NBER WORKING PAPER SERIES

USING INSTRUMENTAL VARIABLES FOR INFERENCE ABOUT POLICY RELEVANT TREATMENT EFFECTS

Magne Mogstad
Andres Santos
Alexander Torgovitsky

Working Paper 23568
http://www.nber.org/papers/w23568

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2017

We thank Liran Einav, Derek Neal, Ed Vytacil, Chris Walters, and seminar participants at several universities and conferences for valuable feedback and suggestions. Bradley Setzler provided excellent research assistance. We are grateful to Pascaline Dupas for her help in accessing the data and in understanding the institutional details. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Using Instrumental Variables for Inference about Policy Relevant Treatment Effects
Magne Mogstad, Andres Santos, and Alexander Torgovitsky
NBER Working Paper No. 23568
July 2017
JEL No. C21,C36

ABSTRACT

We propose a method for using instrumental variables (IV) to draw inference about causal effects for individuals other than those affected by the instrument at hand. Policy relevance and external validity turns on the ability to do this reliably. Our method exploits the insight that both the IV estimand and many treatment parameters can be expressed as weighted averages of the same underlying marginal treatment effects. Since the weights are known or identified, knowledge of the IV estimand generally places some restrictions on the unknown marginal treatment effects, and hence on the values of the treatment parameters of interest. We show how to extract information about the average effect of interest from the IV estimand, and, more generally, from a class of IV-like estimands that includes the two stage least squares and ordinary least squares estimands, among many others. Our method has several applications. First, it can be used to construct nonparametric bounds on the average causal effect of a hypothetical policy change. Second, our method allows the researcher to flexibly incorporate shape restrictions and parametric assumptions, thereby enabling extrapolation of the average effects for compliers to the average effects for different or larger populations. Third, our method can be used to test model specification and hypotheses about behavior, such as no selection bias and/or no selection on gain. To accommodate these diverse applications, we devise a novel inference procedure that is designed to exploit the convexity of our setting. We develop uniformly valid tests that allow for an infinite number of IV--like estimands and a general convex parameter space. We apply our method to analyze the effects of price subsidies on the adoption and usage of an antimalarial bed net in Kenya.

Magne Mogstad
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
and NBER
magne.mogstad@gmail.com

Andres Santos
Department of Economics
University of California - Los Angeles
8283 Bunche Hall, 315 Portola Plaza
Los Angeles, CA, 90095
andres@econ.ucla.edu

Alexander Torgovitsky
Department of Economics
Northwestern University
2001 Sheridan Rd, Room 302
Evanston, IL 60208-2600
a-torgovitsky@northwestern.edu

1 Introduction

In an influential paper, Imbens and Angrist (1994) provided conditions under which an instrumental variables (IV) estimand can be interpreted as the average causal effect for the subpopulation of compliers, i.e. for those whose treatment status would be affected by an exogenous manipulation of the instrument. In some cases, this local average treatment effect (LATE) is of intrinsic interest, for example if the instrument itself represents an intervention or policy change of interest. On the other hand, in many situations, the causal effect for individuals induced to treatment by the instrument at hand might not be representative of the causal effect for those who would be induced to treatment by a given policy change of interest to the researcher. In these cases, the LATE is not the relevant parameter for evaluating the policy change.

In this paper, we show how to use instrumental variables to draw inference about treatment parameters other than the LATE, thereby learning about causal effects for individuals other than those affected by the instrument at hand. Policy relevance and external validity turn on the ability to do this reliably. Our setting is the canonical program evaluation problem with a binary treatment $D \in \{0, 1\}$ and a scalar, real-valued outcome, Y.[1] Corresponding to the two treatment arms are unobservable potential outcomes, Y_0 and Y_1. These represent the realization of Y that would have been experienced by an individual had their treatment status been exogenously set to 0 or 1. The relationship between observed and potential outcomes is given by

$$Y = DY_1 + (1 - D)Y_0. \tag{1}$$

Following Heckman and Vytlacil (1999, 2005), we assume that treatment is determined by the weakly separable selection or choice equation

$$D = \mathbb{1}[\nu(Z) - U \geq 0], \tag{2}$$

where ν is an unknown function, U is a continuously distributed random variable, and Z is a vector of observable regressors. Suppose that Z is independent of (Y_0, Y_1, U), perhaps conditional on some subvector X of Z. Under this assumption, the IV model given by (1)–(2) is equivalent to the IV model used by Imbens and Angrist (1994) and

[1] For discussions of heterogeneous effects IV models with multiple discrete treatments, we refer to Angrist and Imbens (1995), Heckman, Urzua, and Vytlacil (2006), Heckman and Vytlacil (2007b), Heckman and Urzua (2010), Kirkeboen, Leuven, and Mogstad (2016), and Lee and Salanié (2016), among others. Heterogeneous effects IV models with continuous treatments have been considered by Angrist, Graddy, and Imbens (2000), Chesher (2003), Florens, Heckman, Meghir, and Vytlacil (2008), Imbens and Newey (2009), Torgovitsky (2015), Masten (2015), and Masten and Torgovitsky (2016), among others.

many subsequent authors (Vytlacil, 2002). In particular, the instrument monotonicity condition of Imbens and Angrist (1994) is embedded in the separability of U and Z in the latent index $\nu(Z) - U$. An important feature of the model is that treatment effects $Y_1 - Y_0$ can vary across individuals with the same observable characteristics, X, in a way that depends on the unobservable component of treatment choice, U.

Our goal is to develop a method that uses a random sample of (Y, D, Z) together with the structure of the model to draw inference about a parameter of interest, $\beta^\star$, that a researcher has decided is relevant for evaluating a hypothetical policy change or intervention. Our method builds on the work of Heckman and Vytlacil (1999, 2001a,b,c, 2005, 2007a,b). Those authors showed that many different treatment parameters can be expressed in terms of the marginal treatment effect (MTE) function

$$\text{MTE}(u, x) \equiv E\left[Y_1 - Y_0 | U = u, X = x\right]. \tag{3}$$

The MTE can be interpreted as the average treatment effect indexed as a function of an individual's latent propensity to receive treatment, U, and conditional on other covariates, X. Heckman and Vytlacil (2005) showed that common parameters of interest can be expressed as weighted averages of the MTE function, with weights that are either known or identified. They showed that the same is also true of the IV estimand.

These insights suggest that even if the IV estimand is not of direct interest, it still carries information about the underlying MTE function, and hence about the parameter of interest, $\beta^\star$. In particular, since the weights for both the IV estimand and the parameter of interest are identified, knowledge of the IV estimand generally places some restrictions on the unknown MTE function, and hence on the range of values for $\beta^\star$ that are consistent with the data. This can be seen by writing:

$$\begin{aligned}
\underbrace{\beta_{\text{IV}}}_{\text{identified IV estimand}} &\equiv \int \underbrace{\text{MTE}(u)}_{\text{unknown}} \times \underbrace{\omega_{\text{IV}}(u)}_{\text{identified IV weights}} \, du \\
\underbrace{\beta^\star}_{\text{unknown target parameter}} &\equiv \int \underbrace{\text{MTE}(u)}_{\text{unknown}} \times \underbrace{\omega^\star(u)}_{\text{identified target weights}} \, du,
\end{aligned} \tag{4}$$

where we are assuming for the moment that there are no covariates X, just for simplicity. Equation (4) suggests that we can extract information about the parameter of interest, $\beta^\star$, from the IV estimand, β_{IV}, by solving an optimization problem. In particular, $\beta^\star$ must be smaller than

$$\max_{\text{MTE}} \int \text{MTE}(u)\omega^\star(u)\, du \quad \text{subject to} \quad \int \text{MTE}(u)\omega_{\text{IV}}(u)\, du = \beta_{\text{IV}}, \tag{5}$$

where the maximum is taken over a set of potential MTE functions that also incorporates any additional a priori assumptions that the researcher chooses to maintain. Similarly, $\beta^\star$ must be larger than the solution to the analogous minimization problem.

The optimization problem (5) has only the single constraint involving β_{IV}. Using the same logic, one can also include similar constraints for other IV estimands that correspond to different functions of Z. Upon doing so, the bounds on $\beta^\star$ will necessarily tighten, because each new IV estimand reduces the feasible set in (5). We show that, more generally, any cross moment between Y and a known function of D and Z can also be written as a weighted average of the two marginal treatment response (MTR) functions that constitute an MTE function. We refer to this class of cross moments as "IV–like" estimands.

The class of IV–like estimands is general enough to contain the estimands corresponding to any weighted linear IV estimator. This includes, as special cases, the two stage least squares (TSLS), optimal generalized method of moments, and ordinary least squares (OLS) estimands. Each moment in this class provides a different weighted average of the same underlying MTR functions, and therefore carries some distinct information about the possible values of the parameter of interest, $\beta^\star$. We show how these IV–like estimands can be chosen systematically so as to provide the tightest possible bounds on $\beta^\star$.

Our method has several applications. First, it can be used to construct nonparametric bounds on the average causal effect of a hypothetical policy change. Second, our method enables extrapolation of the average effects for compliers to the average effects for different or larger populations. Third, our method can be used to perform tests of model specification and of individual behavior, such as testing the null hypotheses of no selection bias and/or no selection on gains. In all of these applications, our method provides a researcher the *option* to impose parametric and/or shape restrictions, if desired.

To accommodate these diverse applications, we develop a novel inference framework that allows for (but does not require) nonparametric and/or shape constrained specifications for the MTR functions, as well as an infinite number of IV–like estimands. Our approach is specifically designed to take advantage of the convexity of our setting. We show that it satisfies two key requirements. First, it provides uniform size control over a wide class of distributions, a feature which is critically important in partially identified settings (Imbens and Manski, 2004). Second, implementing our procedure involves solving optimization problems for which there exist algorithms that provably converge to the global optimum. The generality of our inference results make them of independent interest, and we state them in a manner that facilitates their portability.

We apply our method using data from Dupas (2014). This data comes from a randomized pricing experiment for a preventative health product, conducted in Kenya. The goal of our empirical analysis is to assess how a class of potential subsidy regimes can promote the use of the health product, and to compare increases in usage to the costs of subsidization. For example, we measure the effect of a policy that offers free provision to each household as compared to a policy under which all households can purchase the product at a given price. This comparison does not correspond to the variation in prices induced by the experiment. As a result, it is not point identified under standard instrumental variables assumptions. However, our method can be used to estimate bounds on the average causal effect of this comparison. Our results show that these bounds can be very informative.

Our paper contributes to several literatures. A large body of work is concerned with using instrumental variables to draw nonparametric inference about treatment parameters other than the LATE. Heckman and Vytlacil (2005) observe that if Z is continuously distributed and has a sufficiently large impact on treatment choices D, so that the propensity score $P(D = 1|Z = z)$ varies over the entire $[0, 1]$ interval, then the MTE function is nonparametrically point identified. As a consequence, any target parameter $\beta^\star$ is also nonparametrically point identified. In practice, however, instruments have limited support and are often discrete or even binary. For these situations, many common target parameters of interest, such as the average treatment effect, are not nonparametrically point identified. Analytic expressions for sharp bounds on the average treatment effect have been derived by Manski (1989, 1990, 1994, 1997, 2003), Balke and Pearl (1997), Heckman and Vytlacil (2001b) and Kitagawa (2009), among others.[2]

Analytic expressions for bounds are useful because they provide intuition on the source and strength of identification. However, it can be difficult to derive analytic bounds for more complicated parameters, such as the policy relevant treatment effects (PRTEs) studied by Heckman and Vytlacil (2001a, 2005) and Carneiro, Heckman, and Vytlacil (2010, 2011). Our methodology is particularly useful in such settings. In addition, our method provides a unified framework for imposing shape restrictions such as monotonicity, concavity, monotone treatment selection (Manski, 1997; Manski and Pepper, 2000, 2009) and separability between observed and unobserved factors in the MTE function (Brinch, Mogstad, and Wiswall, 2015). It can be especially difficult to

[2] Note that Manski's analyses did not impose the separable first stage equation (2), see Heckman and Vytlacil (2001b) and Kitagawa (2009) for further discussion. Also related is work by Shaikh and Vytlacil (2011), Bhattacharya, Shaikh, and Vytlacil (2012), and Mourifié (2015), who augment (2) with a similar assumption for the potential outcomes, (Y_0, Y_1).

derive analytic bounds for treatment parameters that incorporate these types of assumptions in flexible combinations. In contrast, our general computational approach allows one to flexibly adjust the parameter of interest, as well as the maintained assumptions, without requiring additional identification analysis.

In addition, our paper is related to recent work that considers extrapolation in instrumental variables model under additional assumptions. While our method delivers bound on the target parameter in general, these bounds nest important point identification results as special cases. For example, our method nests existing approaches that extrapolate by assuming no unobserved heterogeneity in the treatment effect (Heckman and Robb, 1985; Angrist and Fernandez-Val, 2013), and those that parameterize this unobserved heterogeneity (Heckman, Tobias, and Vytlacil, 2003; Brinch et al., 2015).[3] One attractive feature of our method is that the constraints in (5) require an MTE function to also yield the usual, nonparametrically point identified LATE. Hence, our method allows for extrapolation to other parameters of interest without sacrificing the internal validity of the LATE.

Our paper also relates to a literature on specification tests in settings with instrumental variables. To see this, suppose that (5) is infeasible, so that there does not exist an MTE function that can both satisfy the researcher's assumptions and lead to the observed IV estimand. Then the model is misspecified: Either the researcher's assumptions are invalid, Z is not exogenous, the selection equation (2) is rejected by the data, or some combination of the three. Balke and Pearl (1997) and Imbens and Rubin (1997) noted that (2) has testable implications, while Machado, Shaikh, and Vytlacil (2013), Huber and Mellace (2014), Kitagawa (2015), and Mourifié and Wan (2016) have developed this observation into formal statistical tests. Our method builds on the work of these authors by allowing the researcher to maintain additional assumptions, such as parametric and/or shape restrictions. In addition to testing whether the model is misspecified, our method can also be used to test null hypotheses such as no selection bias and/or no selection on gains.

Lastly, our paper contributes to a growing literature on inference for functionals of partially identified parameters. Our inference procedure is based on a profile statistic. Romano and Shaikh (2008) and Bugni, Canay, and Shi (2015) proposed using profile statistics for moment inequality models, while Chernozhukov, Newey, and Santos (2015) considered models with conditional moment inequalities. Our model contains additional structure not present in moment inequality models.[4] We utilize this spe-

[3] For a completely different Bayesian approach to extrapolation in instrumental variables models, see Chamberlain (2011).

[4] Loosely speaking, in a moment inequality model, the inequalities are random, whereas in our context

cial structure to develop distributional approximations that are uniformly valid under low level conditions, and which only require the parameter space to be convex. An important alternative to profiling test statistics has been recently proposed by Kaido, Molinari, and Stoye (2016), who instead construct confidence regions through an adjusted projection algorithm. Their work, in common with many other papers in the moment inequalities literature, focuses on finite dimensional parameters and a finite number of moment restrictions. Both of these conditions are restrictive for our applications of interest. Our analysis is also related to Beresteanu and Molinari (2008), Bontemps, Magnac, and Maurin (2012), and Kaido and Santos (2014), who also exploit convexity for statistical inference.

The remainder of the paper is organized as follows. In Section 2, we present the model and develop our method for bounding a target parameter of interest while potentially maintaining additional shape constraints. In Section 3, we discuss key applications of our method, which we illustrate in Section 4 through a numerical example. We develop our statistical inference procedure in Section 5. In Section 6, we apply our method to study the effects of price subsidies on the adoption of preventative health products. We provide some concluding remarks in Section 7. Proofs for all results presented in the main text are contained in Appendices A and C.

2 Identification

Throughout this section, we assume that the researcher knows the joint distribution of the observed data (Y, D, Z). We address issues of statistical inference in Section 5.

2.1 Model

Our analysis uses the IV model consisting of (1)–(2), which is also often referred to as the two-sector generalized Roy model. The observable variables in the model are the outcome $Y \in \mathbf{R}$, the binary treatment $D \in \{0, 1\}$, and a vector of observables $Z \in \mathbf{R}^{d_z}$. We decompose Z into $Z = (X, Z_0)$, where $Z_0 \in \mathbf{R}^{d_{z_0}}$ are exogenous instruments and $X \in \mathbf{R}^{d_x}$ are control variables. The unobservables are the potential outcomes (Y_0, Y_1), and the variable U in the selection equation, which represents unobservable factors that affect treatment choice.

We maintain the following assumptions throughout the paper.

Assumptions I

I.1 *$U \perp\!\!\!\perp Z_0 | X$, where $\perp\!\!\!\perp$ denotes (conditional) statistical independence.*

the inequalities are deterministic, because they arise from the specification of the parameter space.

I.2 *$E[Y_d|Z,U] = E[Y_d|X,U]$ and $E[Y_d^2] < \infty$ for $d \in \{0,1\}$.*

I.3 *U is continuously distributed, conditional on X.*

Assumptions I.1 and I.2 require Z_0 to be exogenous with respect to both the selection and outcome processes. Vytlacil (2002) showed that, given I.1, the assumption that the index of the selection equation is additively separable as in (2) is equivalent to the assumption that Z_0 affects D monotonically in the sense introduced by Imbens and Angrist (1994). Hence, I.1 combined with (2) imposes substantive restrictions on choice behavior. Assumption I.2 imposes an exclusion restriction that the conditional means of Y_0 and Y_1 depend on $Z = (Z_0, X)$ only through the covariates X.

Assumption I.3 is a weak regularity condition that enables us to impose a standard normalization. As is well known, equation (2) may be rewritten as

$$D = \mathbb{1}\left[F_{U|X}(U|X) \leq F_{U|X}(\nu(Z)|X)\right] \equiv \mathbb{1}[\widetilde{U} \leq \widetilde{\nu}(Z)], \tag{6}$$

where we are using the notation $F_{U|X}(u|x) \equiv P(U \leq u|X = x)$ and we have defined $\widetilde{U} \equiv F_{U|X}(U|X)$ and $\widetilde{\nu}(Z) \equiv F_{U|X}(\nu(Z)|X)$. Under Assumptions I.1 and I.3, $\widetilde{U}$ is uniformly distributed on $[0,1]$, conditional on $Z = (Z_0, X)$. Working with this normalized model simplifies the analysis and does not affect its empirical content. Hence, we drop the tilde and maintain throughout the paper the normalization that U itself is distributed uniformly over $[0,1]$ conditional on Z. A consequence of this normalization is that

$$p(z) \equiv P(D = 1|Z = z) = F_{U|Z}(\nu(z)|z) = \nu(z), \tag{7}$$

where $p(z)$ is the propensity score.

It is important to observe what is *not* being assumed under Assumptions I. First, we do not impose any conditions on the support of Z: Both the control (X) and exogenous (Z_0) components of Z may be either continuous, discrete and ordered, categorical, or binary. Second, the IV model as specified here allows for rich forms of observed and unobserved heterogeneity. In particular, it allows $Y_1 - Y_0$ to vary not only across individuals with different values of X, but also among individuals with the same X. The treatment D may be statistically dependent with Y_0 (indicating selection bias), or $Y_1 - Y_0$ (indicating selection on the gain), or both, even conditional on X. Third, the model does not specify why individuals make the treatment choice that they do, in contrast to a stylized Roy model in which $D = \mathbb{1}[Y_1 > Y_0]$. However, the model also does not preclude the possibility that individuals choose treatment with full or partial knowledge of the potential outcomes (Y_0, Y_1). Any such knowledge will be reflected through dependence between the potential outcomes, (Y_0, Y_1), and the unobserved component treatment choice, U. Assumption I does not restrict this dependence.

2.2 What We Want to Know: The Target Parameter

As observed by Heckman and Vytlacil (1999, 2005), a wide range of treatment parameters can be written as weighted averages of the underlying MTE function. We use a slight generalization of their observation. Instead of working with the MTE function (3) directly, we consider treatment parameters that can be expressed as functions of the two marginal treatment response (MTR) functions, defined as

$$m_0(u,x) \equiv E\left[Y_0 \mid U = u, X = x\right] \quad \text{and} \quad m_1(u,x) \equiv E\left[Y_1 \mid U = u, X = x\right]. \tag{8}$$

Of course, each pair $m \equiv (m_0, m_1)$ of MTR functions generates an associated MTE function $\text{MTE}(u,x) \equiv m_1(u,x) - m_0(u,x)$. One benefit of working with MTR functions instead of MTE functions is that it allows us to consider parameters that weight m_0 and m_1 asymmetrically. Another benefit is that it allows the researcher to impose assumptions on m_0 and m_1 separately.

We assume that the researcher is interested in a target parameter $\beta^\star$ that can be written for any candidate pair of MTR functions $m \equiv (m_0, m_1)$ as

$$\beta^\star \equiv E\left[\int_0^1 m_0(u,X)\omega_0^\star(u,Z)\,d\mu^\star(u)\right] + E\left[\int_0^1 m_1(u,X)\omega_1^\star(u,Z)\,d\mu^\star(u)\right], \tag{9}$$

where $\omega_0^\star$ and $\omega_1^\star$ are identified weighting functions, and $\mu^\star$ is an integrating measure that is chosen by the researcher and usually taken to be the Lebesgue measure on $[0,1]$. For example, to set $\beta^\star$ to be the average treatment effect (ATE), observe that

$$E[Y_1 - Y_0] = E[m_1(U,X) - m_0(U,X)] = E\left[\int_0^1 m_1(u,X)du\right] - E\left[\int_0^1 m_0(u,X)du\right],$$

take $\omega_1^\star(u,z) = 1$, $\omega_0^\star(u,z) = -1$, and let $\mu^\star$ be the Lebesgue measure on $[0,1]$. Similarly, to set $\beta^\star$ to be the ATE conditional on X lying in some known set $\mathcal{X}^\star$, take

$$\omega_1^\star(u,z) \equiv \omega_1^\star(u,x,z_0) = \frac{\mathbb{1}[x \in \mathcal{X}^\star]}{P(X \in \mathcal{X}^\star)},$$

$\omega_0^\star(u,z) = -\omega_1^\star(u,z)$, and let $\mu^\star$ be as before. The resulting target parameter is then the population average effect of assigning treatment randomly to every individual with covariates $x \in \mathcal{X}^\star$, assuming full compliance.

