1,d_2}(Y, Z, D, V, W, \theta_{d_1}, \theta_{d_2}) \Big). \end{aligned}$$

Let $\mathcal{H}_{d_1,d_2} \equiv \{h_{d_1,d_2,F}(\theta_{d_1}, \theta_{d_2}, \cdot, \cdot) : (\theta_{d_1}, \theta_{d_2}, F) \in \mathcal{F}_d\}$. On the space of all bounded functions mapping $\mathcal{G}_{\text{c-cube}} \times \mathcal{G}_{\text{c-cube}}$ to $\mathbb{R}^{2\times 2}$, which is a superset of $\mathcal{H}_{d_1,d_2}$, we consider the uniform metric ρ as

$$\rho(h, \tilde{h}) \equiv \sup_{g,g^\dagger \in \mathcal{G}_{\text{c-cube}}} \left\| h(g, g^\dagger) - \tilde{h}(g, g^\dagger) \right\|.$$

where h and $\tilde{h}$ are two arbitrary bounded 2×2 matrix valued functions on $\mathcal{G}_{\text{c-cube}} \times \mathcal{G}_{\text{c-cube}}$.

Proposition 9. *For any compact subset $\mathcal{H}_{cpt}$ of $\mathcal{H}_{d_1,d_2}$, we have*

$$\liminf_{n\to\infty} \inf_{\substack{(\theta_{d_1}, \theta_{d_2}, F) \in \mathcal{F}_{d_1,d_2}: \\ h_{d_1,d_2,F}(\theta_{d_1}, \theta_{d_2}) \in \mathcal{H}_{cpt}}} \mathbb{P}_F\Big(\theta_{d_1} - \theta_{d_2} \in CI_{n,1-\alpha}(d_1, d_2) \Big) \geq 1 - \alpha$$

Proof. This proposition is a result of Bonferroni correction and Proposition 8. Fix any $d_1, d_2 \in \mathcal{D}$. Note that, for any $\theta_{d_1} \in [\underline{g}_{d_1}, \overline{g}_{d_2}]$ and $\theta_{d_2} \in [\underline{g}_{d_2}, \overline{g}_{d_2}]$,

$$\begin{aligned} & \mathbb{P}\Big(\theta_{d_1} - \theta_{d_2} \in \text{CI}_{n,1-\alpha}(d_1, d_2) \Big) \\ \geq\ & \mathbb{P}\Big(\theta_{d_1} \in \text{CI}_{n,1-\alpha/2}(d_1) \text{ and } \theta_{d_2} \in \text{CI}_{n,1-\alpha/2}(d_2) \Big) \\ \geq\ & \mathbb{P}\Big(\theta_{d_1} \in \text{CI}_{n,1-\alpha/2}(d_1) \Big) + \mathbb{P}\Big(\theta_{d_2} \in \text{CI}_{n,1-\alpha/2}(d_2) \Big) - 1, \end{aligned}$$

where the last inequality follows from the Fréchet inequality.

Note that for any $g, g^\dagger \in \mathcal{G}_{\text{c-cube}}$, the $(1,1)$ element and $(2,2)$ element of $h_{d_1,d_2,F}(\theta_{d_1}, \theta_{d_2}, g, g^\dagger)$ equal to $h_{d_1,F}(\theta_{d_1}, g, g^\dagger)$ and $h_{d_2,F}(\theta_{d_2}, g, g^\dagger))$ respectively. For each $k \in \{1, 2\}$, define $\mathcal{H}_{cpt,d_k}$ as the subset in the space of all bounded real functions on $\mathcal{G}_{\text{c-cube}} \times \mathcal{G}_{\text{c-cube}}$ which equals the (k,k) element of some h in $\mathcal{H}_{cpt}$. Since $\mathcal{H}_{cpt}$ is compact, $\mathcal{H}_{cpt,d_1}$ and $\mathcal{H}_{cpt,d_2}$ are also compact. By proposition 8, we know, for each $d \in \{d_1, d_2\}$,

$$\liminf_{n\to\infty} \inf_{\substack{(\theta_d, F) \in \mathcal{F}_d: \\ h_{d,F}(\theta_d) \in \mathcal{H}_{cpt}}} \mathbb{P}_F\left(\theta_d \in \text{CI}_{n,1-\alpha/2}(d) \right) \geq 1 - \frac{\alpha}{2}.$$