Table 1: Weights for a Variety of Target Parameters

		Weights		
Target Parameter	**Expression**	$\boldsymbol{\omega_0(u,z) \equiv \omega_0(u,x,z_0)}$	$\boldsymbol{\omega_1(u,z) \equiv \omega_1(u,x,z_0)}$	**Measure** $\boldsymbol{\mu^\star}$
Average Untreated Outcome	$E[Y_0]$	1	0	Leb.$[0,1]$
Average Treated Outcome	$E[Y_1]$	0	1	Leb.$[0,1]$
Average Treatment Effect (ATE)	$E[Y_1 - Y_0]$	-1	1	Leb.$[0,1]$
Average Treatment Effect (ATE) given $X \in \mathcal{X}^\star$	$E[Y_1 - Y_0 \| X \in \mathcal{X}^\star]$	$-\omega_1^\star(u,z)$	$\dfrac{\mathbb{1}[x \in \mathcal{X}^\star]}{P(X \in \mathcal{X}^\star)}$	Leb.$[0,1]$
Average Treatment on the Treated (ATT)	$E[Y_1 - Y_0 \| D = 1]$	$-\omega_1^\star(u,z)$	$\dfrac{\mathbb{1}[u \leq p(z)]}{P(D=1)}$	Leb.$[0,1]$
Average Treatment on the Untreated (ATU)	$E[Y_1 - Y_0 \| D = 0]$	$-\omega_1^\star(u,z)$	$\dfrac{\mathbb{1}[u > p(z)]}{P(D=0)}$	Leb.$[0,1]$
Marginal Treatment Effect at $\overline{u}$	$E[Y_1 - Y_0 \| U = \overline{u}]$	-1	1	Dirac($\overline{u}$)
Local Average Treatment Effect for $U \in [\underline{u}, \overline{u}]$ (LATE($\underline{u}, \overline{u}$))	$E[Y_1 - Y_0 \| U \in [\underline{u}, \overline{u}]]$	$-\omega_1^\star(u,z)$	$\dfrac{\mathbb{1}[u \in [\underline{u}, \overline{u}]]}{(\overline{u} - \underline{u})}$	Leb.$[0,1]$
Policy Relevant Treatment Effect (PRTE) for new policy $(p^\star, Z^\star)$	$\dfrac{E[Y^\star] - E[Y]}{E[D^\star] - E[D]}$	$-\omega_1^\star(u,z)$	$\dfrac{\mathbb{1}[u \leq p^\star(z^\star)] - \mathbb{1}[u \leq p(z)]}{E[p^\star(Z^\star)] - E[p(Z)]}$	Leb.$[0,1]$
Additive PRTE with magnitude α	PRTE with $Z^\star = Z$ and $p^\star(z) = p(z) + \alpha$	$-\omega_1^\star(u,z)$	$\dfrac{\mathbb{1}[u \leq p(z) + \alpha] - \mathbb{1}[u \leq p(z)]}{\alpha}$	Leb.$[0,1]$
Proportional PRTE with magnitude α	PRTE with $Z^\star = Z$ and $p^\star(z) = (1+\alpha)p(z)$	$-\omega_1^\star(u,z)$	$\dfrac{\mathbb{1}[u \leq (1+\alpha)p(z)] - \mathbb{1}[u \leq p(z)]}{\alpha E[p(Z)]}$	Leb.$[0,1]$
PRTE for an additive α shift of the j^{th} component of Z	PRTE with $Z^\star = Z + \alpha e_j$ and $p^\star(z) = p(z)$	$-\omega_1^\star(u,z)$	$\dfrac{\mathbb{1}[u \leq p(z + \alpha e_j)] - \mathbb{1}[u \leq p(z)]}{E[p(Z + \alpha e_j)] - E[p(Z)]}$	Leb.$[0,1]$
Sum of two quantities $\beta_A^\star$, $\beta_B^\star$ with common measure $\mu^\star$	$\beta_A^\star + \beta_B^\star$	$\omega_{A,0}^\star(u,z) + \omega_{B,0}^\star(u,z)$	$\omega_{A,1}^\star(u,z) + \omega_{B,1}^\star(u,z)$	Common $\mu^\star$
Average Selection Bias	$E[Y_0 \| D = 1] - E[Y_0 \| D = 0]$	$\dfrac{\mathbb{1}[u \leq p(z)]}{P(D=1)} - \dfrac{\mathbb{1}[u > p(z)]}{P(D=0)}$	0	Leb.$[0,1]$
Average Selection on the Gain	$E[Y_1 - Y_0 \| D = 1] - E[Y_1 - Y_0 \| D = 0]$	$-\omega_1^\star(u,z)$	$\dfrac{\mathbb{1}[u \leq p(z)]}{P(D=1)} - \dfrac{\mathbb{1}[u > p(z)]}{P(D=0)}$	Leb.$[0,1]$

In Table 1, we provide formulas for the weights $\omega_0^\star$ and $\omega_1^\star$ that correspond to a variety of different treatment parameters. Any of these can be taken to be the target parameter $\beta^\star$. Examples include (i) the average treatment effect for the treated (ATT), i.e. the average impact of treatment for individuals who actually take the treatment; (ii) the average treatment effect for the untreated (ATU), i.e. the average impact of treatment for individuals who do not take treatment; (iii) LATE$[\underline{u}, \overline{u}]$, i.e. the average treatment effect for individuals who would take the treatment if their realization of the instrument yielded $p(z) = \overline{u}$, but not if it yielded $p(z) = \underline{u}$; and (iv) the policy relevant treatment effect (PRTE), i.e. the average impact on Y (either gross or per net individual affected) due to a change from the baseline policy to some alternative policy.

For most of the parameters in Table 1, the integrating measure $\mu^\star$ is taken to be Lebesgue measure on $[0, 1]$. However, researchers are sometimes interested in the MTE function itself. For example, Carneiro et al. (2011) and Maestas, Mullen, and Strand (2013) both report estimates of the MTE function for various values of u. Our specification of $\beta^\star$ accommodates this by replacing $\mu^\star$ with the Dirac measure (i.e., a point mass) at some specified point $\overline{u}$ and taking $\omega_0^\star(u, z) = -\omega_1^\star(u, z) = -1$. The resulting target parameter is the MTE function averaged over X, i.e. $E[m(\overline{u}, X)]$.

2.3 What We Know: IV–Like Estimands

A key point for our method is that a set of identified quantities can also be written in a form similar to (9). Consider, for example, the IV estimand that results from using Z as an instrument for D in a linear instrumental variables regression that includes a constant term, but which does not include any other covariates X. Assuming $\text{Cov}(D, Z) \neq 0$, this estimand is given by

$$\beta_{\text{IV}} \equiv \frac{\text{Cov}(Y, Z)}{\text{Cov}(D, Z)}. \tag{10}$$

For example, if $Z \in \{0, 1\}$ is binary, then β_{IV} reduces to the standard Wald estimand

$$\beta_{\text{IV}} = \frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[D \mid Z = 1] - E[D \mid Z = 0]}. \tag{11}$$

Heckman and Vytlacil (2005) show that β_{IV} can also be written in the form (9) as a weighted average of the MTE function. This observation forms the foundation for our intuition that useful information about $\beta^\star$ can be extracted from knowledge of β_{IV}. The next proposition shows that, more generally, any cross moment of Y with a known or identified function of $(D, Z) \equiv (D, X, Z_0)$ can also be expressed as the weighted sum of the two MTR functions, m_0 and m_1. We refer to such cross moments as IV–like estimands.

Proposition 1. *Suppose that $s : \{0,1\} \times \mathbf{R}^{d_z} \mapsto \mathbf{R}$ is a known or identified function of (D, Z) that is measurable and has a finite second moment. We refer to such a function s as an* IV–like specification *and to $\beta_s \equiv E[s(D,Z)Y]$ as an* IV–like estimand*. If (Y, D) are generated according to* (1) *and* (2) *under Assumptions I, then*

$$\beta_s = E\left[\int_0^1 m_0(u,X)\omega_{0s}(u,Z)\,du\right] + E\left[\int_0^1 m_1(u,X)\omega_{1s}(u,Z)\,du\right], \qquad (12)$$

$$\begin{aligned} \textit{where}\ \ &\omega_{0s}(u,Z) \equiv s(0,Z)\mathbb{1}[u > p(Z)] \\ \textit{and}\ \ &\omega_{1s}(u,Z) \equiv s(1,Z)\mathbb{1}[u \leq p(Z)]. \end{aligned}$$

The weights in Proposition 1 can be shown to reduce to the weights for β_{IV} derived by Heckman and Vytlacil (2005) by taking

$$s(d,z) \equiv s(d,x,z_0) = \frac{z_0 - E[Z_0]}{\text{Cov}(D, Z_0)}, \qquad (13)$$

which is an identified function of D (trivially) and Z. As we elaborate further in Appendix B, Proposition 1 applies more broadly to include any well-defined weighted linear IV estimand that uses some function of D and Z as included and excluded instruments for a set of endogenous variables also constructed from D and Z.[5] For example, the ordinary least squares (OLS) estimand corresponds to taking s to be

$$s(d,z) = \frac{d - E[D]}{\text{Var}(D)}.$$

More generally, any subvector of the TSLS or optimal GMM estimands can also be written as an IV–like estimand. Table 2 contains expressions for some notable IV–like estimands.

2.4 From What We Know to What We Want to Know

We now show how to extract information about the target parameter $\beta^\star$ from the general class of IV-like estimands introduced in Section 2.3. Let $\mathcal{S}$ denote some collection of IV–like specifications (i.e. functions $s : \{0,1\} \times \mathbf{R}^{d_z} \mapsto \mathbf{R}$) chosen by the researcher, that each satisfy the conditions set out in Proposition 1. Corresponding to each $s \in \mathcal{S}$ is an IV–like estimand $\beta_s \equiv E[s(D,Z)Y]$. We assume that the researcher has restricted

[5] The phrases "included" and "excluded" instrument are meant in the sense typically introduced in textbook treatments of the linear IV model without heterogeneity.

the pair of MTR functions $m \equiv (m_0, m_1)$ to lie in some *admissible set* $\mathcal{M}$, which incorporates any a priori assumptions that the researcher wishes to maintain about m, such as parametric or shape restrictions. Our goal is to characterize bounds on the values of the target parameter $\beta^\star$ that could have been generated by MTR functions $m \in \mathcal{M}$ that also could have delivered the collection of identified IV–estimands through (12).

To this end, we denote the weighting expression in Proposition 1 as a linear map $\Gamma_s : \mathcal{M} \mapsto \mathbf{R}$, defined for any IV–like specification $s \in \mathcal{S}$ as

$$\Gamma_s(m) \equiv E\left[\int_0^1 m_0(u, X)\omega_{0s}(u, Z)\, du\right] + E\left[\int_0^1 m_1(u, X)\omega_{1s}(u, Z)\, du\right], \qquad (14)$$

where we recall that $\omega_{0s}(u, z) \equiv s(0, z)\mathbb{1}[u > p(z)]$ and $\omega_{1s}(u, z) \equiv s(1, z)\mathbb{1}[u \leq p(z)]$. By Proposition 1, if (Y, D) are generated according to (1) and (2) under Assumptions I, then the MTR functions $m \equiv (m_0, m_1)$ must satisfy $\Gamma_s(m) = \beta_s$ for every IV–like specification $s \in \mathcal{S}$. As a result, m must lie in the set

$$\mathcal{M}_{\mathcal{S}} \equiv \{m \in \mathcal{M} : \Gamma_s(m) = \beta_s \text{ for all } s \in \mathcal{S}\}.$$

The target parameter, $\beta^\star$, can also be expressed as an identified linear map of the MTR functions. From (9), we define this map as $\Gamma^\star : \mathcal{M} \mapsto \mathbf{R}$, with

$$\Gamma^\star(m) \equiv E\left[\int_0^1 m_0(u, X)\omega_0^\star(u, Z) d\mu^\star(u)\right] + E\left[\int_0^1 m_1(u, X)\omega_1^\star(u, Z) d\mu^\star(u)\right]. \qquad (15)$$

It follows that if (Y, D) is generated according to (1) and (2) under Assumptions I, then the target parameter must belong to the identified set

$$\mathcal{B}^\star_{\mathcal{S}} \equiv \{b \in \mathbf{R} : b = \Gamma^\star(m) \text{ for some } m \in \mathcal{M}_{\mathcal{S}}\}.$$

Intuitively, $\mathcal{B}^\star_{\mathcal{S}}$ is the set of values for the target parameter that could have been generated by MTR functions that are consistent with the IV–like estimands. The next result shows that if $\mathcal{M}$ is convex, then $\mathcal{B}^\star_{\mathcal{S}}$ is an interval that can be characterized by solving two convex optimization problems.

Proposition 2. *Suppose that $\mathcal{M}$ is convex. Then either $\mathcal{M}_{\mathcal{S}}$ is empty, and hence $\mathcal{B}^\star_{\mathcal{S}}$ is empty, or else the closure of $\mathcal{B}^\star_{\mathcal{S}}$ (in $\mathbf{R}$) is equal to the interval $[\underline{\beta}^\star, \overline{\beta}^\star]$, where*

$$\underline{\beta}^\star \equiv \inf_{m \in \mathcal{M}_{\mathcal{S}}} \Gamma^\star(m) \qquad \textit{and} \qquad \overline{\beta}^\star \equiv \sup_{m \in \mathcal{M}_{\mathcal{S}}} \Gamma^\star(m). \qquad (16)$$

Table 2: Notable IV–Like Estimands

Estimand	$\boldsymbol{\beta_s}$	$\boldsymbol{s(D,Z)}$	**Notes**
IV slope	$\frac{\text{Cov}(Y,Z_0)}{\text{Cov}(D,Z_0)}$	$\frac{Z_0 - E[Z_0]}{\text{Cov}(D,Z_0)}$	Z_0 scalar
IV (jth component)	$e_j' E[\widetilde{Z}\widetilde{X}']^{-1} E[\widetilde{Z}Y]$	$e_j' E[\widetilde{Z}\widetilde{X}']^{-1}\widetilde{Z}$	$\widetilde{X} \equiv [1, D, X']'$ $\widetilde{Z} \equiv [1, Z_0, X']'$ Z_0 scalar
OLS slope	$\frac{\text{Cov}(Y,D)}{\text{Var}(D)}$	$\frac{D - E[D]}{\text{Var}(D)}$	—
OLS (jth component)	$e_j' E[\widetilde{X}\widetilde{X}']^{-1} E[\widetilde{X}Y]$	$e_j' E[\widetilde{X}\widetilde{X}']^{-1}\widetilde{X}$	$\widetilde{X} \equiv [1, D, X']'$
TSLS (jth component)	$e_j' \left(\Pi E[\widetilde{Z}\widetilde{X}']\right)^{-1} \left(\Pi E[\widetilde{Z}Y]\right)$	$e_j'(\Pi E[\widetilde{Z}\widetilde{X}'])^{-1}\Pi\widetilde{Z}$	$\Pi \equiv E[\widetilde{X}\widetilde{Z}']E[\widetilde{Z}\widetilde{Z}']^{-1}$ Included variables $\widetilde{X}$ Instruments $\widetilde{Z}$

2.5 Sharpness and Point Identification

The set $\mathcal{M}_{\mathcal{S}}$ consists of all MTR functions in $\mathcal{M}$ that are consistent with the IV–like estimands chosen by the researcher. However, $\mathcal{M}_{\mathcal{S}}$ does not necessarily exhaust all of the information available in the data. In particular, $\mathcal{M}_{\mathcal{S}}$ may contain MTR functions that would be ruled out if $\mathcal{S}$ were expanded to include additional IV–like specifications. If this is the case, then incorporating these further specifications could shrink $\mathcal{B}^{\star}_{\mathcal{S}}$.

We examine this issue by considering the conditional means of Y that would be generated through (1) and (2) under Assumptions I by a given MTR pair $m = (m_0, m_1)$. Whenever $0 < p(Z) < 1$, these conditional means can be written as

$$E[Y|D = 0, Z] = E[Y_0|U > p(Z), Z] = \frac{1}{(1 - p(Z))} \int_{p(Z)}^{1} m_0(u, X)\, du, \quad (17)$$

$$\text{and} \qquad E[Y|D = 1, Z] = E[Y_1|U \leq p(Z), Z] = \frac{1}{p(Z)} \int_{0}^{p(Z)} m_1(u, X)\, du. \quad (18)$$

MTR pairs that (almost surely) satisfy (17)–(18) are compatible with the observed conditional means of Y. Our next result shows that any such MTR pair will be in $\mathcal{M}_{\mathcal{S}}$ for any choice of $\mathcal{S}$. Moreover, we show that if $\mathcal{S}$ is chosen correctly, then $\mathcal{M}_{\mathcal{S}}$ will contain *only* MTR pairs that are compatible with the observed conditional means of Y.

Proposition 3. *Suppose that every $m \in \mathcal{M}$ satisfies $E[\int_0^1 m_d(u,X)^2 du] < \infty$ for $d \in \{0,1\}$. If $\mathcal{S}$ contains functions that satisfy the conditions of Proposition 1, then*

$$\{m \in \mathcal{M} : m \textit{ satisfies } (17) \textit{ and } (18) \textit{ almost surely}\} \subseteq \mathcal{M}_{\mathcal{S}}. \tag{19}$$

Moreover, suppose that $\mathcal{S} \equiv \{s(d,z) = \mathbb{1}[d = d']f(z) \text{ for } (d',f) \in \{0,1\} \times \mathcal{F}\}$, where $\mathcal{F}$ is a collection of functions. If the linear span of $\mathcal{F}$ is norm dense in $\mathbf{L}^2(Z) \equiv \{f : \mathbf{R}^{d_z} \mapsto \mathbf{R} \text{ s.t. } E[f(Z)^2] < \infty\}$, then

$$\{m \in \mathcal{M} : m \textit{ satisfies } (17) \textit{ and } (18) \textit{ almost surely}\} = \mathcal{M}_{\mathcal{S}}. \tag{20}$$

Proposition 3 shows that if $\mathcal{S}$ is a sufficiently rich class of functions, then $\mathcal{M}_{\mathcal{S}}$ is sharp in the sense of being the smallest subset of $\mathcal{M}$ that is compatible with the observed conditional means of Y. It follows that $\mathcal{B}^\star_{\mathcal{S}}$ is also the smallest set of values for the target parameter that are consistent with both the conditional means of Y and the assumptions of the model. For example, if $D \in \{0,1\}$ and $Z \in \{0,1\}$, then (20) holds if we take $\mathcal{F} = \{\mathbb{1}[z = 0], \mathbb{1}[z = 1]\}$, so that

$$\mathcal{S} = \{\mathbb{1}[d = 0, z = 0],\ \mathbb{1}[d = 0, z = 1],\ \mathbb{1}[d = 1, z = 0],\ \mathbb{1}[d = 1, z = 1]\}.$$

The information contained in the corresponding IV–like estimands is the same as that contained in the coefficients of a saturated regression of Y on D and Z. More generally, if Z is continuous, then (20) can be satisfied by taking $\mathcal{F}$ to be certain parametric families of functions of Z. For example, if $Z \in \mathbf{R}^{d_z}$, then one such family is the set of half spaces, $\mathcal{F} = \{\mathbb{1}[z \leq z'] : z' \in \mathbf{R}^{d_z}\}$. Other examples can be found in e.g. Bierens (1990) and Stinchcombe and White (1998).

While we view partial identification as the standard case, we emphasize that our analysis does not preclude point identification. Letting $|\mathcal{S}|$ denote the cardinality of $\mathcal{S}$, notice that the restrictions

$$\Gamma_s(m) = \beta_s \text{ for all } s \in \mathcal{S} \tag{21}$$

constitute a linear system of $|\mathcal{S}|$ equations in terms of m. Thus, if $\mathcal{M}$ is finite dimensional, then point identification of the MTR functions is determined by the rank of this linear system. Note that if the MTR functions are point identified, then any target parameter $\beta^\star$ is also point identified.

These observations about point identification are implicit in the work of Brinch et al. (2015). Those authors show that if $\mathcal{M}$ is restricted to be a set of polynomials, then point identification of the MTR functions can be established by considering regressions of Y on $p(Z)$ and D. Their results allow for Z to be discrete, but require the specification of $\mathcal{M}$ to be no richer than the support of $p(Z)$. For example, if Z is binary, their results require $\mathcal{M}$ to only contain MTR pairs that are linear in u.[6] In contrast, our results allow the researcher to specify $\mathcal{M}$ independently of the data.

In some situations, point identification of the target parameter, $\beta^\star$, can also be established when both $|\mathcal{S}|$ and $\mathcal{M}$ are infinite dimensional. Indeed, relationships similar to (17) and (18) have been used previously to establish point identification of the MTE function. For example, Heckman and Vytlacil (1999, 2001c, 2005) and Carneiro et al. (2010, 2011) show that if Z_0 is continuously distributed, then the MTE is point identified over the support of the propensity score. As a consequence, any target parameter $\beta^\star$ whose weights have support contained within the interior of the support of the propensity score will also be point identified. Proposition 3 implies that the same is true in our framework if $\mathcal{S}$ is chosen to be a sufficiently rich collection of functions.

2.6 Nonparametric Shape Restrictions

Our method allows researchers to easily incorporate nonparametric shape restrictions into their specification of the MTR functions. These restrictions can be imposed either on the MTR functions $m = (m_0, m_1)$ or directly on the MTE function $m_1 - m_0$. For example, to impose the monotone treatment response assumption considered by Manski (1997), i.e. that $Y_1 \geq Y_0$ with probability 1, the set $\mathcal{M}$ should be specified to only contain MTR pairs for which $m_1 - m_0$ is non-negative. Similarly, one could assume that $m(\cdot, x)$ is weakly decreasing for every x. This restriction would reflect the assumption that those more likely to select into treatment (those with small realizations of U) are also more likely to have larger gains from treatment. This is similar to the monotone treatment selection assumption of Manski and Pepper (2000). Maintaining combinations of assumptions simultaneously (e.g. both monotone treatment response and monotone treatment selection) is simply a matter of imposing both restrictions on $\mathcal{M}$ at the same time.

Another type of nonparametric shape restriction that can be used to tighten the bounds is separability between the observed (X) and unobserved (U) components. Although restrictive, separability of this sort is standard (often implicit) in applied work using instrumental variables. In our framework, separability between X and U can be imposed by restricting $\mathcal{M}$ to only contain MTR pairs (m_0, m_1) that can be

[6] Recently, Kowalski (2016) has applied the linear case, studied in depth by Brinch et al. (2015), to analyze the Oregon Health Insurance Experiment.

decomposed as

$$m_d(u, x) = m_d^U(u) + m_d^X(x) \quad \text{for } d = 0, 1, \tag{22}$$

where m_d^U and m_d^X are some other functions that can themselves satisfy some shape restrictions. This type of separability implies that the slopes of the MTR functions with respect to u do not vary with x. Alternatively, it is straightforward to interact u and x fully or partially if complete separability is viewed as too strong of a restriction. Specifications like (22) can also be used to mitigate the curse of dimensionality, for example by specifying $m_d^X(x)$ to be a linear function of x.

2.7 Computing the Bounds

Proposition 2 provides convenient numerical methods for computing bounds on the target parameter. In this section, we focus on a particularly tractable computational approach in which we replace the possibly infinite dimensional admissible set of functions $\mathcal{M}$ by a finite dimensional subset $\mathcal{M}_{\text{fd}} \subseteq \mathcal{M}$. The upper bound for the target parameter with this finite dimensional subset is given by

$$\overline{\beta}^{\star}_{\text{fd}} \equiv \sup_{m \in \mathcal{M}_{\text{fd}}} \{\Gamma^{\star}(m) \text{ s.t. } \Gamma_s(m) = \beta_s \text{ for all } s \in \mathcal{S}\}, \tag{23}$$

while $\underline{\beta}^{\star}_{\text{fd}}$ is defined as the analogous infimum.

Suppose that we specify $\mathcal{M}_{\text{fd}}$ as the finite linear basis

$$\mathcal{M}_{\text{fd}} \equiv \left\{ (m_0, m_1) \in \mathcal{M} : m_d(u, x) = \sum_{k=1}^{K_d} \theta_{dk} b_{dk}(u, x) \text{ for } d \in \{0, 1\} \right\}, \tag{24}$$

where $\{b_{dk}\}_{k=1}^{K_d}$ are known basis functions and $\theta \equiv (\theta_0', \theta_1')'$ parameterizes functions in $\mathcal{M}_{\text{fd}}$ with $\theta_d \equiv (\theta_{d1}, \ldots, \theta_{dK_d})'$. The admissible set $\mathcal{M}$ generates an admissible set

$$\Theta \equiv \left\{ (\theta_0, \theta_1) \in \mathbf{R}^{K_0} \times \mathbf{R}^{K_1} : \left(\sum_{k=1}^{K_0} \theta_{0k} b_{0k}, \sum_{k=1}^{K_1} \theta_{k1} b_{1k} \right) \in \mathcal{M} \right\},$$

for the finite dimensional parameter θ. Using the linearity of the mappings $\Gamma^{\star}$ and Γ_s

defined in (14) and (15), we write (23) as

$$\overline{\beta}^{\star}_{\text{fd}} \equiv \sup_{(\theta_0,\theta_1)\in\Theta} \sum_{k=1}^{K_0} \theta_{0k}\Gamma^{\star}_0(b_{0k}) + \sum_{k=1}^{K_1} \theta_{1k}\Gamma^{\star}_1(b_{1k})$$
$$\text{s.t.} \sum_{k=1}^{K_0} \theta_{0k}\Gamma_{0s}(b_{0k}) + \sum_{k=1}^{K_1} \theta_{1k}\Gamma_{1s}(b_{1k}) = \beta_s \text{ for all } s \in \mathcal{S}, \tag{25}$$

where we have decomposed $\Gamma^{\star}(m) \equiv \Gamma^{\star}_0(m_0) + \Gamma^{\star}_1(m_1)$ with

$$\Gamma^{\star}_d(m_d) \equiv E\left[\int_0^1 m_d(u,X)\omega^{\star}_d(u,Z)d\mu^{\star}(u)\right],$$

and similarly for the map Γ_s. If Θ is a polyhedral set, then (25) is a linear program. Linear programs are used routinely in empirical work involving quantile regressions, e.g. Abadie, Angrist, and Imbens (2002), in part because they can be solved quickly and reliably. Whether a given shape restriction on $\mathcal{M}$ translates into Θ being polyhedral depends on the basis functions. In Appendix F, we discuss the Bernstein polynomial basis, which is particularly attractive in this regard.