Therefore, we conclude that

$$\begin{aligned} & \liminf_{n\to\infty} \inf_{\substack{(\theta_{d_1}, \theta_{d_2}, F) \in \mathcal{F}_{d_1,d_2}: \\ h_{d_1,d_2,F}(\theta_{d_1}, \theta_{d_2}) \in \mathcal{H}_{cpt}}} \mathbb{P}_F\Big(\theta_{d_1} - \theta_{d_2} \in \text{CI}_{n,1-\alpha}(d_1, d_2) \Big) \\ \geq\ & \liminf_{n\to\infty} \inf_{\substack{(\theta_{d_1}, F) \in \mathcal{F}_{d_1}: \\ h_{d_1,F}(\theta_{d_1}) \in \mathcal{H}_{cpt}}} \mathbb{P}_F\left(\theta_{d_1} \in \text{CI}_{n,1-\alpha/2}(d_1) \right) \\ & + \liminf_{n\to\infty} \inf_{\substack{(\theta_{d_2}, F) \in \mathcal{F}_{d_2}: \\ h_{d_2,F}(\theta_{d_2}) \in \mathcal{H}_{cpt}}} \mathbb{P}_F\left(\theta_{d_2} \in \text{CI}_{n,1-\alpha/2}(d_2) \right) - 1 \\ \geq\ & 1 - \alpha. \end{aligned}$$

□

References

Andrews, D. W. K., W. Kim, and X. Shi (2017): "Stata Commands for Testing Conditional Moment Inequalities/Equalities," *Unpublished manuscript*.

Andrews, D. W. K., and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666.

Andrews, D. W. K., and X. Shi (2017): "Inference Based on Many Conditional Moment Inequalities," *Journal of Econometrics*, 196(2), 275–287.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *The Review of Economic Studies*, 67(3), 499–527.

Angrist, J. D., and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90(430), 431–442.

Angrist, J. D., and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics*, 106(4), 979–1014.

Aronow, P. M., and A. Carnegie (2013): "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Political Analysis*, 21(4), 492–506.

Ashenfelter, O., and A. Krueger (1994): "Estimates of the Economic Return to Schooling from a New Sample of Twins," *The American Economic Review*, 84(5), 1157–1173.

Balke, A., and J. Pearl (1997): "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, 92(439), 1171–1176.

Belman, D., and J. S. Heywood (1991): "Sheepskin Effects in the Returns to Education: An Examination of Women and Minorities," *The Review of Economics and Statistics*, 73(4), 720–724.

Beresteanu, A., I. Molchanov, and F. Molinari (2011): "Sharp identification regions in models with convex moment predictions," *Econometrica*, 79(6), 1785–1821.

Beresteanu, A., I. Molchanov, and F. Molinari (2012): "Partial Identification Using Random Set Theory," *Journal of Econometrics*, 166(1), 17–32.

Berger, R. L. (1982): "Multiparameter Hypothesis Testing and Acceptance Sampling," *Technometrics*, 24, 295–300.

Berger, R. L., and J. C. Hsu (1996): "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets," *Statistical Science*, 11(4), 283–319.

Bertsekas, D. P. (1973): "Stochastic optimization problems with nondifferentiable cost functionals," *Journal of Optimization Theory and Applications*, 12(2), 218–231.

Bertsimas, D., and J. N. Tsitsiklis (1997): *Introduction to linear optimization*, Athena Scientific series in optimization and neural computation. Athena Scientific, Belmont, Mass.

Betts, J. (2010): "School Quality and Earnings," *International Encyclopedia of Education*, 2, 313–320.

Bugni, F. A., I. A. Canay, and X. Shi (2017): "Inference for Functions of Partially Identified Parameters in Moment Inequality Models," *Quantitative Economics*, 8(1), 1–38.

Card, D. (1994): "Earnings, Schooling, and Ability Revisited," *NBER Working Paper 4832*.

CARD, D. (1995): *Using Geographic Variation in College Proximity to Estimate the Return to Schooling.* in Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp, ed, by Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky. Toronto: University of Toronto Press.

CARD, D. (2001): "Estimating the Return to Schooling: Progress on some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160.

CARNEIRO, P., J. J. HECKMAN, and E. VYTLACIL (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101(6), 2754–2781.

CHERNOZHUKOV, V., W. KIM, S. LEE, and A. M. ROSEN (2015): "Implementing Intersection Bounds in Stata," *Stata Journal*, 15(1), 21–44.

CHERNOZHUKOV, V., S. LEE, and A. M. ROSEN (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81(2), 667–737.

CHESHER, A. (2002): "Instrumental Values," *CeMMAP Working paper CWP17/02*.

CHESHER, A. (2003): "Identification in nonseparable models," *Econometrica*, 71(5), 1405–1441.

CHESHER, A. (2005): "Nonparametric identification under discrete variation," *Econometrica*, 73, 1525–1550.