In some situations, $\mathcal{M}$ can be replaced by a finite dimensional set $\mathcal{M}_{\text{fd}}$ without affecting the bounds on the target parameter, i.e. while ensuring $\overline{\beta}^{\star}_{\text{fd}} = \overline{\beta}^{\star}$. This can be interpreted as an *exact* computational approach for determining nonparametric bounds on the target parameter. For example, suppose that Z has discrete support and that the weight functions $\omega^{\star}_d(u,z)$ for the target parameter are piecewise constant in u. Then define a partition $\{\mathcal{A}_j\}_{j=1}^J$ of $[0,1]$ such that $\omega^{\star}_d(u,z)$ and $\mathbb{1}[u \leq p(z)]$ are constant (as functions of u) on each $\mathcal{A}_j$.[7] Let $\{x_1,\ldots x_L\}$ denote the support of X and then use

$$b_{jl}(u,x) \equiv \mathbb{1}[u \in \mathcal{A}_j, x = x_l] \text{ for } 1 \leq j \leq J \text{ and } 1 \leq l \leq L \tag{26}$$

as the basis functions employed in the construction of $\mathcal{M}_{\text{fd}}$ in (24), with $K_d = JL$. The basis formed by the functions defined in (26) is known as a constant spline, or a Haar basis.

The element of the Haar basis that provides the best mean squared error approximation to a given function $m_d(u,x)$ can be shown to be

$$\Pi m_d(u,x) \equiv \sum_{j=1}^{J}\sum_{l=1}^{L} E[m_d(U,X)|U \in \mathcal{A}_j, X = x_l]b_{jl}(u,x). \tag{27}$$

[7] For example, take $\mathcal{A}_1 \equiv [u_0,u_1]$ and $\mathcal{A}_j \equiv (u_{j-1},u_j]$ for $2 \leq j \leq J$, where $\{u_j\}_{j=0}^J$ are the ordered unique elements of the union of $\{0,1\}$, $\text{supp}\{p(Z)\}$, and the discontinuity points of $\{\omega^{\star}_d(\cdot,z) : d \in \{0,1\}, z \in \text{supp}\{Z\}\}$.

This corresponds to taking $\theta_{d(j,l)} = E[m_d(U,X)|U \in \mathcal{A}_j, X = x_l]$ for an element of (24), with the slight abuse of notation that $k = (j,l)$. The next proposition uses (27) to show that the Haar basis, despite being a finite basis, reproduces nonparametric bounds on the target parameter.

Proposition 4. *Suppose that Z has discrete support and that $\omega^\star_d(u,z)$ are piecewise constant in u. Let $\{\mathcal{A}_j\}_{j=1}^J$ be a partition of $[0,1]$ such that $\omega^\star_d(u,z)$ and $\mathbb{1}[u \leq p(z)]$ are constant on $u \in \mathcal{A}_j$ for any z. Suppose that $\mathcal{M}_{fd} \subseteq \mathcal{M}$ and that $(\Pi m_0, \Pi m_1) \in \mathcal{M}$ for every $(m_0, m_1) = m \in \mathcal{M}$. Then $\overline{\beta}^\star_{fd} = \overline{\beta}^\star$ and $\underline{\beta}^\star_{fd} = \underline{\beta}^\star$.*

Proposition 4 shows that one can solve the infinite dimensional problems defining $\underline{\beta}^\star$ and $\overline{\beta}^\star$ *exactly* by solving (23) with a Haar basis for $\mathcal{M}_{\text{fd}}$. Besides requiring Z to have discrete support, the result also requires $(\Pi m_0, \Pi m_1) \in \mathcal{M}$ for every $m \in \mathcal{M}$, as well as $\mathcal{M}_{\text{fd}} \subseteq \mathcal{M}$. Intuitively, this requires an MTR pair formed from the Haar basis to itself be admissible, as well as to maintain the restrictions encoded in $\mathcal{M}$. For certain restrictions, such as boundedness or monotonicity, this is immediately implied by (27). We demonstrate the use of the Haar basis in both the numerical illustration in Section 4 and the empirical application in Section 6.

3 Applications of the Method

3.1 Partial Identification of Policy Relevant Treatment Effects

The policy relevant treatment effect (PRTE) is the mean effect of changing from a baseline policy to an alternative policy that provides different incentives to participate in treatment (Heckman and Vytlacil, 1999, 2005). In many situations, this policy comparison does not directly correspond to the variation in treatment induced by the instrument, so the PRTE is not point identified. In such cases, researchers can use our method to construct bounds on the PRTE.

To see how one can use our method to draw inference about PRTEs, consider a policy a that operate by changing factors that affect an agent's treatment decision. We follow Heckman and Vytlacil (1999, 2005) in assuming that a has no direct effect on the potential outcomes (Y_0, Y_1), and in particular that it does not affect the set $\mathcal{M}$ of admissible MTR functions. This assumption is similar to the exclusion restriction. The policy can then be summarized by a propensity score and instrument pair (p^a, Z^a). Treatment choice under policy a is given by

$$D^a \equiv \mathbb{1}[U \leq p^a(Z^a)],$$

where U is the same unobservable term as in the selection equation for the status quo policy, D. The outcome of Y that would be observed under the new policy is therefore

$$Y^a = D^a Y_1 + (1 - D^a) Y_0.$$

The PRTE of policy a_1 relative to another policy a_0 is defined as

$$\text{PRTE} \equiv \frac{E[Y^{a_1}] - E[Y^{a_0}]}{E[D^{a_1}] - E[D^{a_0}]} \tag{28}$$

where we assume that $E[D^{a_1}] \neq E[D^{a_0}]$.[8]

In Table 1, we provide weights $\omega_0^\star$ and $\omega_1^\star$ that can be used to express the PRTE as a target parameter $\beta^\star$ with the form given in (9) for different policies a_0 and a_1. The choice of weights depends on the policies being compared. The way in which different policy comparisons translate into different weights is illustrated in Table 1 through the three specific examples considered by Carneiro et al. (2011). Each of these comparisons is between a hypothetical policy a_1 and the status quo policy a_0, the latter of which is characterized by the pair $(p^{a_0}, Z^{a_0}) = (p, Z)$ observed in the data. The comparisons are: (i) an additive α change in the propensity score, i.e. $p^{a_1} = p + \alpha$; (ii) a proportional $(1 + \alpha)$ change in the propensity score, i.e. $p^{a_1} = (1 + \alpha)p$; and (iii) an additive α shift in the distribution the jth component of Z, i.e. $Z^{a_1} = Z + \alpha e_j$, where e_j is the jth unit vector. The first and second of these represent policies that increase (or decrease) participation in the treatment by a given amount α or a proportional amount $(1 + \alpha)$. The third policy represents the effect of shifting the distribution of a variable that impacts treatment choice. In all of these definitions, α is a quantity that could either be hypothesized by the researcher, estimated from some auxiliary choice model, or predicted from the estimated $p(Z)$ under parametric assumptions.

After choosing the weights that correspond to the policy comparison of interest, the procedure in Section 2.7 can be used to estimate bounds for the PRTE. These bounds can be fully nonparametric, but they can also incorporate a priori parametric or shape restrictions if desired. Our statistical inference results in Section 5 allow us to build confidence intervals for the target parameter. In Section 6, we apply these insights to evaluate the PRTEs of alternative subsidy regimes for malaria bed nets.

[8] If this assumption is concerning, one can also define the PRTE as $E[Y^{a_1}] - E[Y^{a_0}]$, see Heckman and Vytlacil (2001a) or pp. 380–381 of Carneiro et al. (2010). Our approach directly applies to this definition as well.

3.2 Extrapolation of Local Average Treatment Effects

Imbens and Angrist (1994) showed that the LATE is point identified under the assumptions considered in this paper. As argued by Imbens (2010, pp. 414–415), it is desirable to report both the LATE, with its high degree of internal validity, but possibly limited external validity, and extrapolations of the LATE to larger or different populations. We now show how to use our method to perform this type of extrapolation, thereby allowing researchers to assess the sensitivity of a given LATE estimate to an expansion (or contraction) of the complier subpopulation.

To extrapolate the LATE, it is useful to connect the LATE parameter to the PRTE. To see the relationship, suppose that there are no covariates X, i.e. $Z = Z_0$, and suppose that Z_0 is binary. Consider the PRTE that results from comparing a policy a_1 under which every agent receives $Z = 1$ against a policy a_0 under which every agent receives $Z = 0$. Choices under these policies are

$$D^{a_0} \equiv \mathbb{1}[U \leq p(0)] \quad \text{and} \quad D^{a_1} \equiv \mathbb{1}[U \leq p(1)],$$

where $p(1) > p(0)$ are the propensity score values in the observed data. The PRTE for this policy comparison is

$$\frac{E[Y^{a_1}] - E[Y^{a_0}]}{E[D^{a_1}] - E[D^{a_0}]} = \frac{E\left[(D^{a_1} - D^{a_0})(Y_1 - Y_0)\right]}{p(1) - p(0)} = E\left[Y_1 - Y_0 \mid U \in (p(0), p(1)]\right], \quad (29)$$

where we used $D^{a_1} - D^{a_0} = \mathbb{1}[U \in (p(0), p(1)]]$. The right-hand side of (29) is precisely the LATE as defined by Imbens and Angrist (1994).

Extrapolation of the LATE amounts to changing $p(0)$, $p(1)$, or both. For example, suppose that the researcher wants to examine the sensitivity of the LATE to an expansion of the complier subpopulation that includes individuals with lower willingness to pay for treatment. This sensitivity check corresponds to shifting $p(1)$ to $p(1) + \alpha$ for $\alpha > 0$. Arguing as in (29), the extrapolated LATE can be shown to be

$$\text{PRTE}(\alpha) = E\left[Y_1 - Y_0 \mid U \in (p(0), p(1) + \alpha]\right]. \quad (30)$$

This PRTE is still a LATE as defined by Heckman and Vytlacil (2005), but one that is not point identified by the IV estimand unless $\alpha = 0$.

While PRTE(α) is not point identified, we show in Table 1 that it can be expressed as a target parameter $\beta^\star$ in the form (9). As a result, we can use our approach to bound PRTE(α). To gain some intuition into such approach, we write PRTE(α) as

$$\begin{aligned}&\text{PRTE}(\alpha)\\ &\quad= \left(\frac{p(1)-p(0)}{\alpha+p(1)-p(0)}\right)\text{LATE} + \left(\frac{\alpha}{\alpha+p(1)-p(0)}\right)\int_{p(1)}^{p(1)+\alpha}\{m_1(u)-m_0(u)\}\,du,\end{aligned}$$

where LATE is the usual point identified LATE in (29). This decomposition shows that the conclusions that can be drawn about PRTE(α) depend on what can be said about $m_1(u) - m_0(u)$ for $u \in [p(1), p(1) + \alpha]$. Therefore, restrictions on the set $\mathcal{M}$ of admissible MTR pairs translate directly into restrictions on the possible values of PRTE(α). For example, if we possess an a priori bound on the support of Y, e.g. Y is binary, then even nonparametric bounds can be informative about PRTE(α).

Our method allows a researcher to formally and transparently balance their desire for robust assumptions against their desire for broader extrapolation. Stronger assumptions are reflected through a more restrictive set of admissible MTR pairs, $\mathcal{M}$. Less ambitious extrapolations are reflected through smaller values of α. For a given α, more restrictive specifications of $\mathcal{M}$ yield smaller bounds, while for a given specification of $\mathcal{M}$, smaller values of α also yield smaller bounds. Both margins can be smoothly adjusted, with point identification obtained as a limiting case as $\alpha \to 0$. We illustrate this tradeoff in our numerical example in Section 4.

3.3 Testable Implications

If the set $\mathcal{M}_{\mathcal{S}}$ is empty, then the model is misspecified: There does not exist a pair of MTR functions m that is both admissible ($m \in \mathcal{M}$), and which could have generated the observed data. If $\mathcal{M}$ is an unrestricted class of functions, then this is attributable to a falsification of selection equation (2) together with Assumptions I. The testable implications of these assumptions for the IV model are well-known, see e.g. Balke and Pearl (1997), Imbens and Rubin (1997) or Kitagawa (2015). On the other hand, if other restrictions have been placed on $\mathcal{M}$, then misspecification could be due either to the failure of Assumptions I, or the specification of $\mathcal{M}$, or both. Our inference results in Section 5 provide a formal test of the null hypothesis that $\mathcal{M}_{\mathcal{S}}$ is nonempty.

This test can also be used to test a variety of null hypotheses about the underlying MTR functions. For example, Table 1 reports the weights that correspond to the quantity $E[Y_0|X, D = 1] - E[Y_0|X, D = 0]$. This quantity is often described as a measure of selection bias, since it captures the extent to which average untreated outcomes differ solely on the basis of treatment status, conditional on observables. The weights provide a linear mapping $\beta^{\star}_{\text{sel}}(m)$ for the amount of selection bias under an MTR pair m. Suppose that we restrict $\mathcal{M}$ to contain only MTR pairs m for which $\beta^\star_{\text{sel}}(m) = 0$. Then rejecting the null hypothesis that $\mathcal{M}_{\mathcal{S}}$ is nonempty could be interpreted as rejecting the hypothesis that there is no selection bias, at least as long as Assumptions I and any other restrictions on $\mathcal{M}$ are not deemed suspect.

Alternatively, one might be interested in testing the joint hypothesis that there is no selection on unobservables; that is, no selection bias and no selection on the gain. Weights for a typical measure of selection on the gain are provided in Table 1. These too provide a linear mapping $\beta^\star_{\text{gain}}$ on the set of MTR pairs $\mathcal{M}$. The hypothesis that there is no selection on unobservables would be rejected if we were to reject the null hypothesis that $\mathcal{M}_{\mathcal{S}}$ is nonempty when $\mathcal{M}$ is restricted to contain only MTR pairs m for which $\beta^\star_{\text{sel}}(m) = \beta^\star_{\text{gain}}(m) = 0$.

4 Numerical Illustration

In this section, we illustrate how to use our method to construct nonparametric bounds on treatment parameters of interest, and how shape restrictions and parametric assumptions can be used to tighten these bounds.

4.1 The Data Generating Process

We consider a simple example with a trinary instrument, $Z \in \{0, 1, 2\}$, with $P(Z = 0) = .5$, $P(Z = 1) = .4$, and $P(Z = 2) = .1$. The propensity score is specified as $p(0) = .35, p(1) = .6$, and $p(2) = .7$. We take the outcome $Y \in \{0, 1\}$ to be binary and restrict $\mathcal{M}$ to contain only MTR pairs that are each bounded between 0 and 1. The data is generated using the MTR functions

$$\begin{aligned} & m_0(u) = .6b^2_0(u) + .4b^2_1(u) + .3b^2_2(u) \\ \text{and} \quad & m_1(u) = .75b^2_0(u) + .5b^2_1(u) + .25b^2_2(u), \end{aligned} \tag{31}$$

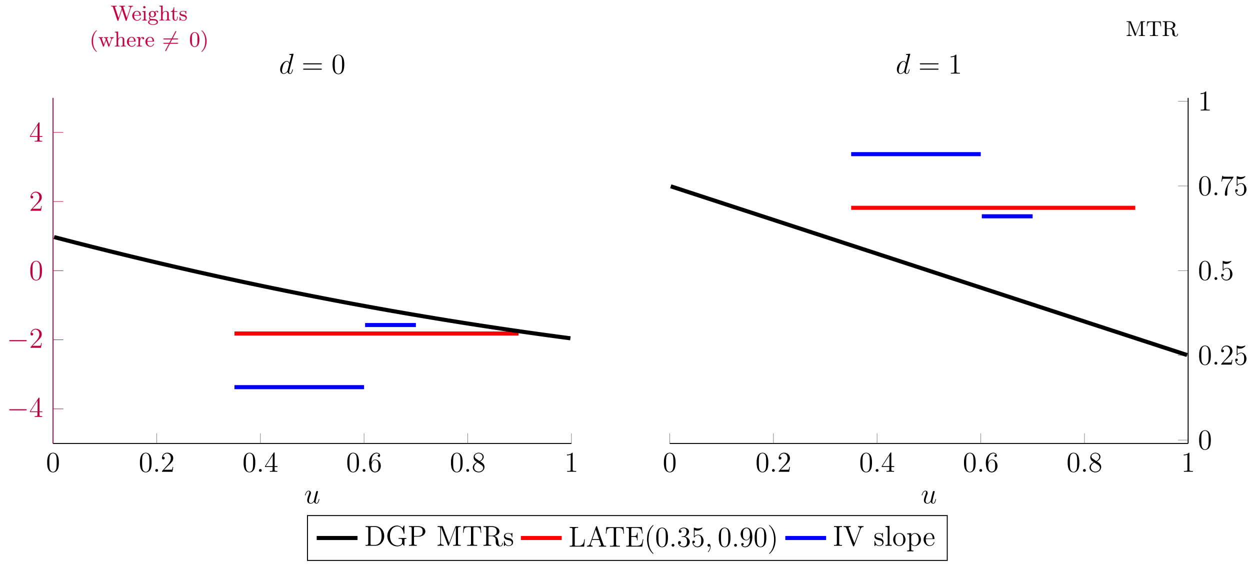
where b^2_k is the kth Bernstein basis polynomial of degree 2.[9]

4.2 IV Estimand, Weights, and Parameter of Interest

Figure 1 contains two plots with two vertical axes. The left plot is for $d = 0$, while the right plot is for $d = 1$, and both vertical axes apply to both plots. The left axis measures weight functions, which are indicated with colored curves and, for the sake of clarity, are not drawn over regions where they are zero. The blue weights correspond

[9] Appendix F contains the definition of the Bernstein polynomials, along with a discussion of some useful properties of the Bernstein polynomial basis.

Figure 1: MTRs Used in the Data Generating Process (DGP)



to ω_{ds} when $s(D, Z)$ is taken to be (13) so that β_s is the IV slope coefficient estimand (10) from using Z as an instrument for D. These weights are positive between the smallest and largest values of the propensity score, i.e. from $p(0) = .35$ to $p(2) = .7$, and they change value at $p(1) = .6$.

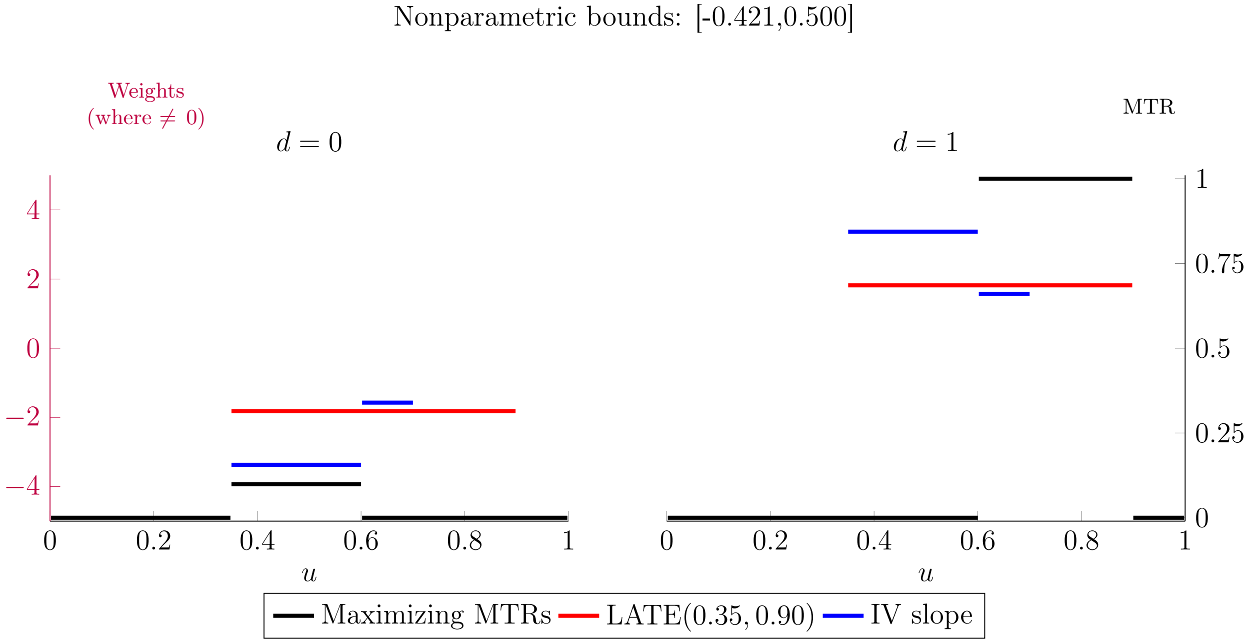
As shown by Imbens and Angrist (1994), three LATEs are nonparametrically point identified in this setting: LATE$(.35, .6)$, LATE$(.35, .7)$ and LATE$(.6, .7)$. Suppose that the researcher wants to examine the sensitivity of these average causal effects to an expansion of the complier subpopulation. Then they might be interested in the target parameter

$$\text{LATE}(.35, .9) \equiv E\left[Y_1 - Y_0 | U \in [.35, .9]\right].$$

The weights for this parameter are drawn in red in Figure 1. As shown in Table 1, the weights are constant over $[.35, .9]$ with magnitude $(.9 - .35)^{-1} \approx 1.81$.

The right vertical axis in Figure 1 measures MTR functions for both the $d = 0$ and $d = 1$ plots. The MTR functions that were used to generate the data, i.e. (31), are plotted in black. These MTR functions imply a value of approximately .074 for the IV slope coefficient. By Proposition 1, this value is equal to the integral of the product of the black and blue curves, summed over the $d = 0$ and $d = 1$ plots. Similarly, these MTR function imply a value of approximately .046 for LATE$(.35, .9)$ through an analogous sum of integrals using the red curves.

Figure 2: Maximizing MTRs When Using Only the IV Slope Coefficient



4.3 Bounds on the Target Parameter

We now illustrate how the researcher can extract information about the target parameter LATE$(.35, .9)$ from the class of IV-like estimands introduced in Section 2.3.

Figure 2 is like Figure 1, except the MTR functions that are plotted yield the nonparametric upper bound on LATE$(.35, .9)$. The nonparametric upper bounds are computed using the Haar basis discussed in Section 2.7.[10] The pair $m \equiv (m_0, m_1)$ in this plot is generated by trying to make m_1 as large as possible, and m_0 as a small as possible, on the support of the red weights, while still yielding a value of .074 for the IV slope coefficient determined by the blue weights. These MTR functions imply a value of .5 for the target parameter LATE$(.35, .9)$, which is the largest value that is consistent with the IV slope estimand. There are multiple pairs of MTR functions with this property. In particular, notice that neither weights are positive over the region $[0, .35]$, so that MTR pairs may be freely adjusted on this region without changing the implied values of the IV slope estimand or LATE$(.35, .9)$. The lower bound of $-.421$ indicated in Figure 2 is obtained through an analogous minimization problem that follows the same logic as the upper bound.

Figure 3 repeats this exercise while including a second IV–like estimand. The second

[10] As discussed in Section 2.7, this amounts to solving two linear programs. We used Gurobi (Gurobi Optimization, 2015) to solve all linear programs reported in the paper.