CHESHER, A., and A. ROSEN (2017): "Incomplete English Auction Models with Heterogeneity," *CeMMAP Working paper CWP27/17*.

CHETVERIKOV, D. (2013): "Testing Regression Monotonicity in Econometric Models," *Working Paper*.

DEARDEN, L., J. FERRI, and C. MEGHIR (2002): "The Effect of School Quality on Educational Attainment and Wages," *The Review of Economics and Statistics*, 84(1), 1–20.

DEATON, A. S. (2009): "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development," *NBER Working Paper*, 14690.

EKELAND, I., A. GALICHON, and M. HENRY (2010): "Optimal Transportation and the Falsifiability of Incompletely Specified Models," *Economic Theory*, 42(2), 355–374.

GALICHON, A., and M. HENRY (2006): "Inference in Incomplete Models," *Working Paper*.

GALICHON, A., and M. HENRY (2011): "Set identification in models with multiple equilibria," *The Review of Economic Studies*, 78(4), 1264–1298.

GINTHER, D. K. (2000): "Alternative Estimates of the Effect of Schooling on Earnings," *The Review of Economics and Statistics*, 82(1), 103–116.

HECKMAN, J. J., and S. URZUA (2010): "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify," *Journal of Econometrics*, 156(1), 27–37.

HECKMAN, J. J., S. URZUA, and E. J. VYTLACIL (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88(3), 389–432.

HECKMAN, J. J., and E. VYTLACIL (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73(3), 669–738.

HOROWITZ, J. L., and C. F. MANSKI (1995): "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63(2), 281–302.

HSU, Y.-C., C.-A. LIU, and X. SHI (2018): "Testing Generalized Regression Monotonicity," *Econometric Theory, Forthcoming*.

HUBER, M., and G. MELLACE (2015): "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints," *The Review of Economics and Statistics*, 97(2), 398–411.

Hungerford, T., and G. Solon (1987): "Sheepskin Effects in the Returns to Education," *The Review of Economics and Statistics*, 69(1), 175–177.

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

Jaeger, D. A., and M. E. Page (1996): "Degrees Matter: New Evidence on Sheepskin Effects in the Returns to Education," *The Review of Economics and Statistics*, 78(4), 720–724.

Kaido, H., F. Molinari, and J. Stoye (2017): "Confidence Intervals for Projections of Partially Identified Parameters," *Working Paper*.

Kédagni, D., and I. Mourifié (2017): "Generalized Instrumental Inequalities: Testing the IV Independence Assumption," *Unpublished manuscript*.

Kitagawa, T. (2009): "Identication Region of the Potential Outcome Distributions under Instrument Independence," *CeMMAP working paper CWP30/09*.

Kitagawa, T. (2010): "Testing for Instrument Independence in the Selection Model," *Unpublished manuscript*.

Kitagawa, T. (2015): "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063.

Lee, D. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 76(3), 1071–1102.

Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Reviews, Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association*, 80(2), 319–323.

Manski, C. F. (1994): "The Selection Problem," *in Advances Economics, Sixth World Congress, C. Sims (ed.), Cambridge University Press*, 1, 143–170.

Manski, C. F. (2003): *Partial Identification of Probability Distributions*. Springer-Verlag, New York.

Manski, C. F., and J. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010.

Masten, M. A., and A. Poirier (2017): "Inference on Breakdown Frontiers," *Working Paper*.

Masten, M. A., and A. Poirier (2018): "Identification of Treatment Effects under Conditional Partial Independence," *Econometrica*, 86(1), 317–351.

Mourifié, I., M. Henry, and R. Méango (2018): "Sharp Bound and Testability of a Roy Model Model of STEM Major Choices," *Unpublished manuscript*.

Mourifié, I., and Y. Wan (2017): "Testing Local Average Treatment Effect Assumptions," *The Review of Economics and Statistics*, 99(2), 305–313.

Oreopoulos, P. (2006): "Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter," *American Economic Review*, 96(1), 152–175.

Pearl, J. (1994): "On the Testability of Causal Models with Latent and Instrumental Variables," *Uncertainty in Articial Intelligence*, 11, 435–443.

Pearl, J. (2011): "Principal Stratification–a Goal or a Tool?," *The International Journal of Biostatistics*, 7(1), 20.

Russell, T. (2017): "Sharp Bounds on Functionals of the Joint Distribution in the Analysis of Treatment Effects," *Unpublished manuscript*.

Schennach, S. M. (2014): "Entropic Latent Variable Integration Via Simulation," *Econometrica*, 82(1), 345–385.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, 2 edn.