Figure 3: Maximizing MTRs When Using Both the IV and OLS Slope Coefficients

Nonparametric bounds: [-0.411,0.500]

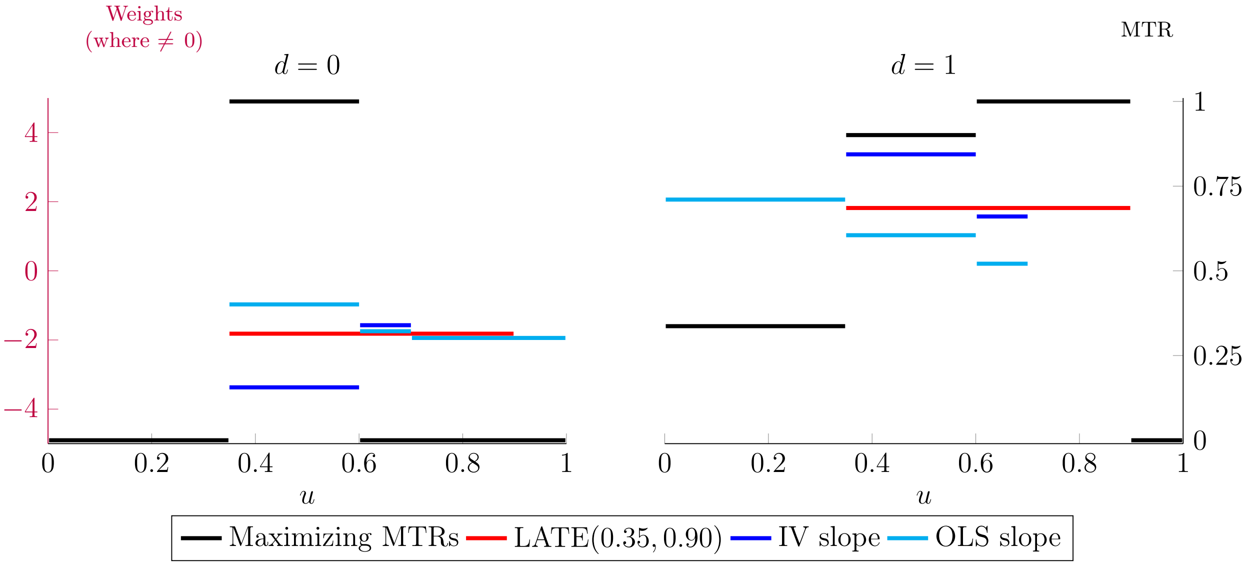


Figure 4: Maximizing MTRs When Breaking the IV Slope into Two Components

Nonparametric bounds: [-0.320,0.407]

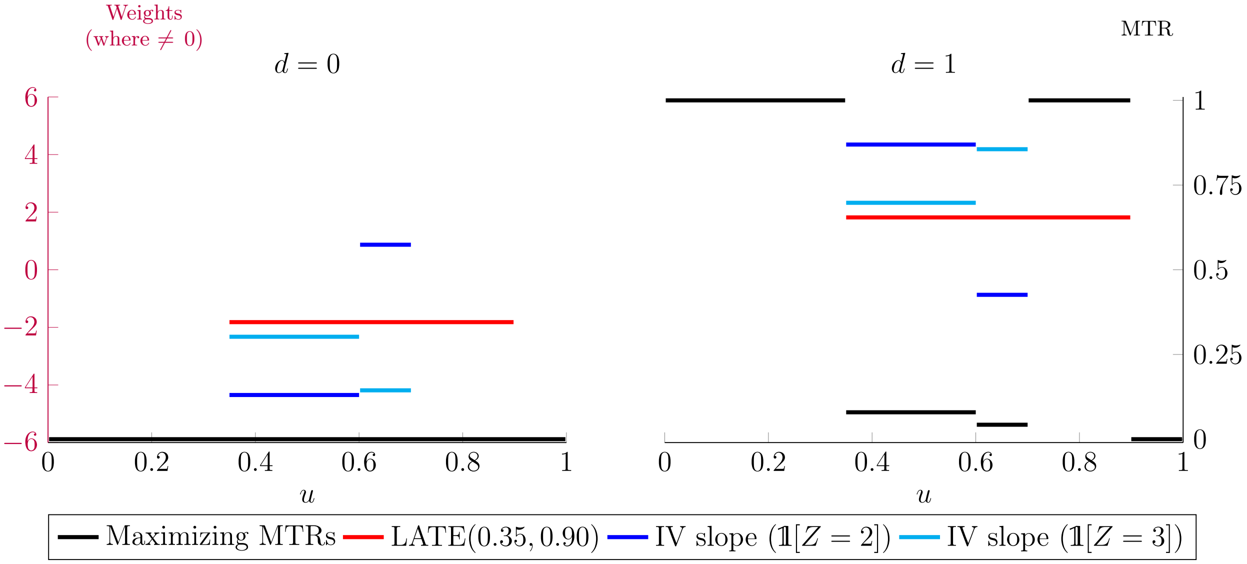
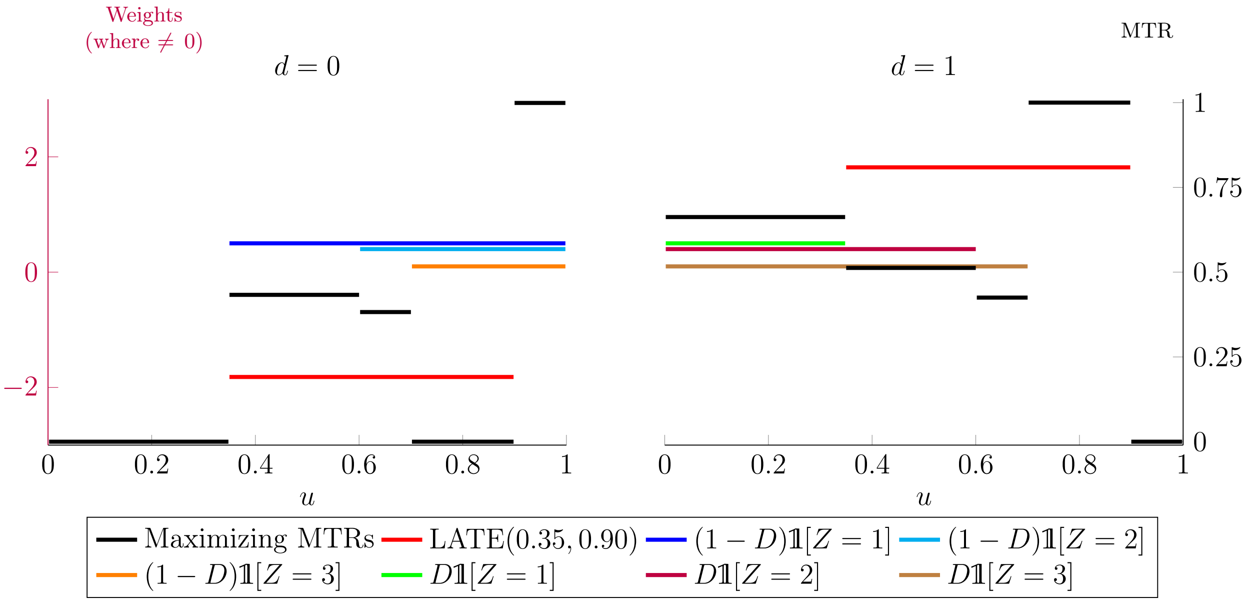


Figure 5: Maximizing MTRs When Using All IV–like Estimands (Sharp Bounds)

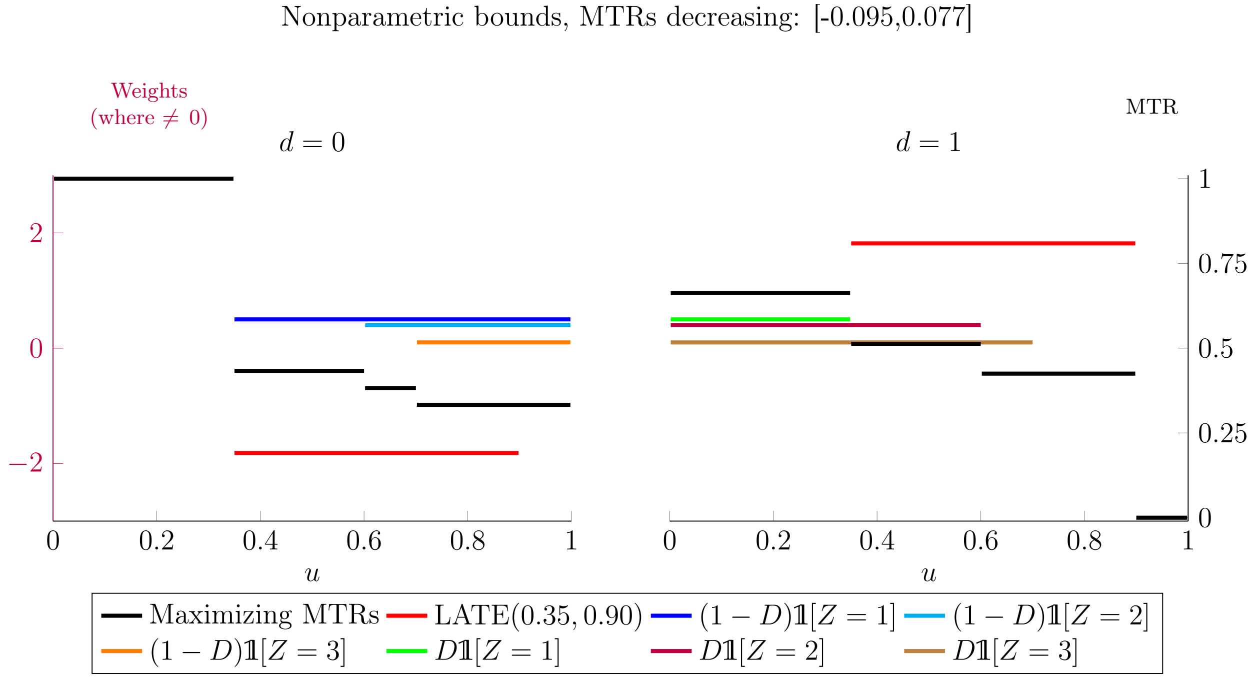
Nonparametric bounds: [-0.138,0.407]



estimand is the OLS slope coefficient, whose weights are drawn in light blue. Notice that, whereas the blue and red weights are symmetric between $d = 0$ and $d = 1$ in the sense of having different signs, but the same magnitude, the light blue weights for the OLS slope coefficient are asymmetric. A maximizing or minimizing MTR pair must yield the implied values for both the IV slope coefficient and the OLS slope coefficient, which is approximately .253 in the simulation. In this DGP, the additional constraint from the OLS slope coefficient has no effect on the upper bound of LATE(.35, .9), but does tighten the lower bound slightly from $-.421$ to $-.411$. In Figure 4, instead of using Z as a single instrument, we split Z into two binary indicators, $\mathbb{1}[Z = 2]$ and $\mathbb{1}[Z = 3]$, to create two IV slope estimands. This tightens both bounds on LATE(.35, .9) considerably. The tightest possible nonparametric bounds are obtained in Figure 5, which includes a collection of six IV–like specifications that is rich enough to satisfy the conditions of Proposition 3. The resulting bounds of $[-.138, .407]$ are the sharp nonparametric bounds for LATE(.35, .9).

The bounds in Figure 5 can be tightened considerably by imposing nonparametric shape restrictions. For example, in Figure 6, the MTR functions are restricted to be decreasing like the DGP MTR functions shown in Figure 1. This tightens the sharp identified set for LATE(.35, .9) by ruling out non-decreasing MTR pairs like the one shown in Figure 5.

Figure 6: Maximizing MTRs When Restricted to be Decreasing



Even tighter bounds can be obtained by also requiring the MTR functions to be smooth. This may be a desirable a priori assumption if one believes that the jump-discontinuous MTR pairs in Figure 6 are sufficiently poorly behaved as to be an unlikely description of the relationship between selection unobservables, U, and potential outcomes, Y_0 and Y_1. For example, in Figure 7, the MTR functions are restricted to be decreasing and characterizable by a polynomial of order 10 or lower. This eliminates the possibility of non-smooth MTR functions like those in Figure 6, and in this example reduces the bounds to $[0, 0.067]$.

4.4 Tradeoffs between Tightness and the Target Parameter

Figure 8 illustrates how the bounds change as the parameter of interest changes. In particular, instead of LATE$(.35, .9)$, we construct bounds on

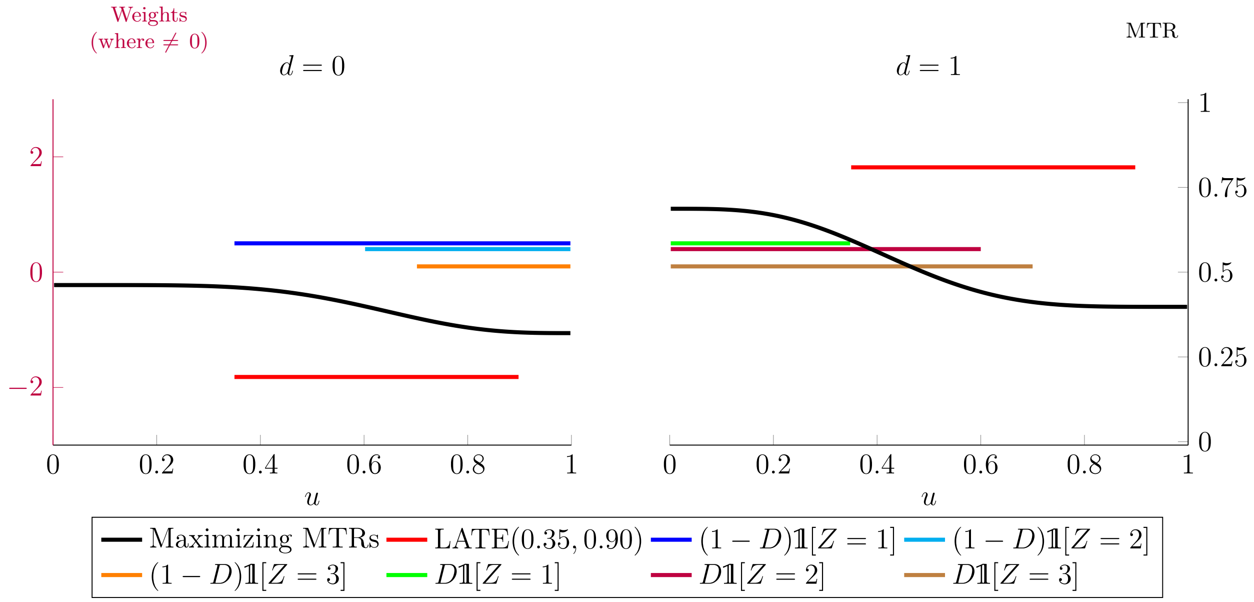
$$\text{LATE}(.35, \overline{u}) \equiv E\left[Y_1 - Y_0 | U \in [.35, \overline{u}]\right],$$

for different values of $\overline{u}$, using the same specification as in Figure 7. Sharp lower and upper bounds on this parameter are given by the blue and red curves, respectively.

As evident from Figure 8, the bounds collapse to a point for $\overline{u} = .6$ and $.7$, i.e the two other points of support for the propensity score. For these values of $\overline{u}$, LATE$(.35, \overline{u})$

Figure 7: Maximizing MTRs When Further Restricted to be a 10th Order Polynomial

Order 9 polynomial bounds, MTRs decreasing: [0.000,0.067]



is a usual point identified LATE as in Imbens and Angrist (1994). For other values of $\overline{u}$ this parameter is not point identified, such as for $\overline{u} = .9$, which is indicated by the dotted vertical line. For $\overline{u}$ between .6 and .7 the bounds are very tight, as shown in the magnified region. As $\overline{u}$ decreases from .6 or increases above .7, the bounds widen. This reflects the increasing difficulty of drawing inference about a parameter the more dissimilar it is from what was observed in the data.

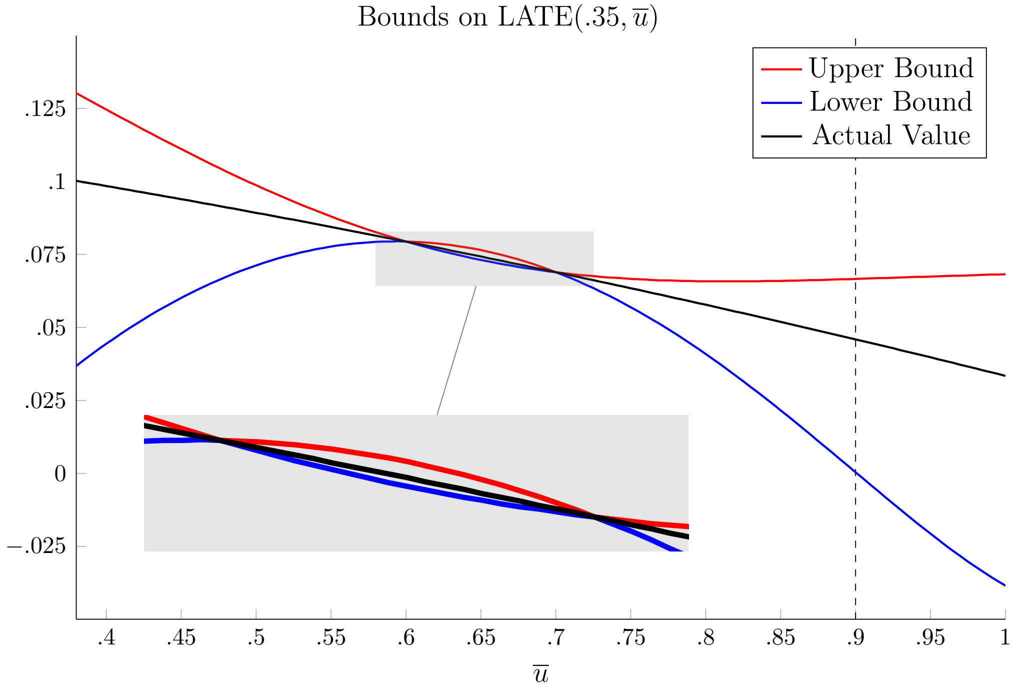
5 Statistical Inference

In this section, we develop a general testing procedure that enables us to conduct statistical inference for the methods and applications described in Sections 2 and 3.

5.1 Notation and Null Hypothesis

Our results will be sufficiently flexible to allow for possible nonparametric specifications of $\mathcal{M}$, potentially with additional shape restrictions, as well as a finite or infinite collection of IV–like specifications, $\mathcal{S}$. To accommodate these possibilities formally, we abstract from finite dimensional Euclidean spaces and work with general complete normed vector spaces, i.e. Banach spaces. For the MTR functions, we assume that $\mathcal{M}$ is a subset of a Banach space $\mathbf{M}$ with norm $\|\cdot\|_{\mathbf{M}}$.

Figure 8: Bounds on a Family of PRTEs



For the IV–like estimands, we identify the relationship between β_s and its specification $s \in \mathcal{S}$ with a function $s \mapsto \beta_s$ that maps each $s \in \mathcal{S}$ into $\beta_s \equiv E[s(D, Z)Y]$. As a notational device, it will sometimes be convenient to extend the domain of the map $s \mapsto \beta_s$. For example, when testing whether a target parameter $\beta^\star \in \mathbf{R}$ is equal to zero, we let $\beta \equiv \{\beta_s : s \in \mathcal{S}\} \cup \{0\}$ so that β then represents the collection of IV–like estimands together with the hypothesized value for the target parameter. On the other hand, when conducting a specification test, we only need to set $\beta = \{\beta_s : s \in \mathcal{S}\}$. As another example, if we were testing the joint hypothesis of no unobserved heterogeneity, we might augment β with two additional elements—one for selection bias and the other for selection on the gain—as per our discussion in Section 3.3. In order to accommodate these different applications, as well as a finite or infinite number of IV–like specifications $\mathcal{S}$, we will view β as an element of a Banach space $\mathbf{B}$ with norm $\|\cdot\|_{\mathbf{B}}$.

A primary concern in our development of a statistical procedure is uniform validity in the underlying distribution of the data. As argued by Imbens and Manski (2004), uniformity considerations are of particular concern when conducting inference in the presence of partial identification. In order to discuss uniformity formally, we now explicitly denote the dependence of various quantities on the distribution of the data, P. So, for example, we now write the IV–like estimand $\beta_s \equiv E[s(D, Z)Y]$ as $\beta_{P,s}$ to emphasize its dependence on the distribution of (Y, D, Z). Similarly, we now write β

with a subscript as β_P. We assume that P lies in a set $\mathbf{P}$ of possible distributions which satisfy regularity conditions that are introduced subsequently.

As discussed in Section 2, both the IV–like estimands and the target parameter are linear functions of the admissible MTR pair $m = (m_0, m_1) \in \mathcal{M} \subseteq \mathbf{M}$. We denote these linear functions by a single map $\Gamma_P : \mathbf{M} \mapsto \mathbf{B}$. For example, if $\mathcal{S} = \{s_1, \ldots, s_{|\mathcal{S}|}\}$ is finite and $\beta_P = (\beta_{P,s_1}, \ldots, \beta_{P,s_{|\mathcal{S}|}})'$, then $\mathbf{B} = \mathbf{R}^{|\mathcal{S}|}$ and

$$\Gamma_P(m) \equiv (\Gamma_{P,s_1}(m), \ldots, \Gamma_{P,s_{|\mathcal{S}|}}(m))',$$

where $\Gamma_{P,s}$ is as defined in (14), but now carries a P subscript to emphasize its dependence on the distribution of the data. To test whether a target parameter is equal to a given hypothesized value $\beta_0^\star$, we let $\beta_P = (\beta_{P,s_1}, \ldots, \beta_{P,s_{|\mathcal{S}|}}, \beta_0^\star)'$ and

$$\Gamma_P(m) \equiv (\Gamma_{P,s_1}(m), \ldots, \Gamma_{P,s_{|\mathcal{S}|}}(m), \Gamma_P^\star(m))', \tag{32}$$

where $\Gamma_P^\star$ is as defined in (15), but now carries a P subscript as well.

Given the flexibility of the parameter $\beta_P \in \mathbf{B}$ and the map $\Gamma_P : \mathbf{M} \mapsto \mathbf{B}$, we can encompass the applications discussed in Section 3 as special cases of the null hypothesis

$$H_0 : P \in \mathbf{P}_0 \qquad H_1 : P \in \mathbf{P} \setminus \mathbf{P}_0, \tag{33}$$

where the set $\mathbf{P}_0$ is defined as

$$\mathbf{P}_0 \equiv \{P \in \mathbf{P} : \Gamma_P(m) = \beta_P \text{ for some } m \in \mathcal{M}\}. \tag{34}$$

The set $\mathbf{P}_0$ consists of distributions of the observed data for which there exists an admissible MTR pair that generates β_P. In the following, we propose a test of (33) that provides uniform size control over the set of distributions $\mathbf{P}$.

By specifying β_P appropriately, we can use our test of (33) for a variety of purposes. For example, a confidence region for a target parameter can be obtained by setting Γ_P as in (32) and conducting test inversion of (33) for $\beta_P = ((\beta_{P,s_1}, \ldots, \beta_{P,s_{|\mathcal{S}|}}, \beta_0^\star)'$ over different hypothesized values $\beta_0^\star$ of the target parameter. Alternatively, to conduct a specification test that employs an infinite collection of moments, we would let $\beta_P = \{\beta_{P,s} : s \in \mathcal{S}\}$ and $\Gamma_P(m) = \{\Gamma_{P,s}(m) : s \in \mathcal{S}\}$.[11]

[11] We discuss this application more formally in Example 5.2.

5.2 The Test Statistic

The applications in Section 3 highlight both the wide array of empirically relevant hypotheses encompassed by (33) and the importance of being flexible in the definitions of $\beta_P \in \mathbf{B}$ and $\mathcal{M} \subseteq \mathbf{M}$. To ensure that our general results apply to such examples, we will simply assume that we posses estimators $\hat{\beta} \in \mathbf{B}$ for the parameter β_P and $\hat{\Gamma} : \mathbf{M} \mapsto \mathbf{B}$ for the map $\Gamma_P : \mathbf{M} \mapsto \mathbf{B}$. Given such estimators, we then construct a minimum distance test statistic for the null hypothesis in (33) as

$$T_n \equiv \inf_{m \in \mathcal{M}_n} \sqrt{n} \|\hat{\beta} - \hat{\Gamma}(m)\|_{\mathbf{B}}, \tag{35}$$

where $\mathcal{M}_n$ is a subset of $\mathcal{M}$ that grows dense in $\mathcal{M}$. When $\mathcal{M}$ is finite dimensional, we can set $\mathcal{M}_n = \mathcal{M}$. We note that, provided $\hat{\Gamma}$ is linear and $\mathcal{M}_n$ is convex, as they are in our applications of interest, T_n is the solution to a convex optimization problem. In Appendix G.2, we describe situations in which (35) can be reformulated as a linear program.

Our analysis uses some properties of convex sets. In order to introduce these properties, we need some additional notation. Let $\mathbf{B}^*$ denote the dual space of $\mathbf{B}$, i.e.

$$\mathbf{B}^* \equiv \{b^* : \mathbf{B} \mapsto \mathbf{R} \text{ s.t. } b^* \text{ is linear and continuous}\},$$

endowed with the norm $\|b^*\|_{\mathbf{B}^*} \equiv \sup_{\|b\|_{\mathbf{B}} \leq 1} |b^*(b)|$. By definition, every $b^* \in \mathbf{B}^*$ is a linear map on $\mathbf{B}$. Note too, that every $b \in \mathbf{B}$ also induces a linear map on $\mathbf{B}^*$ given by $b^* \mapsto b^*(b)$. We emphasize this bilinear relationship with the notation $\langle b^*, b \rangle \equiv b^*(b)$. The weak topology on $\mathbf{B}$ is then the weakest topology under which all of the linear maps $b^* \in \mathbf{B}^*$ (i.e. all $\langle b^*, \cdot \rangle$) are continuous. The weak topology is important for our purposes because it arises naturally in the study of both linear maps and convex sets.

Let $\mathbf{M}^*$ denote the dual space of $\mathbf{M}$. The adjoint of a linear map $\Gamma_P : \mathbf{M} \mapsto \mathbf{B}$ is defined as the unique linear map $\Gamma_P^* : \mathbf{B}^* \mapsto \mathbf{M}^*$ that satisfies

$$\langle b^*, \Gamma_P(m) \rangle = \langle \Gamma_P^*(b^*), m \rangle \tag{36}$$

for every $b^* \in \mathbf{B}^*$ and $m \in \mathbf{M}$. For example, if $\mathbf{M} = \mathbf{R}^{d_m}$ and $\mathbf{B} = \mathbf{R}^{d_\beta}$, then Γ_P can be identified with a $d_\beta \times d_m$ matrix, and its adjoint Γ_P^* is simply its $d_m \times d_\beta$ transpose. Similarly, we denote the adjoint of $\hat{\Gamma}$ as $\hat{\Gamma}^*$. If $\Gamma_P : \mathbf{M} \mapsto \mathbf{B}$ is continuous, then the bilinear maps $(b^*, m) \mapsto \langle b^*, \Gamma_P(m) \rangle$ and $(b^*, m) \mapsto \langle \Gamma_P^*(b^*), m \rangle$ are bounded on any bounded subsets $\mathcal{D} \times \mathcal{M} \subset \mathbf{B}^* \times \mathbf{M}$. Hence, these maps are elements of

$$\ell^\infty(\mathcal{D} \times \mathcal{M}) \equiv \left\{ f : \mathcal{D} \times \mathcal{M} \mapsto \mathbf{R} \text{ s.t. } \sup_{(b,m) \in \mathcal{D} \times \mathcal{M}} |f(b,m)| < \infty \right\}. \tag{37}$$

For our purposes, it will be useful to let $\mathcal{D}$ be the unit ball in $\mathbf{B}^*$, so we define

$$\mathcal{D} \equiv \{b^* \in \mathbf{B}^* : \|b^*\|_{\mathbf{B}^*} \leq 1\}.$$

Intuitively, one can interpret $b^* \in \mathcal{D}$ as a "direction" in the original space $\mathbf{B}$.

For any convex set $\mathcal{C} \subseteq \mathbf{B}$, we define its support function $\nu(\cdot, \mathcal{C}) : \mathcal{D} \mapsto \mathbf{R}$ as

$$\nu(b^*, \mathcal{C}) \equiv \sup_{b \in \mathcal{C}} \langle b^*, b \rangle. \tag{38}$$

Intuitively, $\nu(b^*, \mathcal{C})$ indicates how far one can move in the direction b^* while staying within the set $\mathcal{C}$. Letting $\hat{\Gamma}(\mathcal{M}_n) \equiv \{b \in \mathbf{B} : b = \hat{\Gamma}(m) \text{ for some } m \in \mathcal{M}_n\}$, we obtain by duality (see, e.g., Theorem 5.13.1 of Luenberger (1969))

$$T_n = \sup_{b^* \in \mathcal{D}} \sqrt{n} \left\{ \langle b^*, \hat{\beta} \rangle - \nu(b^*, \hat{\Gamma}(\mathcal{M}_n)) \right\} = \sup_{b^* \in \mathcal{D}} \inf_{m \in \mathcal{M}_n} \sqrt{n} \left\{ \langle b^*, \hat{\beta} - \hat{\Gamma}(m) \rangle \right\}. \tag{39}$$

In other words, the minimum distance test statistic T_n can also be computed by finding the direction $b^* \in \mathcal{D}$ for which $\hat{\beta}$ is the farthest away from its projection onto $\hat{\Gamma}(\mathcal{M}_n)$. We will heavily rely on the dual characterization in (39) in our analysis.

We illustrate our results with two examples that we will return to throughout our discussion. The first example is extremely simple and intended to exposit the nature of our statistical approximations, while the second example uses infinite dimensional spaces, and is used to clarify the more abstract aspects of our analysis.

Example 5.1. Suppose that there are no covariates ($Z = Z_0$), $Y \in \{0, 1\}$ is binary, we have chosen a single IV–like specification ($\mathcal{S} = \{s\}$), and we have modeled the MTR functions as $m_d(u) = \theta_d u$ for $\theta_d \in \mathbf{R}$ and $d = 0, 1$. A given MTR function (m_0, m_1) is then fully characterized by a pair (θ_0, θ_1). Therefore, we set $\mathbf{M} = \mathbf{R}^2$ and define

$$\mathcal{M} \equiv \{(\theta_0, \theta_1) \in \mathbf{R}^2 : \theta_d \in [0, 1] \text{ for } d \in \{0, 1\}\}, \tag{40}$$

where the restriction $\theta_d \in [0, 1]$ reflects $Y \in \{0, 1\}$. Suppose that our goal is to test whether the specified model is compatible with the single chosen IV–like restriction (recall $\mathcal{S} = \{s\}$). To do this, we set $\mathbf{B} = \mathbf{R}$, so that $\beta_P = E[s(D, Z)Y]$, and use some calculus to write the map $\Gamma_P : \mathbf{R}^2 \mapsto \mathbf{R}$ as

$$\Gamma_P(m) = \sum_{d\in\{0,1\}} E\left[\int_0^1 \theta_d u\, \omega_{ds}(u,Z)\, du\right]$$
$$= \begin{bmatrix} \frac{1}{2}E[s(0,Z)(1-p^2(Z))] \\ \frac{1}{2}E[s(1,Z)p^2(Z)] \end{bmatrix}' \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \equiv \begin{bmatrix} \Gamma_{P,0} \\ \Gamma_{P,1} \end{bmatrix}' \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \equiv \Gamma_P \theta. \tag{41}$$

Hence, the linear map Γ_P is just the two dimensional row vector $\Gamma_P \equiv (\Gamma_{P,0}, \Gamma_{P,1})$ defined in (41). Given a parametric or nonparametric estimator $\hat{p}(Z)$ for $p(Z)$, we then set $\hat{\beta} \in \mathbf{R}$ to be the sample analog of $E[s(D,Z)Y]$ and take $\hat{\Gamma} \equiv (\hat{\Gamma}_0, \hat{\Gamma}_1) \in \mathbf{R}^2$ to be the sample analog of Γ_P. For this example, the duality result in (39) states that

$$T_n = \inf_{m=(\theta_0,\theta_1)'} \left\{ \sqrt{n}|\hat{\beta} - \hat{\Gamma}(m)| \text{ s.t. } \theta_d \in [0,1] \right\}$$
$$= \sup_{-1\le b^*\le 1} \inf_{\theta_0,\theta_1} \left\{ \sqrt{n} b^*(\hat{\beta} - \theta_0\hat{\Gamma}_0 - \theta_1\hat{\Gamma}_1) \text{ s.t. } \theta_d \in [0,1] \right\}, \tag{42}$$

where we used that $\mathbf{B}^* = \mathbf{B} = \mathbf{R}$, and $\mathcal{D} = \{b^* \in \mathbf{R} : |b^*| \le 1\}$. ■

Example 5.2. As in Example 5.1, we assume that there are no covariates ($Z = Z_0$) and that $Y \in \{0,1\}$ is binary. However, now we consider testing the null hypothesis that agents more likely to select into treatment also have a higher expected benefit from treatment. To this end, we take $\mathcal{M}$ to be the infinite dimensional set of functions

$$\mathcal{M} = \{(m_0, m_1) \text{ s.t. } m_d : [0,1] \mapsto [0,1],\ m_1 - m_0 \text{ monotonically increasing}\}, \tag{43}$$

and $\mathcal{M}_n \subset \mathcal{M}$ a finite dimensional subset, such as pairs of Bernstein polynomials of finite order $K < \infty$; see Appendix F. For any $m = (m_0, m_1)$, we let $\|m\|^2_{\mathbf{M}} \equiv \int (m_0(u)^2 + m_1(u)^2) du$ and note that $\mathcal{M} \subset \mathbf{M}$ for $\mathbf{M} \equiv \mathbf{L}^2([0,1]) \times \mathbf{L}^2([0,1])$ where $\mathbf{L}^2([0,1]) \equiv \{f : [0,1] \mapsto \mathbf{R} \text{ s.t. } \int_0^1 f(u)^2\, du < \infty\}$.

We also consider an infinite set of IV–like specifications $\mathcal{S}$ that satisfies the conditions of Proposition 3. By Stinchcombe and White (1998), such a set $\mathcal{S}$ can be of the form

$$\mathcal{S} = \{s(\cdot;\lambda) : \lambda \in \Lambda \subset \mathbf{R}^{d_\lambda}\}, \tag{44}$$

where $s(\cdot;\lambda) : \{0,1\} \times \mathbf{R}^{d_z} \mapsto \mathbf{R}$ is known and $\lambda \in \Lambda$ for $\Lambda \subset \mathbf{R}^{d_\lambda}$ a finite dimensional compact parameter space.[12] We can then identify β_P with a function on Λ via

$$\beta_P(\lambda) \equiv E[s((D,Z);\lambda)Y]. \tag{45}$$

[12] E.g., take $s((d,z);\lambda) = (1-d)\exp\{\lambda_0 + z'\lambda_1\} + d\exp\{\lambda_2 + z'\lambda_3\}$ with $\lambda = (\lambda_0, \lambda_1, \lambda_2, \lambda_3)$.

Provided $\lambda \mapsto E[s((D,Z);\lambda)]$ is continuous, we can set $\mathbf{B} = \mathbf{C}(\Lambda)$ with $\mathbf{C}(\Lambda) \equiv \{f : \Lambda \mapsto \mathbf{R} \text{ s.t. } f \text{ is continuous}\}$ and $\|\cdot\|_{\mathbf{B}} = \|\cdot\|_\infty$ with $\|f\|_\infty \equiv \sup_{\lambda\in\Lambda} |f(\lambda)|$. The map $\Gamma_P : \mathcal{M} \mapsto \mathbf{C}(\Lambda)$ assigns to each MTR pair (m_0, m_1) the continuous function on Λ defined by

$$\Gamma_P(m)(\lambda) \equiv E\left[\int_{p(Z)}^{1} m_0(u)s_0((0,Z);\lambda)du\right] + E\left[\int_{0}^{p(Z)} m_1(u)s_1((0,Z);\lambda)du\right]. \quad (46)$$

As in Example 5.1, given an estimator $\hat{p}(Z)$ for the propensity score $p(Z)$, we may obtain estimators $\hat{\beta}$ and $\hat{\Gamma}$ through the sample analogs of (45) and (46). The dual space of $\mathbf{B} = \mathbf{C}(\Lambda)$ is the set of signed Borel measures with bounded variation, i.e. $\mathbf{T}(\Lambda) \equiv \{\tau : \int_\Lambda d|\tau|(\lambda) \le 1\}$.[13] For this example, the duality result in (39) states that

$$\begin{aligned} T_n &= \inf_{m\in\mathcal{M}_n} \sqrt{n}\|\hat{\beta} - \hat{\Gamma}(m)\|_\infty \\ &= \sup_{\tau\in\mathbf{T}(\Lambda)} \inf_{m\in\mathcal{M}_n} \sqrt{n}\int (\hat{\beta}(\lambda) - \hat{\Gamma}(m)(\lambda))d\tau(\lambda) \\ &= \sup_{\lambda\in\Lambda} \inf_{m\in\mathcal{M}_n} \sqrt{n}|\hat{\beta}(\lambda) - \hat{\Gamma}(m)(\lambda)|, \end{aligned} \quad (47)$$

where the final equality follows after noting that, for any $m \in \mathcal{M}_n$, the optimal τ puts measure plus or minus one on the $\lambda \in \Lambda$ that maximizes $|\hat{\beta}(\lambda) - \hat{\Gamma}(m)(\lambda)|$. ■

5.3 Distributional Approximation

In this section, we obtain an approximation for the distribution of the test statistic, T_n. To do this, we maintain the following assumptions.

Assumptions S

S.1 *$\mathcal{M} \subseteq \mathbf{M}$ is convex and compact in the weak topology.*

S.2 *The maps $\Gamma_P : \mathbf{M} \mapsto \mathbf{B}$ and $\hat{\Gamma} : \mathbf{M} \mapsto \mathbf{B}$ are linear and continuous.*

S.3 *There are tight and centered jointly Gaussian processes $(\mathbb{G}_{P,\beta}, \mathbb{G}_{P,\Gamma}) \in \mathbf{B}\times\boldsymbol{\ell}^\infty(\mathcal{D}\times\mathcal{M})$ such that $\sqrt{n}\{\hat{\Gamma} - \Gamma_P\} = \mathbb{G}_{P,\Gamma} + O_p(\delta_n^c)$ and $\sqrt{n}\{\hat{\beta} - \beta_P\} = \mathbb{G}_{P,\beta} + O_p(\delta_n^c)$ uniformly in $P \in \mathbf{P}$ for some sequence $\delta_n^c \downarrow 0$.*

S.4 *The Gaussian processes $(\mathbb{G}_{P,\beta}, \mathbb{G}_{P,\Gamma})$ satisfy*

$$\sup_{P\in\mathbf{P}} E[\|\mathbb{G}_{P,\beta}\|_{\mathbf{B}}] < \infty \quad \textit{and} \quad \sup_{P\in\mathbf{P}} E\left[\sup_{m\in\mathcal{M}} \|\mathbb{G}_{P,\Gamma}(m)\|_{\mathbf{B}}\right] < \infty.$$

[13] See, e.g., Corollary 14.15 in Aliprantis and Border (2006).

S.5 *For every $m \in \mathcal{M}$ there is a $\Pi_n m \in \mathcal{M}_n \subseteq \mathcal{M}$ and a sequence $\delta_n^s \downarrow 0$ such that*

$$\sup_{P\in\mathbf{P}} \sqrt{n}\|\Gamma_P(m - \Pi_n m)\|_{\mathbf{B}} \leq \delta_n^s$$

$$\textit{and} \quad \sup_{P\in\mathbf{P}} E\left[\sup_{m\in\mathcal{M}} \|\mathbb{G}_{P,\Gamma}(m) - \mathbb{G}_{P,\Gamma}(\Pi_n m)\|_{\mathbf{B}}\right] \leq \delta_n^s.$$

Assumption S.1 is our main requirement for $\mathcal{M}$, which is that it is convex and compact with respect to the weak topology. If $\mathbf{M}$ is reflexive (i.e. $\mathbf{M} = (\mathbf{M}^*)^*$), as in Example 5.2, then S.1 is satisfied if $\mathcal{M}$ is convex, closed, and bounded under $\|\cdot\|_{\mathbf{M}}$. Assumption S.2 formalizes our requirement that the maps Γ_P and $\hat{\Gamma}$ are linear and continuous. Our main statistical assumption is S.3, which can be interpreted as requiring that a central limit theorem applies to the estimators $\hat{\beta}$ and $\hat{\Gamma}$ uniformly in $P \in \mathbf{P}$ at a rate $\delta_n^c \downarrow 0$. In our applications, $\sqrt{n}\{\hat{\beta} - \beta_P\}$ and $\sqrt{n}\{\hat{\Gamma} - \Gamma_P\}$ are asymptotically equivalent to empirical processes and Assumption S.3 can hold with rates as fast as $\delta_n^c = \log(n)/\sqrt{n}$ (Koltchinskii, 1994; Rio, 1994); see also Chernozhukov et al. (2015). Assumption S.4 imposes moment conditions on the processes $(\mathbb{G}_{P,\beta}, \mathbb{G}_{P,\Gamma})$. Assumption S.5 requires the approximation error δ_n^s introduced from employing $\mathcal{M}_n$ in place of $\mathcal{M}$ to vanish sufficiently fast. We note that, due to Assumption S.3, there are no constraints on the rate of growth of the sieve $\mathcal{M}_n$.

Under the null hypothesis, there exists an $m \in \mathcal{M}$ such that $\Gamma_P(m) = \beta_P$. By the definition of the support function, it follows that

$$\langle b^*, \beta_P\rangle - \nu(b^*, \Gamma_P(\mathcal{M})) \leq 0 \quad \text{for all } b^* \in \mathcal{D}. \tag{48}$$

Since $0 \in \mathcal{D}$, there is always a $b^* \in \mathcal{D}$ for which (48) holds with equality. This suggests that the supremum in the dual representation of T_n (see (39)) is attained at a b^* in

$$\mathcal{D}_P(\kappa_n^u) \equiv \{b^* \in \mathcal{D} : \langle b^*, \beta_P\rangle - \nu(b^*, \Gamma_P(\mathcal{M})) \geq -\kappa_n^u\}, \tag{49}$$

for any positive sequence κ_n^u that converges to zero, but not too quickly. The set $\mathcal{D}_P(\kappa_n^u)$ shares a similarity with the moment inequalities literature, in which the set of moment inequalities that are "close" to binding plays a crucial role in inference; see Canay and Shaikh (2016) and the references cited therein. Analogously, we use $\mathcal{D}_P(\kappa_n^u)$ to define the distributional approximation

$$\mathbb{U}_{P,n}(\kappa_n^u) \equiv \sup_{b^*\in\mathcal{D}_P(\kappa_n^u)} \inf_{m\in\mathcal{M}} \left\{\langle b^*, \mathbb{G}_{P,\beta} - \mathbb{G}_{P,\Gamma}(m)\rangle \text{ s.t. } \sqrt{n}\langle b^*, \beta_P - \Gamma_P(m)\rangle \leq \delta_n^s\right\}.$$

Our next result provides some properties of $\mathbb{U}_{P,n}(\kappa_n^u)$ as an approximation to T_n.

Theorem 1. *Suppose that Assumptions S.1–S.5 hold. Let κ_n^u be any sequence for which $\sqrt{n}\kappa_n^u \uparrow \infty$, and define $\mathcal{D}_P(\kappa_n^u)$ as in (49). Then, uniformly over $P \in \mathbf{P}$,*

$$T_n = \sup_{b^* \in \mathcal{D}_P(\kappa_n^u)} \inf_{m \in \mathcal{M}} \{\langle b^*, \mathbb{G}_{P,\beta} - \mathbb{G}_{P,\Gamma}(m)\rangle + \sqrt{n}\langle b^*, \beta_P - \Gamma_P(m)\rangle\} + o_p(1). \tag{50}$$

Moreover, there exists a sequence $\xi_n = O_p(\delta_n^c + \delta_n^s)$ such that for any $P \in \mathbf{P}_0$,

$$T_n \leq \mathbb{U}_{P,n}(\kappa_n^u) + \xi_n, \tag{51}$$

uniformly over $P \in \mathbf{P}$.

The first result in Theorem 1 provides an asymptotic expansion for our test statistic, T_n. We will use this expansion to assess the quality of our subsequent bounds. In particular, (50) can be used to conclude that the quantiles of the pointwise asymptotic distribution of T_n can fail to deliver uniform size control without additional assumptions; see Example 5.1 below. As a result, we will instead construct a test with a critical value based on the quantiles of $\mathbb{U}_{P,n}(\kappa_n^u)$. By showing that $\mathbb{U}_{P,n}(\kappa_n^u)$ is an asymptotic upper bound for T_n, Theorem 1 lays the foundations for establishing that such a test can deliver uniform size control. It is worth noting that the results of Theorem 1 hold for any sequence κ_n^u that satisfies the stated conditions. Thus, Theorem 1 actually establishes a family (indexed by sequences κ_n^u) of uniformly valid upper bounds for T_n.

We now revisit Example 5.1, which is useful for clarifying the nature of our approximations, and then Example 5.2, which helps illustrate the content of our assumptions.

Example 5.1 (continued)**.** Consider a sequence of distributions $P_{n,\gamma}$, such that $\beta_{P_{n,\gamma}} = (1 + \gamma/\sqrt{n})$ and $\Gamma_{P_{n,\gamma}} = (1, \gamma/\sqrt{n})$ for some $\gamma \geq 0$. Suppose for simplicity that $\mathbb{Z}_\beta$ and $\mathbb{Z}_\Gamma$ are independent standard normal random variables, and that

$$\sqrt{n}\{\hat{\beta} - \beta_{P_{n,\gamma}}\} = \mathbb{Z}_\beta + o_p(1) \qquad \sqrt{n}\{\hat{\Gamma} - \Gamma_{P_{n,\gamma}}\} = (0, \mathbb{Z}_\Gamma) + o_p(1). \tag{52}$$

Direct calculation shows that for any $\kappa_n^u \downarrow 0$ that satisfies $\sqrt{n}\kappa_n^u \uparrow \infty$, (50) simplifies to

$$T_n = \max\{\min\{\mathbb{Z}_\beta + \gamma, \mathbb{Z}_\beta - \mathbb{Z}_\Gamma\}, 0\} + o_p(1). \tag{53}$$

It is important to notice that the quantiles of the distributional approximation in (53) are *increasing* in γ. As a result, any critical value c_n that satisfies

$$\lim_{n\to\infty} P_{n,0}(T_n > c_n) = \alpha \tag{54}$$

will deliver an asymptotic rejection probability *larger* than α along any contiguous sequence $P_{n,\gamma}$ with $\gamma > 0$. This result contrasts with, e.g., Andrews and Soares (2010), in which the pointwise limit is also the "least favorable." Employing the upper bound $\mathbb{U}_{P,n}(\kappa_n^u)$ addresses this difficulty. In particular,

$$\mathbb{U}_{P_{n,\gamma},n}(\kappa_n^u) = \max\{\mathbb{Z}_\beta - \mathbb{Z}_\Gamma, 0\} + o_p(1) \tag{55}$$

for any sequence $\kappa_n^u \downarrow 0$. In this example, (55) corresponds to the "least favorable" value of the local parameter γ in (53), i.e. $\gamma = +\infty$. ■

Example 5.2 (continued)**.** Since $\mathbf{M}$ is reflexive, Assumption S.1 is satisfied whenever $\mathcal{M}$ is closed and bounded, which is the case for $\mathcal{M}$ in (43). Assumption S.2 is satisfied for $\Gamma_P : \mathcal{M} \mapsto \mathbf{B}$ as in (46) (and its plug-in analog) by Jensen's inequality, provided that $\mathcal{S}$ has a square integrable envelope. Sufficient conditions for Assumption S.3 can be found by establishing a (uniform over $P \in \mathbf{P}$) asymptotic expansion

$$\sqrt{n}\{\hat{\Gamma} - \Gamma_P\}(m)(\lambda) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(D_i, Z_i, m, \lambda) + o_p(1) \tag{56}$$

and ensuring that $\{f(y,d,z) = ys(d,z) : s \in \mathcal{S}\}$ and $\{f(d,z) = \psi(d,z,m,\lambda) : (m,\lambda) \in \mathcal{M} \times \Lambda\}$ are suitably Donsker classes.[14] Assumption S.4 reduces to a uniform (in $P \in \mathbf{P}$) moment bound on the supremum of the approximating Gaussian processes. To verify Assumption S.5, one can apply results on sieve approximation errors, such as those available in Chen (2007) and the references cited therein. ■

5.4 Bootstrap Approximation

The important implication of Theorem 1 is that the quantiles of $\mathbb{U}_{P,n}(\kappa_n^u)$ are valid critical values for the test statistic T_n. We now address the problem of estimating these quantiles. For clarity of exposition, we divide our analysis into two steps, each of which addresses a distinct challenge.

[14] Such requirement may necessitate further restricting $\mathcal{M}$ by, e.g., imposing a bound on derivatives.

5.4.1 An Infeasible Approximation

We first derive an infeasible bootstrap approximation for the distribution of $\mathbb{U}_{P,n}(\kappa_n^u)$. To this end, we first rewrite $\mathbb{U}_{P,n}(\kappa_n^u)$ as the saddle point problem

$$\begin{aligned}\mathbb{U}_{P,n}(\kappa_n^u) = \sup_{b^*\in\mathcal{D}} \inf_{m\in\mathcal{M}} &\langle b^*, \mathbb{G}_{P,\beta} - \mathbb{G}_{P,\Gamma}(m)\rangle \\ \text{s.t. (i) } &\langle b^*, \beta_P\rangle - \nu(b^*, \Gamma_P(\mathcal{M})) \geq -\kappa_n^u, \\ \text{(ii) } &\sqrt{n}\langle b^*, \beta_P - \Gamma_P(m)\rangle \leq \delta_n^s, \end{aligned} \tag{57}$$

where constraint (i) corresponds to $b^* \in \mathcal{D}_P(\kappa_n^u)$, as defined in (49). Characterizing $\mathbb{U}_{P,n}(\kappa_n^u)$ as (57) emphasizes that its distribution is determined by only three essential unknowns: The objective function and the two constraints of optimization.

The objective function in (57) depends on the unknown distribution P of the data only through the asymptotic distribution of the chosen estimators, i.e. $(\mathbb{G}_{P,\beta}, \mathbb{G}_{P,\Gamma})$. While the bootstrap is not valid for estimating the distribution of T_n, it can nonetheless often consistently estimate the distribution of these estimators (Fang and Santos, 2014). We therefore assume the existence of suitable estimators $(\hat{\mathbb{G}}_\beta, \hat{\mathbb{G}}_\Gamma)$ of the distribution of $(\mathbb{G}_{P,\beta}, \mathbb{G}_{P,\Gamma})$ as follows.

Assumptions S (continued)

S.6 $(\hat{\mathbb{G}}_\beta, \hat{\mathbb{G}}_\Gamma) = (\mathbb{G}^{bs}_{P,\beta}, \mathbb{G}^{bs}_{P,\Gamma}) + O_p(\delta_n^c)$ *in* $\mathbf{B} \times \ell^\infty(\mathcal{D}\times\mathcal{M})$ *uniformly in* $P \in \mathbf{P}$*, with* $(\mathbb{G}^{bs}_{P,\beta}, \mathbb{G}^{bs}_{P,\Gamma})$ *independent of* $\{(Y_i, D_i, Z_i)\}_{i=1}^\infty$ *and equal in law to* $(\mathbb{G}_{P,\beta}, \mathbb{G}_{P,\Gamma})$.

S.7 $\mathcal{M}_n \subseteq \mathcal{M}$ *is convex and closed.*

Assumption S.6 is our main bootstrap requirement. It requires the existence of a consistent bootstrap procedure that estimates the law of $(\mathbb{G}_{P,\beta}, \mathbb{G}_{P,\Gamma})$ uniformly in $P \in \mathbf{P}$ at the rate δ_n^c, i.e. the same rate as in Assumption S.3. Typically, for standard bootstrap analogs $\hat{\beta}^{\text{bs}}$ and $\hat{\Gamma}^{\text{bs}}$ of $\hat{\beta}$ and $\hat{\Gamma}$ respectively, the estimators $\hat{\mathbb{G}}_\beta$ and $\hat{\mathbb{G}}_\Gamma$ would correspond to setting $\hat{\mathbb{G}}_\beta = \sqrt{n}\{\hat{\beta}^{\text{bs}} - \hat{\beta}\}$ and $\hat{\mathbb{G}}_\Gamma = \sqrt{n}\{\hat{\Gamma}^{\text{bs}} - \hat{\Gamma}\}$. However, note that Assumption S.6 also allows for estimation of the law of $(\mathbb{G}_{P,\beta}, \mathbb{G}_{P,\Gamma})$ through alternative resampling procedures, such as a score or weighted bootstrap, subsampling, or the m out of n bootstrap. Assumption S.7 imposes the regularity condition that $\mathcal{M}_n$ is closed and convex.

Having estimators $(\hat{\mathbb{G}}_\beta, \hat{\mathbb{G}}_\Gamma)$ enables us to mimic the stochastic behavior of the objective function in (57) by simply employing a plug-in sample analog, i.e. by replacing $\langle b^*, \mathbb{G}_{P,\beta} - \mathbb{G}_{P,\Gamma}(m)\rangle$ with $\langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m)\rangle$. A similar approach is also effective for handling constraint (i) in (57), which corresponds to imposing $b^* \in \mathcal{D}_P(\kappa_n^u)$. Specifically, we replace constraint (i) in (57) with the constraint that $b^* \in \hat{\mathcal{D}}_n$, where

$$\hat{\mathcal{D}}_n \equiv \{b^* \in \mathcal{D} : \langle b^*, \hat{\beta}\rangle - \nu(b^*, \hat{\Gamma}(\mathcal{M}_n)) \geq -\hat{\kappa}_n^u\} \tag{58}$$

and $\hat{\kappa}_n^u$ is a bandwidth selected by the researcher. We discuss data-driven choices of $\hat{\kappa}_n^u$ in Section 5.5.2. However, we note here that setting $\hat{\mathcal{D}}_n = \mathcal{D}$ ($\hat{\kappa}_n^u = \infty$) is always a valid choice.

Turning to (ii) in (57), we note that using a simple plug-in analog of this constraint can lead to a lack of size control. Instead, we construct a possibly random set $\hat{\mathcal{M}}_n$ that, heuristically, should satisfy the following condition asymptotically:

$$\{m \in \mathcal{M} : \Gamma_P(m) = \beta_P\} \subseteq \hat{\mathcal{M}}_n. \tag{59}$$

Under the null hypothesis, there exists an $m_P \in \mathcal{M}$ such that $\Gamma_P(m_P) = \beta_P$, and therefore any $m \in \hat{\mathcal{M}}_n$ that satisfies $\langle b^*, \Gamma_P(m_P - m)\rangle \leq 0$ must also satisfy constraint (ii) in (57). These observations imply that under the null hypothesis

$$\{m \in \hat{\mathcal{M}}_n : \langle b^*, \Gamma_P(m)\rangle = \nu(b^*, \Gamma_P(\hat{\mathcal{M}}_n))\} \subseteq \{m \in \mathcal{M} : \langle b^*, \beta_P - \Gamma_P(m)\rangle \leq 0\} \tag{60}$$

whenever (59) holds. While setting $\hat{\mathcal{M}}_n = \mathcal{M}$ guarantees condition (59), for power considerations it may be advisable to instead set

$$\hat{\mathcal{M}}_n \equiv \{m \in \mathcal{M}_n : \|\hat{\beta} - \hat{\Gamma}(m)\|_{\mathbf{B}} \leq \hat{\kappa}_n^m\} \tag{61}$$

for some bandwidth $\hat{\kappa}_n^m$ that is chosen by the researcher.[15] Note that taking $\hat{\kappa}_n^m = \infty$ corresponds to choosing $\hat{\mathcal{M}}_n = \mathcal{M}$. While we allow $\hat{\kappa}_n^u$ and $\hat{\kappa}_n^m$ to differ from each other, these bandwidths are in fact closely related and can often be selected through a common procedure. See Section 5.5.2 for additional details.

At this point, our discussion suggests that the distribution of

$$I_n(\Gamma_P) \equiv \sup_{b^* \in \hat{\mathcal{D}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \left\{\langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m)\rangle \text{ s.t. } \langle b^*, \Gamma_P(m)\rangle = \nu(b^*, \Gamma_P(\hat{\mathcal{M}}_n))\right\}, \tag{62}$$

conditional on the data, should approximate the distribution of $\mathbb{U}_{P,n}(\kappa_n^u)$. Intuitively, $I_n(\Gamma_P)$ imitates (57) but replaces constraint (i) with $b^* \in \hat{\mathcal{D}}_n$, and replaces constraint (ii) with $m \in \{\tilde{m} \in \hat{\mathcal{M}}_n : \langle b^*, \Gamma_P(\tilde{m})\rangle = \nu(b^*, \Gamma_P(\hat{\mathcal{M}}_n))\}$. The bootstrap statistic

[15] Loosely speaking, we should expect the minimizer of $T_n = \inf_{m \in \mathcal{M}_n} \|\hat{\beta} - \hat{\Gamma}(m)\|_{\mathbf{B}}$ to "converge" to $\{m \in \mathcal{M} : \beta_P = \Gamma_P(m)\}$. We emphasize, however, that the set of distributions $\mathbf{P}$ that we consider allows for the possibility that the set $\{m \in \mathcal{M} : \beta_P = \Gamma_P(m)\}$ is not consistently estimable uniformly in $P \in \mathbf{P}$.

$I_n(\Gamma_P)$ is infeasible due to its dependence on the unknown map $\Gamma_P : \mathcal{M} \mapsto \mathbf{B}$. We address this challenge in the next section. Nevertheless, establishing the properties of $I_n(\Gamma_P)$ provides us with a useful intermediate result before analyzing the bootstrap statistic itself. For this purpose, we impose the following requirements on $\hat{\kappa}_n^u$ and $\hat{\kappa}_n^m$:

Assumptions S (continued)

S.8 *$\hat{\kappa}_n^u$ satisfies*

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}} P\left(\frac{\hat{\kappa}_n^u}{\kappa_n^u} > (1+\delta)\right) = 1$$

for some $\delta > 0$ and some non-random sequence κ_n^u such that $\sqrt{n}\kappa_n^u \uparrow \infty$.

S.9 *$\hat{\kappa}_n^m$ satisfies $\liminf_{C\uparrow\infty} \liminf_{n\to\infty} \inf_{P\in\mathbf{P}} P(\sqrt{n}\hat{\kappa}_n^m \geq C) = 1$ for every $C > 0$.*

Assumption S.8 allows $\hat{\kappa}_n^u$ to be random, but requires it to be (asymptotically) larger than the nonrandom sequence κ_n^u, which determines the random variable $\mathbb{U}_{P,n}(\kappa_n^u)$ whose distribution we aim to approximate. Assumption S.9 requires $\sqrt{n}\hat{\kappa}_n^m$ to diverge to infinity asymptotically as well. By allowing $\hat{\kappa}_n^u$ and $\hat{\kappa}_n^m$ to be random, Assumptions S.8 and S.9 enable us to employ data-driven choices of bandwidths. We discuss possible choices in Section 5.5.2.

Let $\mathbb{U}_{P,n}^{\text{bs}}(\kappa_n^u)$ be the random variable defined by

$$\mathbb{U}_{P,n}^{\text{bs}}(\kappa_n^u) \equiv \sup_{b^*\in\mathcal{D}_P(\kappa_n^u)} \inf_{m\in\mathcal{M}} \left\{\langle b^*, \mathbb{G}_{P,\beta}^{\text{bs}} - \mathbb{G}_{P,\Gamma}^{\text{bs}}(m)\rangle \text{ s.t. } \sqrt{n}\langle b^*, \beta_P - \Gamma_P(m)\rangle \leq \delta_n^s\right\}.$$

Notice that, given Assumption S.6, the random variables $\mathbb{U}_{P,n}^{\text{bs}}(\kappa_n^u)$ and $\mathbb{U}_{P,n}(\kappa_n^u)$ share the same distribution, and hence the same quantiles. However, in contrast to $\mathbb{U}_{P,n}(\kappa_n^u)$, the random variable $\mathbb{U}_{P,n}^{\text{bs}}(\kappa_n^u)$ is independent of the data, and hence its quantiles conditional on the data are equal to the *unconditional* quantiles of $\mathbb{U}_{P,n}(\kappa_n^u)$ that we desire for inference. These observations motivate our next intermediate result.

Lemma 5.1. *Suppose that Assumptions S.1–S.7 and S.9 hold. Then, for any sequence κ_n^u that satisfies Assumption S.8, and for which $\sqrt{n}\kappa_n^u \uparrow \infty$, there exists a sequence $\xi_n^{bs} \in \mathbf{R}$ such that $\xi_n^{bs} = O_p(\delta_n^c)$ uniformly over $P \in \mathbf{P}_0$, and such that*

$$\mathbb{U}_{P,n}^{bs}(\kappa_n^u) \leq I_n(\Gamma_P) + \xi_n^{bs} \tag{63}$$

for any $P \in \mathbf{P}_0$.

Together with Theorem 1, Lemma 5.1 suggests that the quantiles of the bootstrap statistic $I_n(\Gamma_P)$, conditional on the data, can be used as critical values for the test statistic, T_n. In the next section, we use Lemma 5.1 to establish an analogous result for a feasible bootstrap statistic. Before proceeding, we revisit Example 5.1 to illustrate the approximations in Lemma 5.1, and Example 5.2 to clarify our assumptions.

Example 5.1 (continued)**.** We continue to consider the sequence $P_{\gamma,n}$, which satisfies (52) with $\beta_{P_{\gamma,n}} = (1 + \gamma/\sqrt{n})$ and $\Gamma_{P_{\gamma,n}} = (1, \gamma/\sqrt{n})$ for some $\gamma > 0$. Let

$$\hat{\mathbb{G}}_\beta = \mathbb{Z}_\beta^{\text{bs}} + o_p(1) \qquad \hat{\mathbb{G}}_\Gamma = (0, \mathbb{Z}_\Gamma^{\text{bs}}) + o_p(1) \tag{64}$$

where $\mathbb{Z}_\beta^{\text{bs}}$ and $\mathbb{Z}_\Gamma^{\text{bs}}$ are independent standard normal random variables. Provided that $\hat{\kappa}_n^u$ satisfies Assumption S.8 and converges in probability to zero, it is straightforward to verify that $\hat{\mathcal{D}}_n$ converges in probability to $[0,1]$, which is the set of "directions" $b^* \in [-1,1]$ satisfying $\langle b^*, \beta_{P_{\gamma,n}}\rangle = \nu(b^*, \Gamma_{P_{\gamma,n}}(\mathcal{M}))$. Similarly, provided that $\hat{\kappa}_n^m$ converges in probability to zero, it follows that with probability tending to one along $P_{\gamma,n}$

$$\{(\theta_0, \theta_1) = m \in \mathcal{M} : m = (1, \theta_1) \text{ for } \theta_1 \in [0,1]\} \subset \hat{\mathcal{M}}_n. \tag{65}$$

However, (65) implies $\{m \in \hat{\mathcal{M}}_n : \langle b^*, \Gamma_P(m)\rangle = \nu(b^*, \Gamma_P(\hat{\mathcal{M}}_n))\} = \{(1,1)\}$, and hence

$$I_n(\Gamma_{P_{\gamma,n}}) = \sup_{b^* \in [0,1]} \langle b^*, \mathbb{Z}_\beta^{\text{bs}} - \mathbb{Z}_\Gamma^{\text{bs}}\rangle + o_p(1) = \mathbb{U}^{\text{bs}}_{P_{\gamma_n},n}(\kappa_n^u) + o_p(1), \tag{66}$$

where the final equality follows from (55). Therefore, in this simple example, $I_n(\Gamma_P)$ and $\mathbb{U}^{\text{bs}}_{P,n}(\kappa_n^u)$ are asymptotically equivalent. ■

Example 5.2 (continued)**.** As in the verification of Assumption S.3, sufficient conditions for establishing Assumption S.6 can be found by guaranteeing that an asymptotically linear expansion such as (56) also holds for the bootstrap estimators. For verifying Assumption S.7, note that convex sets are closed in the weak topology of $\mathbf{M}$ if and only if they are closed under $\|\cdot\|_{\mathbf{M}}$. ■

5.4.2 The Bootstrap Statistic

The main obstacle to using the statistic $I_n(\Gamma_P)$ for inference is that it depends on the unknown linear map $\Gamma_P : \mathcal{M} \mapsto \mathbf{B}$. We now address this challenge and construct a feasible bootstrap statistic. Before proceeding, we present a lemma that makes the dependence of $I_n(\Gamma_P)$ on the unknown map Γ_P more transparent. Recall from (36) that $\Gamma_P^* : \mathbf{B}^* \mapsto \mathbf{M}^*$ denotes the adjoint of Γ_P.

Lemma 5.2. *Suppose that Assumptions S.1–S.5, S.7, and S.9 are satisfied. Then*

$$I_n(\Gamma_P) = \sup_{b^* \in \hat{\mathcal{D}}_n} \sup_{\tilde{m} \in \hat{\mathcal{M}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \left\{ \langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m) \rangle \ \textit{s.t.} \ \langle \Gamma_P^*(b^*), m - \tilde{m} \rangle \geq 0 \right\} \tag{67}$$

with probability tending to one, uniformly over $P \in \mathbf{P}_0$.

Lemma 5.2 shows that Γ_P enters the optimization problem that defines $I_n(\Gamma_P)$ solely through the bilinear constraint involving b^*, m, and $\tilde{m}$. Given an estimator, $\hat{\Gamma}$, it would be natural to approximate $I_n(\Gamma_P)$ by simply employing the plug-in analog $I_n(\hat{\Gamma})$. However, the feasible set in (67) changes discontinuously in Γ_P. As a result of this, using the quantiles of $I_n(\hat{\Gamma})$ as critical values can fail to control size. For this reason, we instead consider the least favorable critical value obtained from a "neighborhood" of our estimator $\hat{\Gamma}$.

Defining an appropriate neighborhood of $\hat{\Gamma}$ can be challenging when $\mathbf{M}$ and/or $\mathbf{B}$ are infinite dimensional. In particular, Assumption S.3 is too weak to guarantee that $\hat{\Gamma}$ is consistent for Γ_P with respect to the operator norm.[16] Instead, we build neighborhoods using the weak (operator) topology. Let $\mathbf{M}_n$ be the closed linear span of $\mathcal{M}_n$ (in $\mathbf{M}$), and let $\mathbf{M}_n^*$ denote its dual space. For any $b^* \in \mathbf{B}^*$, define

$$\begin{aligned} \hat{\mathcal{G}}_n(b^*) &\equiv \{g \in \mathbf{M}_n^* : |\langle g - \hat{\Gamma}^*(b^*), v \rangle| \leq \hat{\kappa}_n^g \text{ for all } v \in \mathcal{V}_n\} && (68) \\ &= \{g \in \mathbf{M}_n^* : |\langle g, v \rangle - \langle b^*, \hat{\Gamma}(v) \rangle| \leq \hat{\kappa}_n^g \text{ for all } v \in \mathcal{V}_n\}, && (69) \end{aligned}$$

where $\hat{\kappa}_n^g \downarrow 0$ and $\mathcal{V}_n \subseteq \mathbf{M}$ are chosen by the researcher. Typically, we take $\mathcal{V}_n \subseteq \mathcal{M}_n$ and use specification (69) in place of (68) so as to avoid having to compute the adjoint $\hat{\Gamma}^*$. We provide guidance for choosing $\hat{\kappa}_n^g$ in Section 5.5.2.

We can now define our feasible bootstrap statistic as

$$T_n^{\text{bs}} \equiv \sup_{b^* \in \hat{\mathcal{D}}_n} \sup_{(g, \tilde{m}) \in \hat{\mathcal{G}}_n(b^*) \times \hat{\mathcal{M}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \left\{ \langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m) \rangle \text{ s.t. } \langle g, m - \tilde{m} \rangle \geq 0 \right\}. \tag{70}$$

While (70) looks like an unwieldy optimization problem, we show in Appendix G.3 that it can be reformulated as a bilinear program. While bilinear programs are not convex, they can be provably solved to global optimality by algorithms that combine McCormick relaxations (McCormick, 1976) with spatial branch-and-bound strategies.[17]

[16] More precisely, $\sup_{\|m\|_{\mathbf{M}} \leq 1} \|\{\hat{\Gamma} - \Gamma_P\}(m)\|_{\mathbf{B}}$ need not converge to zero in probability.

[17] A good software implementation of these types of algorithms is the BARON solver developed by Tawarmalani and Sahinidis (2005). Although not provably convergent to global optimality, we have found that more common locally optimal solvers, such as KNITRO (Byrd, Nocedal, and Waltz, 2006), almost always find the global optimum for our problem, and much more quickly than BARON. We use KNITRO

For our purposes, an attractive property of these algorithms is that termination after a set amount of time always provides an upper bound on the actual optimal value, T_n^{bs}. Critical values constructed from these upper bounds may be conservative, however they will still provide size control.

We use the following two assumptions to establish the properties of T_n^{bs}.

Assumptions S (continued)

S.10 $\hat{\kappa}_n^g$ *satisfies* $\liminf_{C\uparrow\infty}\liminf_{n\to\infty}\inf_{P\in\mathbf{P}} P(\sqrt{n}\hat{\kappa}_n^g \geq C) = 1$ *for every* $C > 0$.

S.11 *There exists a* $\lambda \in \mathbf{R}$ *and an* $m_c \in \mathcal{M}_n$ *such that* $\mathcal{V}_n \subseteq \lambda(\mathcal{M}_n - m_c)$ *for every* n.

Assumption S.10 requires the bandwidth $\hat{\kappa}_n^g$ (which can be data-dependent) to not converge to zero at a rate faster than $\sqrt{n}$. Assumption S.11 relates the set of functions $\mathcal{V}_n$ used to construct $\hat{\mathcal{G}}_n(b^*)$ to the sieve $\mathcal{M}_n$. This requirement is imposed as a simple sufficient condition to ensure that the process $\langle b^*, \sqrt{n}\{\hat{\Gamma} - \Gamma_P\}(v)\rangle$ is asymptotically tight on $\ell^\infty(\mathcal{D} \times \mathcal{V})$ for $\mathcal{V} = \bigcup_{n=1}^\infty \mathcal{V}_n$. For applications in which $\mathbf{B}$ and $\mathbf{M}$ are finite dimensional, the sets $\mathcal{V}_n$ may chosen so that $\hat{\mathcal{G}}_n(b^*) = \{g : \|g - \hat{\Gamma}(v)\| \leq \hat{\kappa}_n^g\}$, where $\|\cdot\|$ is the standard Euclidean norm. Notice that Assumption S.11 does not place any requirements on the "rate of growth" of $\mathcal{V}_n$.

The next theorem describes the properties of our proposed bootstrap statistic.

Theorem 2. *Suppose that Assumptions S.1–S.7 and S.9–S.11 are satisfied. Then, for any sequence* κ_n^u *that satisfies Assumption S.8, as well as* $\sqrt{n}\kappa_n^u \uparrow \infty$*, there exists a sequence* $\xi_n^{bs} \in \mathbf{R}$*, with* $\xi_n^{bs} = O_p(\delta_n^c)$ *uniformly in* $P \in \mathbf{P}_0$*, and such that*

$$\mathbb{U}_{P,n}^{bs}(\kappa_n^u) \leq T_n^{bs} + \xi_n^{bs} \tag{71}$$

for any $P \in \mathbf{P}_0$.

Theorems 1 and 2 build the foundation for testing the null hypothesis in (33) by comparing the test statistic T_n to critical values obtained from the quantiles of the bootstrap statistic T_n^{bs}, conditional on the data. We establish the properties of such a test in the next section.

Our choice of T_n^{bs} is informed by considerations of computational reliability in realistically sized problems. For simple problems, a less conservative, but computationally more challenging critical value may also be available. To see this, start by defining the $1 - \alpha$ quantile of $I_n(\Gamma)$ (conditional on the data) as

$$q_{1-\alpha}(\Gamma) \equiv \inf_{c\in\mathbf{R}} \left\{c \text{ s.t. } P(I_n(\Gamma)|\{Y_i, D_i, Z_i\}_{i=1}^n) \geq 1 - \alpha\right\}. \tag{72}$$

for our empirical application in Section 6, but we have checked a subset of the results using BARON and found them to be nearly identical.

In principle, it is possible to use the least favorable critical value given by

$$\sup_{\Gamma:\mathbf{M}\mapsto\mathbf{B}} \left\{ q_{1-\alpha}(\Gamma) \text{ s.t. } |\langle \Gamma^*(b^*) - \hat{\Gamma}^*(b^*), v\rangle| \leq \hat{\kappa}_n^g \text{ for all } (b^*, v) \in \hat{\mathcal{D}}_n \times \mathcal{V}_n \right\}. \tag{73}$$

Intuitively, by using T_n^{bs}, we are computing the quantile of the maximum over a neighborhood of $\hat{\Gamma}$, instead of the maximum *of the quantile* over such a neighborhood. Whenever (73) is solvable, it will yield a critical value that provides better power than T_n^{bs}. We illustrate this in Example 5.1 below. Our recommendation is to use (73) when feasible, but we focus on T_n^{bs} due to its wider (computational) applicability.

We conclude this section by revisiting Examples 5.1 and 5.2.

Example 5.1 (continued)**.** We return again to the sequences $P_{\gamma,n}$, which satisfied (52) and (65) with $\beta_{P_{\gamma},n} = (1+\gamma/\sqrt{n})$ and $\Gamma_{P_{\gamma},n} = (1, \gamma/\sqrt{n})$. In this simple example, $\mathbf{M} = \mathbf{R}^2$ and $\mathbf{B} = \mathbf{R}$, so that $\hat{\Gamma}^* : \mathbf{R} \mapsto \mathbf{R}^2$ and $\Gamma_P^* : \mathbf{R} \mapsto \mathbf{R}^2$ are given by

$$\Gamma_P^*(b^*) = b^* \left(1, \frac{\gamma}{\sqrt{n}}\right)' \qquad \text{and} \qquad \hat{\Gamma}^*(b^*) = b^* \left(1, \frac{\mathbb{Z}_\Gamma + \gamma}{\sqrt{n}}\right)' + o_p(1) \tag{74}$$

for any $b^* \in \mathbf{R}$. Setting $\mathcal{V}_n = \{(1,0), (0,1)\}$ implies that for any $b^* \geq 0$,

$$\left\{ (g_1, g_2)' = g \in \mathbf{R}^2 : g = (1, c)' \text{ and } |c| \leq \frac{1}{\sqrt{n}} \right\} \subseteq \hat{\mathcal{G}}_n(b^*) \tag{75}$$

with probability tending to one along $P_{\gamma,n}$. For notational clarity, we also define

$$\mathbb{L}(g) \equiv \sup_{(b^*, \tilde{m}) \in [0,1] \times \hat{\mathcal{M}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \{ \langle b^*, \mathbb{G}_{P,\beta}^{\mathrm{bs}} - \mathbb{G}_{P,\Gamma}^{\mathrm{bs}}(m) \rangle \text{ s.t. } \langle g, m - \tilde{m} \rangle \geq 0 \}, \tag{76}$$

and note that the set inclusion in (65) implies that

$$\mathbb{L}(g) = \begin{cases} \max\{0, \mathbb{Z}_\beta^{\mathrm{bs}} - \mathbb{Z}_\Gamma^{\mathrm{bs}}\} & \text{if } g_2 > 0 \\ \max\{0, \mathbb{Z}_\beta^{\mathrm{bs}} - (\mathbb{Z}_\Gamma^{\mathrm{bs}})_+\} & \text{if } g_2 = 0 \\ \max\{0, \mathbb{Z}_\beta^{\mathrm{bs}}\} & \text{if } g_2 < 0 \end{cases} \tag{77}$$

for any $(1, g_2)' = g \in \mathbf{R}^2$, where $(\mathbb{Z}_\Gamma^{\mathrm{bs}})_+ = \max\{\mathbb{Z}_\Gamma^{\mathrm{bs}}, 0\}$. We conclude that

$$T_n^{\mathrm{bs}} = \max\{0, \mathbb{Z}_\beta^{\mathrm{bs}}, \mathbb{Z}_\beta^{\mathrm{bs}} - \mathbb{Z}_\Gamma^{\mathrm{bs}}\} + o_p(1).$$

Note that, at least for this example, the less conservative critical values given in (73) can actually be solved for analytically. These critical values correspond to the quantiles of $\max\{0, \mathbb{Z}_\beta^{\mathrm{bs}} - \mathbb{Z}_\Gamma^{\mathrm{bs}}\}$, which are also equal to the quantiles of $\mathbb{U}_{P,n}(\kappa_n^u)$. ■

Example 5.2 (continued)**.** We use this example to illustrate the computation of T_n^{bs} in a more abstract setting. Recall that $\mathbf{B} = \mathbf{C}(\Lambda)$ and $\mathbf{M} = \mathbf{L}^2([0,1]) \times \mathbf{L}^2([0,1])$ so that $\mathbf{B}^* = \mathbf{T}(\Lambda)$ is the set of signed Borel measures of bounded variation, while $\mathbf{M}^* = \mathbf{L}^2([0,1]) \times \mathbf{L}^2([0,1])$. Hence, for this example we obtain

$$\begin{aligned}
\hat{\mathcal{D}}_n &= \left\{ b^* : \int_\Lambda d|b^*|(\lambda) \leq 1 \text{ and } \int_\Lambda \hat{\beta}(\lambda) db^*(\lambda) \geq \sup_{m \in \mathcal{M}_n} \int_\Lambda \hat{\Gamma}(m) db^*(\lambda) - \hat{\kappa}_n^u \right\}, \\
\hat{\mathcal{M}}_n &= \left\{ m \in \mathcal{M}_n : |\hat{\beta}(\lambda) - \hat{\Gamma}(m)(\lambda)| \leq \hat{\kappa}_n^m \text{ for all } \lambda \in \Lambda \right\}, \\
\hat{\mathcal{G}}_n(b^*) &= \left\{ g \in \mathbf{M}_n : \left| \int_0^1 \left[\sum_{d=0}^1 g_d(u) v_d(u) \right] du - \int_\Lambda \hat{\Gamma}(v)(\lambda) db^*(\lambda) \right| \leq \hat{\kappa}_n^g \ \forall v \in \mathcal{V}_n \right\},
\end{aligned}$$

where we used the fact that $\mathbf{M}_n = \mathbf{M}_n^*$. Therefore, the bootstrap statistic reduces to

$$\begin{aligned}
T_n^{\text{bs}} = \sup_{b^* \in \hat{\mathcal{D}}_n} \sup_{(g,\tilde{m}) \in \hat{\mathcal{G}}_n(b^*) \times \hat{\mathcal{M}}_n} \inf_{m \in \hat{\mathcal{M}}_n} & \int_\Lambda \{\hat{\mathbb{G}}_\beta(\lambda) - \hat{\mathbb{G}}_\Gamma(m)(\lambda)\} db^*(\lambda) \\
\text{s.t. } & \int_0^1 \sum_{d=0}^1 \left[g_d(u)(m_d(u) - \tilde{m}_d(u)) \right] du \geq 0. \qquad (78)
\end{aligned}$$

We note that, even though b^* is infinite dimensional, it is possible to instead optimize T_n^{bs} over a sieve for $\mathcal{D}$. We provide a formal development of this sieve approach for $\mathcal{D}$ in Appendix E. A natural choice of a sieve for $\mathcal{D}$ would be $\mathcal{D}_n \equiv \{b^*(\mathcal{A}) = \sum_{j=1}^J w_j 1\{\lambda_j \in \mathcal{A}\} : \sum_{j=1}^J |w_j| \leq 1\}$ for $\{\lambda_j\}_{j=1}^J \subseteq \Lambda$. Similarly, using $\mathcal{M}_n$ in place of $\mathcal{M}$ renders all parameters finite dimensional and all constraints bilinear. See Appendix G for additional discussion of computation. ■

5.5 The Test

In this section, we apply the results obtained in Sections 5.3 and 5.4 to construct a consistent test. We then discuss choosing bandwidths for this test.

5.5.1 Basic Properties

Theorem 1 implies that the *unconditional* quantiles of $\mathbb{U}_{P,n}(\kappa_n^u)$ could be used as critical values for the test statistic T_n. While $\mathbb{U}_{P,n}(\kappa_n^u)$ is infeasible, Theorem 2 suggests that the quantiles of T_n^{bs} *conditional* on the data can be used instead. We therefore define the critical value $\hat{c}_{1-\alpha}$ for our test to be given by

$$\hat{c}_{1-\alpha} \equiv \inf \left\{ c : P\left(T_n^{\text{bs}} \leq c | \{Y_i, D_i, Z_i\}_{i=1}^n \right) \geq 1 - \alpha \right\}. \qquad (79)$$

In the following, it will also be useful to define the quantiles of $\mathbb{U}_{P,n}(\kappa_n^u)$ itself, which we denote by

$$c_{1-\alpha}(\mathbb{U}_{P,n}(\kappa_n^u)) \equiv \inf\{c : P(\mathbb{U}_{P,n}(\kappa_n^u) \leq c) \geq 1-\alpha\}. \tag{80}$$

As usual, a final regularity condition is required to establish that a distributional approximation also delivers consistent estimators of the desired quantiles; see, e.g., Romano and Shaikh (2012) and Chernozhukov, Lee, and Rosen (2013). In the present context, this regularity condition is the following.

Assumptions S (continued)

S.12 *There is a $\delta > 0$ and sequence ζ_n with $\zeta_n(\delta_n^c + \delta_n^s) = o(1)$ and*

$$\sup_{\eta\in[0,\delta]} \sup_{0<\epsilon<1} \sup_{P\in\mathbf{P}_0} \frac{1}{\epsilon} P\Big(|\mathbb{U}_{P,n}(\kappa_n^u) - c_{1-\alpha-\eta}(\mathbb{U}_{P,n}(\kappa_n^u))| \leq \epsilon \Big) \leq \zeta_n.$$

Notice that if ζ_n is bounded, then Assumption S.12 corresponds to the requirement that the cumulative distribution functions of $\mathbb{U}_{P,n}(\kappa_n^u)$ are uniformly (in n) continuous in a neighborhood of their $1-\alpha$ quantiles. However, by allowing ζ_n to diverge, Assumption S.12 also allows the distribution of $\mathbb{U}_{P,n}(\kappa_n^u)$ to become increasingly discontinuous. The "rate" of this loss of continuity (ζ_n) must be slower than the rate of convergence of our stochastic approximation ($\delta_n^c + \delta_n^s$).

We now state our final and main result on statistical inference.

Theorem 3. *Suppose that Assumptions S.1–S.7 and S.9–S.11 are satisfied. Then, for any sequence κ_n^u that satisfies Assumption S.8 and S.12,*

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} P(T_n > \hat{c}_{1-\alpha}) \leq \alpha.$$

Moreover, under the same assumptions,

$$\lim_{n\to\infty} P(T_n > \hat{c}_{1-\alpha}) = 1.$$

for any $P \in \mathbf{P} \setminus \mathbf{P}_0$.

Theorem 3 says that our proposed test is consistent and controls size uniformly in $P \in \mathbf{P}_0$. We note that the class $\mathbf{P}_0$ over which size control is ensured is fairly broad. In particular, the class $\mathbf{P}_0$ is such that the set of solutions $\{m \in \mathcal{M} : \|\beta_P - \Gamma_P(m)\|_{\mathbf{B}} = 0\}$ cannot even be consistently estimated uniformly in $\mathbf{P}_0$. We believe that this quality is particularly important when $\mathcal{M}$ is defined by linear inequality constraints and the target parameter is partially identified. Moreover, we emphasize that we have not placed any conditions that prohibit shape restrictions on $\mathcal{M}$, other than requiring $\mathcal{M}$ to be a convex set. As such, we believe Theorem 3 may be of independent interest for applications other than the one studied in this paper. Finally, we note that in establishing Theorem 3, we have not assumed that the data is i.i.d. Instead, we have assumed that there is a bootstrap $(\hat{\mathbb{G}}_\beta, \hat{\mathbb{G}}_\Gamma)$ that is capable of handling the dependence structure in the data. This allows for clustering and other types of dependence.

5.5.2 Bandwidth Guidance

Constructing the bootstrap statistic T_n^{bs} requires specifying the bandwidths $\hat{\kappa}_n^u$, $\hat{\kappa}_n^m$, and $\hat{\kappa}_n^g$. While Assumptions S.8, S.9, and S.10 do not provide much direction in choosing these bandwidths, it is clear that their values should be related to the distribution of the data. In the following, we provide some guidance on the selection of these quantities. In future work, we hope to formalize a data-driven procedure for determining these bandwidths by applying the insights of Romano, Shaikh, and Wolf (2014). However, at present we keep the discussion informal.

The sequence κ_n^u and its feasible counterpart $\hat{\kappa}_n^u$ were introduced in the derivation of the distributional approximation in Theorem 1. The role of κ_n^u is related to

$$\hat{b}^* \in \arg\max_{b^* \in \mathcal{D}} \sqrt{n} \left\{ \langle b^*, \hat{\beta} \rangle - \nu(b^*, \hat{\Gamma}(\mathcal{M}_n)) \right\}, \tag{81}$$

which represents the direction in which $\hat{\beta}$ is farthest away from the set $\hat{\Gamma}(\mathcal{M}_n)$. In particular, while κ_n^u is allowed to converge to zero, it must do so slowly enough for

$$\liminf_{n\to\infty} \inf_{P \in \mathbf{P}_0} P\left(\hat{b}^* \in \mathcal{D}_P(\kappa_n^u)\right) = 1, \tag{82}$$

where $\mathcal{D}_P(\kappa_n^u)$ is the set defined in (49). A bound for the probability in (82) is

$$\liminf_{n\to\infty} \inf_{P \in \mathbf{P}_0} P\left(\hat{b}^* \in \mathcal{D}_P(\kappa_n^u)\right) \geq \liminf_{n\to\infty} \inf_{P \in \mathbf{P}_0} P\left(\sup_{m \in \mathcal{M}_n} \|\mathbb{G}_{P,\beta} - \mathbb{G}_{P,\Gamma}(m)\|_{\mathbf{B}} \leq \sqrt{n}\kappa_n^u \right), \tag{83}$$

where we are assuming that $\delta_n^s = 0$ for simplicity. This suggests taking $\hat{\kappa}_n^u$ to be

$$\hat{\kappa}_n^u = \frac{1}{\sqrt{n}} \times \inf\left\{ c : P\left(\sup_{m \in \mathcal{M}_n} \|\hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m)\|_{\mathbf{B}} \leq c | \{Y_i, D_i, Z_i\}_{i=1}^n \right) \geq 1 - \alpha_n \right\} \tag{84}$$

for some sequence $\alpha_n \downarrow 0$.

The bandwidth $\hat{\kappa}_n^m$ was introduced in Assumption S.9 with the purpose of ensuring that $\hat{\mathcal{M}}_n$ contained the set $\{m \in \mathcal{M} : \|\beta_P - \Gamma_P(m)\|_{\mathbf{B}} = 0\}$ with probability tending to one. In analogy to (83), it is possible to show that

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}_0} P\left(\Pi_n m \in \hat{\mathcal{M}}_n \text{ for all } m \in \mathcal{M} \text{ s.t. } \Gamma_P(m) = \beta_P\right)$$
$$\geq \liminf_{n\to\infty} \inf_{P\in\mathbf{P}_0} P\left(\sup_{m\in\mathcal{M}_n} \|\mathbb{G}_{P,\beta} - \mathbb{G}_{P,\Gamma}(m)\|_{\mathbf{B}} \leq \sqrt{n}\hat{\kappa}_n^m\right). \tag{85}$$

From (83) and (85), we see that the choices of $\hat{\kappa}_n^u$ and $\hat{\kappa}_n^m$ are closely related. For instance, given that $\hat{\kappa}_n^u$ is selected according to (84), a simple rule is to let $\hat{\kappa}_n^m = \hat{\kappa}_n^u + T_n/\sqrt{n}$. Here, the addition of $T_n/\sqrt{n}$ to $\hat{\kappa}_n^u$ ensures that $\hat{\mathcal{M}}_n \neq \emptyset$.

Lastly, recall that $\hat{\kappa}_n^g$ was introduced in the construction of neighborhoods for the adjoint $\Gamma_P^* : \mathbf{B}^* \mapsto \mathbf{M}^*$ in the weak (operator) topology. A natural choice for $\hat{\kappa}_n^g$ is

$$\hat{\kappa}_n^g = \frac{1}{\sqrt{n}} \times \inf\left\{c : P\left(\sup_{b^*\in\hat{\mathcal{D}}_n} \sup_{v\in\mathcal{V}_n} |\langle b^*, \hat{\mathbb{G}}_\Gamma(v)\rangle| \leq c | \{Y_i, D_i, Z_i\}_{i=1}^n\right) \geq 1 - \alpha_n\right\}, \tag{86}$$

where again $\alpha_n \downarrow 0$. Together, (84), $\hat{\kappa}_n^m = \hat{\kappa}_n^u + T_n/\sqrt{n}$, and (86) provide a simple heuristic way to relate bandwidth selection to features of the distribution of the data. In Appendix G.4, we show that implementing these bandwidth choices amounts to solving a large number of small mixed integer linear programs. We leave a more detailed analysis of bandwidth selection to future work.

6 The Efficacy of Price Subsidies for Bed Nets

In this section, we apply our method to analyze how price subsidies affect the adoption and usage of a preventative health product.

6.1 Background, Data, and Experiment

According to the World Health Organization (WTO), approximately 5.9 million children under the age of 5 died in 2015. WTO estimates that a majority of these early childhood deaths could be prevented or treated if households were regularly using existing health products, such as deworming medication, mosquito nets, water treatment solution, or latrines.[18] An important, and largely unanswered, question for developing countries is how to design cost effective policies that promote access to (and usage of) preventive health products.

While highly subsidizing health products has been shown to markedly increase access in developing countries, researchers and policymakers alike have expressed con-

[18] See http://www.who.int/mediacentre/factsheets/fs178/en/.

cerns about the cost effectiveness of such policies (see e.g. Cohen and Dupas (2010) and Dupas and Zwane (2016)). One concern is the financial cost of subsidizing inframarginal consumers who would have still purchased the product under a smaller subsidy. Another concern is that households that are unwilling to pay a monetary price for a product might also be unwilling to pay the non-monetary costs associated with using the product on a regular basis.[19] On the other hand, charging a higher price to screen out non-users may exclude poor or credit-constrained individuals who would benefit from using the product.[20]

The goal of our empirical analysis is to assess these tradeoffs by evaluating the effects of potential subsidy regimes on usage of a preventative health product, taking into account differences in subsidization costs across the regimes. Building on Dupas (2014), we use data from a randomized controlled experiment in Kenya that randomly assigned prices for a new and improved antimalarial bed net called the Olyset net.[21] The experimenters randomized prices across a total of 1,200 households in six villages. Households had a three month opportunity to purchase the Olyset net at their assigned subsidized price. Prices for the net varied from 0 to 250 Kenyan schillings (250 Ksh, or approximately $3.80), which is roughly twice the average daily wage for agricultural work in the area. Seventeen different prices were offered in total, but each area was assigned only four or five of these prices. For example, if an area was assigned the price set (Ksh 50, 100, 150, 200, 250), then all of the study households in the area were randomly assigned to one of these five prices. Price sets for every area included low, medium, and high prices. Only two areas had a price set that included free provision for some households.

Two months after the experiment, Dupas (2014) collected data on household purchase and usage of the Olyset net.[22] Usage was assessed by whether a household stated having started using the net, and whether the net was observed hanging above their bedding at the time of the visit. Table 1 in Dupas (2014) presents summary statistics of household characteristics and their correlation with the randomized price assignment. This table suggests that randomization was successful in making the price assignment

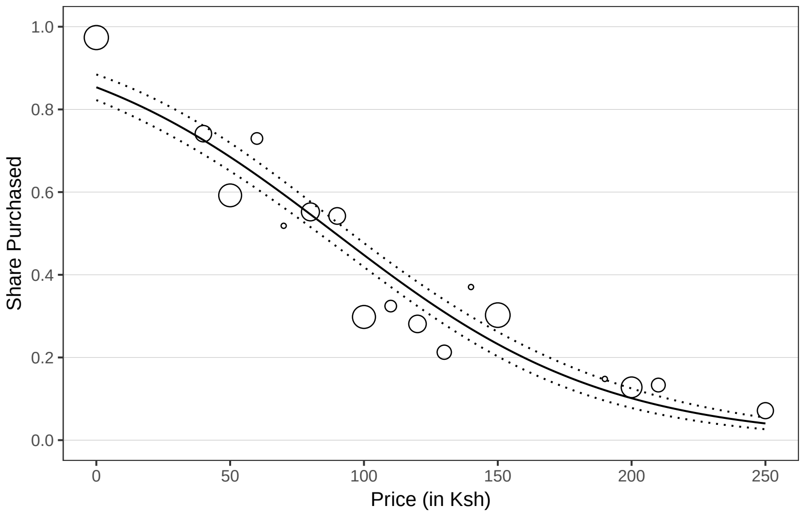
[19] See, for example, Ashraf, Berry, and Shapiro (2010), who study the provision of chlorine solution for water treatment. They find that higher prices tend to screen out those who use the product less.

[20] For example, Dupas and Zwane (2016) study alternative screening mechanisms for the provision of chlorine solution. They find that higher prices may prevent many households that would use the product from obtaining it.

[21] Dupas (2014) uses this data to examine how short-run subsidies affect long-run adoption through learning. We refer to Dupas (2014) for a detailed discussion of the background, data, and experimental design.

[22] To study effects on long-run adoption, Dupas also conducted a one year follow-up. However, the follow-up survey only included a subset of the villages. We therefore focus on the short-run effects.

Figure 9: Impact of Price on the Household's Purchase of Bed Net



Notes: This figure plots purchase rates against prices. The size of the circles reflect the relative size of the sample at each price point. The lines show predicted values from logit regressions of the households' decision to purchase the bed net on the randomly assigned prices. The dashed lines indicate 90% confidence intervals.

independent of observable baseline characteristics.

Figure 9 replicates Dupas' experimental estimates of the impact of price on the household's purchase decision. This figure plots purchase rates for the Olyset net against the price assignment. The sizes of the circles reflect the relative sample sizes at each price point. The lines indicate predicted values (and confidence intervals) from logistic regressions of the households' purchase decisions on their randomly assigned prices. The demand function is quite steep. The likelihood of purchasing the bed net increases from .04 to over .23 as the price decreases from Ksh 250 to 150, reaching nearly .70 at Ksh 50.

6.2 Evaluating a Class of Subsidy Regimes

We use the randomly assigned prices as instruments, and use their exogenous variation, together with our method, to study the effectiveness of different subsidy regimes.[23] As discussed in Section 3.1, we can use our method to compute bounds on PTREs that measure the causal effect on usage of one subsidy regime, a_0, relative to another subsidy

[23] A possible threat to the exclusion restriction (Assumption I.2) is the psychological "sunk cost" effect, whereby individuals who have paid more for a product feel more compelled to use it. However, recent experiments conducted in developing countries find no evidence of a sunk cost effect in settings with health products (Cohen and Dupas (2010) and Ashraf et al. (2010)).

regime, a_1. For example, consider the effect of a policy that offers free provision of the Olyset net to all households, compared to a policy under which households have the option to buy the net at a given price. This comparison does not directly correspond to the variation in prices induced by Dupas' experiment, and is therefore not point identified under standard instrumental variables assumptions. Nevertheless, our method can be used to establish bounds, which we now show can be quite informative.

In the notation of Section 2.1, Z is the randomly assigned price, D is an indicator for whether the household purchases the Olyset net, and Y is an indicator for whether the household uses the net. We consider PRTEs that contrast a regime a_0 under which the propensity score is constant at p^{a_0}, to a regime a_1 with a constant propensity score of p^{a_1}. In Table 3, we focus on two particular choices of p^{a_0} and p^{a_1}. The first choice is $p^{a_0} = 0$ and $p^{a_1} = 1$, which can be thought of as the contrast between a regime a_1 with free provision of the Olyset set, and another regime a_0 under which there is no access to the Olyset net. In this case, the PRTE we consider is just the average treatment effect (ATE). For the second choice, the a_0 regime is still free provision, but in the a_1 regime the Olyset net is offered at a price of Ksh 150, which is roughly the observed market price a year after the experiment (Dupas, 2014). To implement this PRTE, we use the estimated propensity score to predict p^{a_0} and p^{a_1}. As discussed in Section 3.2, the PRTE in this case can be interpreted as the LATE from a hypothetical experiment in which households are either freely provided an Olyset net or able to purchase one at 150 Ksh.

Table 3 demonstrates the way in which our method allows the researcher to transparently substitute the strength of their assumptions with the strength of their conclusions. Comparing the bounds across columns clarifies how the strength of the conclusions (the width of the bounds) depends on two aspects. First, for a fixed set of assumptions (indexed here by the Bernstein polynomial order, K), bounds based on a broader class of IV-like estimates are substantially more informative. In the first five columns, we restrict attention to the information contained in the IV estimand that uses the propensity score $p(Z)$ as an instrument for D. By comparison, the next five columns demonstrate that both the OLS and IV estimands carry independent information. The last five columns demonstrate that the bounds can be tightened considerably by using richer specifications of the multi-valued price instrument.

The other aspect that affects the strength of the conclusions is the set of maintained assumptions on the MTR functions, $m = (m_0, m_1)$. In columns (1)-(4), (6)-(9) and (11)-(14), we model m_0 and m_1 with Bernstein polynomials of order K for various

Table 3: The Effects of Purchase on Net Usage

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
	Information Specification														
Intercept	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Linear in $p(Z)$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
OLS						✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$1(Z \leq 50)$											✓	✓	✓	✓	✓
$1(Z \leq 150)$											✓	✓	✓	✓	✓
Panel A.	**Population Average Treatment Effect**														
K (polynomial order)	2	6	10	20	NP	2	6	10	20	NP	2	6	10	20	NP
Bounds															
Lower	.6521	.4646	.3857	.3275	.2533	.6521	.4956	.4700	.4537	.3954	$\emptyset$	.6365	.5602	.5269	.4487
Upper	.6772	.7269	.7362	.7445	.7515	.6521	.7269	.7362	.7445	.7515	$\emptyset$	.7104	.7178	.7229	.7253
90% Confidence Interval															
Lower	.5486	.3761	.2995	.2421		.4282	.4032	.3511	.3204		.5206	.4130	.3652	.3260	
Upper	.7462	.8019	.8102	.8139		.7516	.8093	.8179	.8209		.7491	.7910	.7941	.7978	
Panel B.	**PRTE at Free Provision versus a Price of 150 Ksh**														
K (polynomial order)	2	6	10	20	NP	2	6	10	20	NP	2	6	10	20	NP
Bounds															
Lower	.6600	.5881	.5626	.5444	.4817	.6600	.5881	.5626	.5444	.4856	$\emptyset$	.6758	.6506	.6214	.5573
Upper	.7049	.8140	.8469	.8817	.9732	.6600	.7085	.7172	.7275	.7941	$\emptyset$	.6895	.6988	.7140	.7492
90% Confidence Interval															
Lower	.5417	.5005	.4695	.4479		.3890	.3472	.3414	.3320		.5079	.4755	.4584	.4281	
Upper	.7686	.9161	.9519	.9746		.7732	.9263	.9616	.9838		.7713	.9093	.9291	.9511	
	Specifications of the IV-like Estimands														
Intercept	$s(d,z) = 1$					$s(d,z) = 1$					$s(d,z) = 1$				
Linear in $p(Z)$	$s(d,z) = p(z)$					$s(d,z) = p(z)$					$s(d,z) = p(z)$				
OLS						$s(d,z) = d$					$s(d,z) = d$				
$1(Z \leq 50)$											$s(d,z) = 1(z \leq 50)$				
$1(Z \leq 150)$											$s(d,z) = 1(z \leq 150)$				

Notes: This table reports bounds and 90% confidence intervals for the effects of purchase on usage of the Olyset net. We estimate the propensity score, p, using the fitted logistic regression from Figure 9. K denotes the order of the Bernstein polynomial specification for the MTR functions. The confidence intervals are based on 200 bootstrap replicates, and the tuning parameters are specified as 0.05.

choices of K.[24] In columns (5), (10), and (15), we use the Haar basis (constant spline) specification discussed in Section 2.7, which was shown in Proposition 4 to provide exact nonparametric bounds in the sample. As evident from Table 3, the stronger the restrictions, the tighter the bounds.

At one extreme, it is possible to specify K to achieve point identification. As shown in Brinch et al. (2015), the information used in column (6)-(15) are sufficient to point identify the MTE (and hence any PRTE), provided that $K = 2$, so that the MTR functions are quadratic. For example, the results in column (6) imply a 65% usage rate among individuals induced to purchase by a change to free provision from a price of Ksh 150. However, the bounds for this specification are empty in column (11), suggesting that $K = 2$ might be too restrictive in our setting.

At the other extreme, one can impose few or no restrictions on m_0 and m_1, which yields wider, but more robust bounds. However, even with a very flexible specification, the bounds remain quite informative. For example, with a 10th order Bernstein polynomial ($K = 10$), the results in column (13) show that the ATE can be bounded between .56 and .72, whereas the PRTE of free provision compared to a price of Ksh 150 is bounded between .65 and .70. These results imply that the usage rate in the overall population is relatively similar to that among individuals induced to purchase by a change to free provision from a price of Ksh 150.

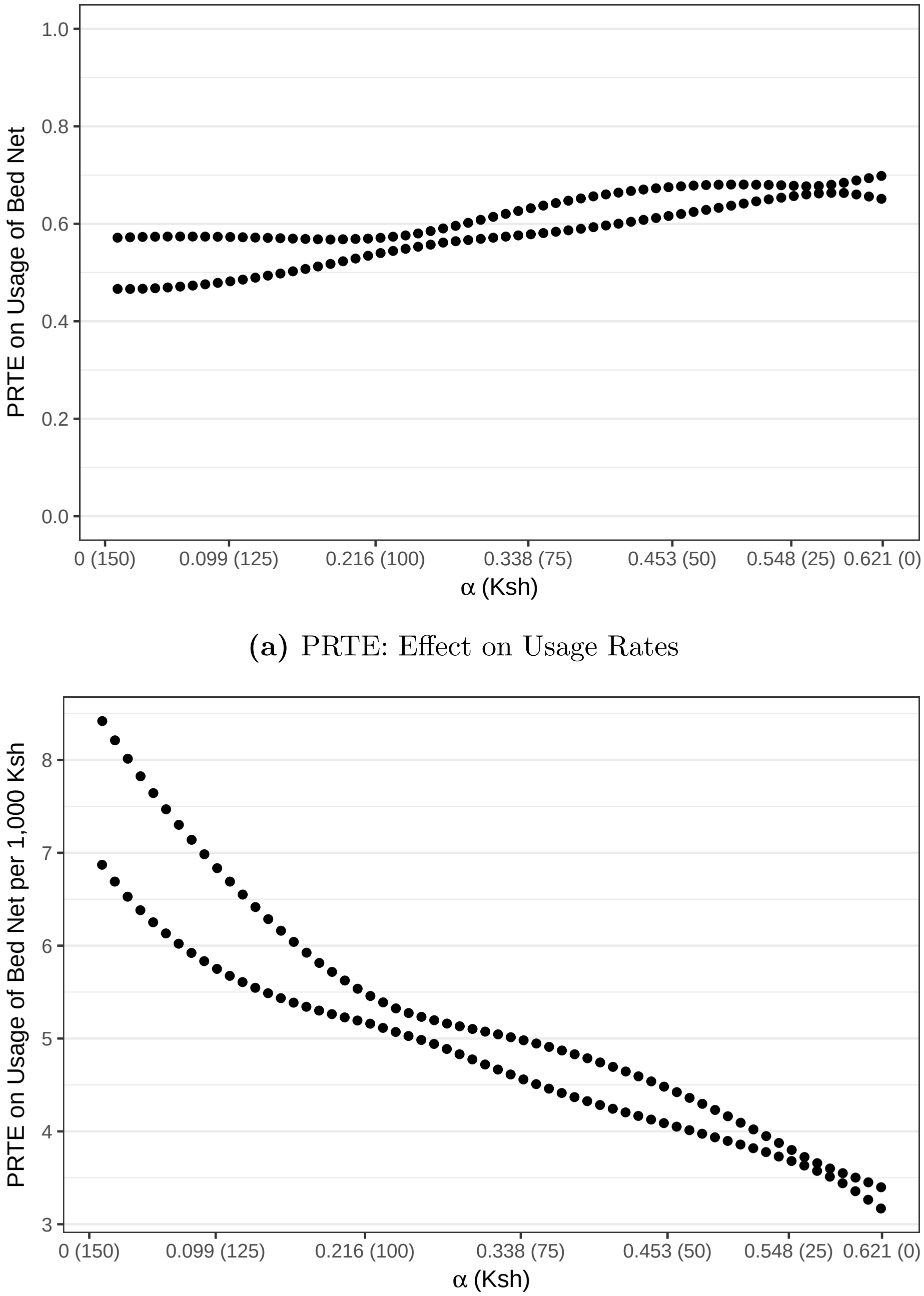
Figure 10 reports bounds on a range of PRTEs. Each PRTE takes p^{a_0} to be the propensity score associated with a regime in which all households can purchase the bed net at a uniform price of Ksh 150.[25] The alternative policy is specified by $p^{a_1} = p^{a_0} + \alpha$. Panel A plots bounds on this PRTE as a function of α. The predicted price levels associated with a given α (predicted using the logistic regression in Figure 9) are indicated in parentheses. To be conservative, we set $K = 10$ as in column (13) of Table 3, which allows a very flexible functional form for m_0 and m_1.

If households that are unwilling to pay larger monetary prices for the Olyset net are less likely to use the net after purchasing it, then we would expect to see the PRTEs declining in α. In contrast, Panel A of Figure 10 suggests that the PRTEs are *increasing* in α. This finding is consistent with higher prices excluding poor or credit constrained individuals who would use an Olyset net if they were able to purchase one. For example, among those induced to purchase the net by lowering prices from Ksh 150 to Ksh 80, the usage rate is bounded between .57 and .62. By comparison, among the larger set of individuals induced to purchase by a change from Ksh 150 to free provision, the usage rate is bounded between .65 and .70.

[24] See Appendix F for a definition and discussion of the Bernstein polynomials.

[25] That is, $p^{a_0} = \hat{p}(150)$, where $\hat{p}$ is the estimated logistic model plotted in Figure 9.

Figure 10: Bounds on Policy Relevant Treatment Effects



(a) PRTE: Effect on Usage Rates

(b) PRTE: Effects Relative to Costs

Notes: Panel A displays bounds on usage for a range of PRTEs. Each PRTE takes $p^{a_0} = \hat{p}(150)$, where $\hat{p}$ is the fitted logistic regression from Figure 9. This corresponds to a policy regime under which all households can purchase the Olyset bed net at a uniform price of Ksh 150. The alternative policy regime is $p^{a_1} = p^{a_0} + \alpha$. The x-axis shows how the PRTE varies with changes in α or prices in Ksh (z), where α and z are related through $\alpha = \hat{p}(z) - p^{a_0}$. Panel B divides the PRTE on usage by the PRTE on subsidy costs. The specification of the IV-like estimands and Bernstein polynomial order correspond to column (13) in Panel B of Table 3.

Even if a subsidy regime with low prices leads to relatively high usage among those who purchase the Olyset net, it still comes at the cost of subsidizing inframarginal consumers who would have also purchased the net at a higher price. To compare the effects on usage to the costs of lowering prices, we divide the PRTE on usage by the PRTE on subsidy costs. Panel B of Figure 10 plots the results, which show that subsidy regimes with low prices or free provision induce relatively few individuals to use the Olyset net per Ksh spent. For example, if prices are lowered from Ksh 150 to 80, then each 1,000 Ksh in subsidy costs induces at least 4.70 households (lower bound) to use the bed net. By comparison, moving from a regime with a price of Ksh 150 to one with free provision induces at most 3.39 households (upper bound) to use the bed net for every 1,000 Ksh in subsidy costs.[26]

7 Conclusion

We proposed a method for using instrumental variables (IVs) to draw inference about treatment parameters other than the LATE. Our method uses the observation that both the IV estimand and many treatment parameters can be expressed as weighted averages of the same underlying marginal treatment effects. Since the weights are known or identified, this observation implies that knowledge of the IV estimand places some restrictions on the unknown marginal treatment response functions, and hence on the possible values of other treatment parameters of interest. We showed how to extract this information from a class of objects described as IV-like estimands, which includes the TSLS and OLS estimands, among others. An important aspect of our method is that it allows the researcher a large degree of flexibility in choosing a parameter of interest, and in choosing auxiliary identifying assumptions that can be used to help tighten their empirical conclusions. Another important feature is that it is computationally straightforward.

We considered three main applications of our method. We showed that it can be used for counterfactual policy analysis, including extrapolation away from the variation in treatment induced by the instrument at hand. In addition, we showed that the general framework facilitates tests of both model specification and economic behavior. To implement our method, we developed a novel inference procedure that exploits the convexity of our setting, while remaining uniformly valid and computationally reliable under weak conditions. We applied our method to analyze how price subsidies affect

[26] Of course, high subsidization or free provision of the Olyset net may still be optimal depending on its effects on health outcomes. Unfortunately, we do not have the data needed to assess the effects on health outcomes.

the adoption and usage of antimalarial bed nets in Kenya. Our results suggest that generous subsidy regimes encourage usage among individuals who would otherwise not use a bed net, albeit at a potentially high cost of subsidizing other inframarginal individuals.

The overall message from our analysis is that it is possible, both in theory and in practice, to use IVs to draw informative inference about a wide range of treatment parameters other than the LATE. This enables researchers to learn about causal effects for a broad range of individuals, not just those who are affected by the instrument observed in the data. The ability to do this is critical to ensuring that estimates obtained through IV strategies are both externally valid and relevant to policy.

A Proofs for Section 2

Proof of Proposition 1. Using equation (1), we first note that

$$\beta_s = E[s(D,Z)DY_1] + E[s(D,Z)(1-D)Y_0]. \tag{87}$$

Using equation (2) with Assumptions I.1 and I.2, observe that the first term of (87) can be written as

$$\begin{aligned} E[s(D,Z)DY_1] &= E[s(D,Z)\mathbb{1}[U \leq p(Z)]E[Y_1|U,Z]] \\ &\equiv E[s(1,Z)\mathbb{1}[U \leq p(Z)]m_1(U,X)], \end{aligned} \tag{88}$$

where the first equality follows because $s(D,Z)D$ is a deterministic function of (U,Z), and the second equality uses the definition of m_1 and I.2, together with the identity that

$$s(D,Z)\mathbb{1}[U \leq p(Z)] \equiv s\left(\mathbb{1}[U \leq p(Z)], Z\right)\mathbb{1}[U \leq p(Z)] = s(1,Z)\mathbb{1}[U \leq p(Z)].$$

Using the normalization that $U|Z$ is uniformly distributed on $[0,1]$ for any realization of Z, it follows from (88) that

$$\begin{aligned} E[s(D,Z)DY_1] &= E\left[E\left[s(1,Z)\mathbb{1}[U \leq p(Z)]m_1(U,Z)|Z\right]\right] \\ &= E\left[\int_0^1 s(1,Z)\mathbb{1}[u \leq p(Z)]m_1(u,Z)\,du\right] \\ &\equiv E\left[\int_0^1 \omega_{1s}(u,Z)m_1(u,Z)\,du\right]. \end{aligned}$$

The claimed result follows after applying a symmetric argument to the second term on the right hand side of equation (87). *Q.E.D.*

Proof of Proposition 2. Since $\Gamma_s : \mathcal{M} \mapsto \mathbf{R}$ is linear for every $s \in \mathcal{S}$, it follows from convexity of $\mathcal{M}$ that $\mathcal{M}_{\mathcal{S}}$ is convex as well. (Note that the empty set is trivially convex.) If $\mathcal{M}_{\mathcal{S}}$ is empty, then by definition we also have $\mathcal{B}^\star_{\mathcal{S}} = \emptyset$. On the other hand, if $\mathcal{M}_{\mathcal{S}} \neq \emptyset$, then the linearity of $\Gamma^\star : \mathcal{M} \mapsto \mathbf{R}$ implies that $\mathcal{B}^\star_{\mathcal{S}} \equiv \Gamma^\star(\mathcal{M}_{\mathcal{S}}) \subseteq \mathbf{R}$ is a convex set, and so its closure is $[\inf_{m\in\mathcal{M}_{\mathcal{S}}} \Gamma^\star(m), \sup_{m\in\mathcal{M}_{\mathcal{S}}} \Gamma^\star(m)] \equiv [\underline{\beta}^\star, \overline{\beta}^\star]$. *Q.E.D.*

Proof of Proposition 3. For notational simplicity, we define the set

$$\mathcal{M}_{\text{id}} \equiv \{m \in \mathcal{M} : m \text{ satisfies (17) and (18) almost surely}\}.$$

For any $m \equiv (m_0, m_1) \in \mathcal{M}_{\text{id}}$ and $s \in \mathcal{S}$ we obtain from the definition of β_s that

$$\beta_s = E[s(D,Z)E[Y|D,Z]] = \sum_{d\in\{0,1\}} E[\mathbb{1}[D=d]s(d,Z)E[Y|D=d,Z]]. \tag{89}$$

Examining the $d = 0$ term in the summation, we obtain

$$\begin{aligned} E[\mathbb{1}[D=0]s(0,Z)E[Y|D=0,Z]] &= E\left[\mathbb{1}[D=0]s(0,Z)\frac{1}{1-p(Z)}\int_{p(Z)}^1 m_0(u,X)du\right] \\ &= E\left[\mathbb{1}[D=0]\frac{1}{1-p(Z)}\int_0^1 m_0(u,X)\omega_{0s}(u,Z)du\right] \\ &= E\left[\int_0^1 m_0(u,X)\omega_{0s}(u,Z)du\right], \end{aligned} \tag{90}$$

where the first equality follows from $m \in \mathcal{M}_{\text{id}}$ satisfying (17), the second equality uses the definition $\omega_{0s}(u,z) = s(0,z)1\{u > p(z)\}$, and the final equality is implied by $P(D=0|Z) = 1 - p(Z)$. By analogous arguments, we also obtain

$$E[\mathbb{1}[D=1]s(1,Z)E[Y|D=1,Z]] = E\left[\int_0^1 m_1(u,X)\omega_{1s}(u,Z)du\right]. \tag{91}$$

Together, (89), (90), and (91) imply that $\Gamma_s(m) = \beta_s$. In particular, since $s \in \mathcal{S}$ and $m \in \mathcal{M}_{\text{id}}$ were arbitrary, we conclude $\mathcal{M}_{\text{id}} \subseteq \mathcal{M}_S$ as claimed.

Next, suppose $\mathcal{S} = \{s(d,z) = \mathbb{1}[d = d']f(z) : (d', f) \in \{0,1\} \times \mathcal{F}\}$ and that the closed linear span of $\mathcal{F}$ is equal to $L^2(Z)$. Then note that for any $m \in \mathcal{M}_{\mathcal{S}}$ and $s \in \mathcal{S}$ with the structure $s(d,z) = \mathbb{1}[d=0]f(z)$ we obtain by definition of β_s and Γ_s that

$$E[Y\mathbb{1}[D=0]f(Z)] \equiv \beta_s = \Gamma_s(m) = E\left[\int_{p(Z)}^1 m_0(u,X)du f(Z)\right] \tag{92}$$

where the second equality follows from $m \in \mathcal{M}_{\mathcal{S}}$ and the final equality is due to $\omega_{0s}(u,z) \equiv 1\{u > p(z)\}s(0,z)$ and $s(0,z) = \mathbb{1}[0=0]f(z)$. Furthermore, define

$$\Delta(Z) \equiv E[Y\mathbb{1}[D=0]|Z] - \int_{p(Z)}^1 m_0(u,X)du \tag{93}$$

and note that (92) implies that $E[\Delta(Z)f(Z)] = 0$ for all $f \in \mathcal{F}$. Since $E[Y_0^2] < \infty$ by Assumption I.2 and $E[\int m_d^2(u,X)du] < \infty$, Jensen's inequality implies that $\Delta \in L^2(Z)$. Thus, since $E[\Delta(Z)f(Z)] = 0$ for all $f \in \mathcal{F}$ and the closed linear span of $\mathcal{F}$ equals $L^2(Z)$, we conclude that $\Delta(Z) = 0$ almost surely. Equivalently, since $P(D = 0|Z) =$

$1 - p(Z)$ by definition of $p(Z)$, we obtain whenever $1 - p(Z) > 0$ that

$$E[Y|D = 0, Z] = \frac{1}{1 - p(Z)} \int_{p(Z)}^{1} m_0(u, X) du \tag{94}$$

almost surely, i.e. m_0 satisfies (17). Analogous arguments imply that m_1 satisfies (18). Since $(m_0, m_1) = m \in \mathcal{M}_{\mathcal{S}}$ was arbitrary, we conclude that $\mathcal{M}_{\mathcal{S}} \subseteq \mathcal{M}_{\text{id}}$, which together with (19) establishes (20). *Q.E.D.*

Proof of Proposition 4. We prove the proposition for the upper bound of the target parameter. The proof for the lower bound follows by identical arguments.

Observe that since $\mathcal{M}_{\text{fd}} \subseteq \mathcal{M}$, we can immediately conclude that $\overline{\beta}^\star_{\text{fd}} \leq \overline{\beta}^\star$. As for the opposite inequality, let $\{z_1, \ldots, z_K\}$ denote the discrete support of Z. Then notice that for any $(m_0, m_1) = m \in \mathcal{M}$ we have that

$$\begin{aligned}
\Gamma_{ds}(m_d) &= E\left[\sum_{j=1}^{J}\sum_{k=1}^{K} \mathbb{1}[U \in \mathcal{A}_j, Z = z_k] m_d(U, X)\omega_{ds}(U, Z)\right] \\
&= E\left[\sum_{j=1}^{J}\sum_{k=1}^{K} E[m_d(U, X)|U \in \mathcal{A}_j, Z = z_k]\omega_{ds}(U, Z)\mathbb{1}[U \in \mathcal{A}_j, Z = z_k]\right] \\
&= E\left[\sum_{j=1}^{J}\sum_{l=1}^{L} E[m_d(U, X)|U \in \mathcal{A}_j, X = x_l]\omega_{ds}(U, Z)\mathbb{1}[U \in \mathcal{A}_j, X = x_l]\right].
\end{aligned} \tag{95}$$

The first equality here follows from $\{\mathcal{A}_j\}_{j=1}^{J}$ being a partition of $[0, 1]$. The second equality follows because $\omega_{ds}(u, z)$ is constant on each set $\{u \in \mathcal{A}_j, z = z_k\}$ given the assumption that $\mathbb{1}[u \leq p(z)]$ is constant $\mathcal{A}_j$—recall Proposition 1. The third equality in (95) follows from X being a subvector of $Z = (Z_0, X)$.

Given the definition of $b_{jl}(u, x)$ in (26), we can conclude from (95) that $\Gamma_{ds}(m_d) = \Gamma_{ds}(\Pi m_d)$ for Πm_d as defined in (27). An identical argument also yields $\Gamma^\star_d(m_d) = \Gamma^\star_d(\Pi m_d)$. Hence, given the assumption that $\Pi m \equiv (\Pi m_0, \Pi m_1) \in \mathcal{M}_{\text{fd}}$, we have

$$\begin{aligned}
\overline{\beta}^\star &\equiv \sup_{m \in \mathcal{M}} \{\Gamma^\star(m) \text{ s.t. } \Gamma_s(m) = \beta_s \text{ for all } s \in \mathcal{S}\} \\
&= \sup_{m \in \mathcal{M}} \{\Gamma^\star(\Pi m) \text{ s.t. } \Gamma_s(\Pi m) = \beta_s \text{ for all } s \in \mathcal{S}\} \\
&\leq \sup_{m \in \mathcal{M}_{\text{fd}}} \{\Gamma^\star(m) \text{ s.t. } \Gamma_s(m) = \beta_s \text{ for all } s \in \mathcal{S}\} \equiv \overline{\beta}^\star_{\text{fd}}.
\end{aligned} \tag{96}$$

We conclude that $\overline{\beta}^\star_{\text{fd}} = \overline{\beta}^\star$. *Q.E.D.*

B MTR Weights for Linear IV Estimands

In this appendix, we show that linear IV estimands are a special case of our notion of an IV–like estimand. For the purpose of this discussion, we adopt some of the standard textbook terminology regarding "endogenous variables" and "included" and "excluded" instruments in the context of linear IV models without heterogeneity. Consider a linear IV specification with endogenous variables $\widetilde{X}_1$, included instruments $\widetilde{Z}_1$, and excluded instruments $\widetilde{Z}_2$. We let $\widetilde{X} \equiv [\widetilde{X}_1, \widetilde{Z}_1]'$ and $\widetilde{Z} \equiv [\widetilde{Z}_2, \widetilde{Z}_1]'$. We assume that both $E[\widetilde{Z}\widetilde{Z}']$ and $E[\widetilde{Z}\widetilde{X}']$ have full rank.

As long as these two conditions hold, all of the variables in $\widetilde{X}$ and $\widetilde{Z}$ can be functions of (D, Z). Usually, one would expect that $\widetilde{X}_1$ would include D and possibly some interactions between D and other covariates X. The instruments, $\widetilde{Z}$, would usually consist of functions of the vector Z, which contains X, by notational convention. The included portion of $\widetilde{Z}$, i.e. $\widetilde{Z}_1$, would typically also include a constant term as one of its components. However, whether $\widetilde{Z}$ is actually "exogenous" in the usual sense of the linear instrumental variables model is not relevant to our definition of an IV–like estimand or the derivation of the weighting expression (12). In particular, OLS is nested as a linear IV specification through the case in which $\widetilde{Z}_1 = [1, D]'$ and both $\widetilde{X}_1$ and $\widetilde{Z}_2$ are empty vectors.

It may be the case that $\widetilde{Z}$ has dimension larger than $\widetilde{X}$, as in "overidentified" linear models. In such cases, a positive definite weighting matrix Π is used to generate instruments $\Pi\widetilde{Z}$ that have the same dimension as $\widetilde{X}$. A common choice of Π is the two-stage least squares weighting $\Pi_{\text{TSLS}} \equiv E[\widetilde{X}\widetilde{Z}']E[\widetilde{Z}\widetilde{Z}']^{-1}$ which has as its rows the first stage coefficients corresponding to linear regressions of each component of $\widetilde{X}$ on the entire vector $\widetilde{Z}$. We assume that Π is a known or identified non-stochastic matrix with full rank. This covers Π_{TSLS} and the optimal weighting under heteroskedasticity (optimal GMM) as particular cases given standard regularity conditions. The instrumental variables estimator that uses $\Pi\widetilde{Z}$ as an instrument for $\widetilde{X}$ in a regression of Y on $\widetilde{X}$ has corresponding estimand

$$\beta_{\text{IV},\Pi} \equiv \left(\Pi E[\widetilde{Z}\widetilde{X}']\right)^{-1} \left(\Pi E[\widetilde{Z}Y]\right) = E\left[\left(\Pi E[\widetilde{Z}\widetilde{X}']\right)^{-1} \Pi\widetilde{Z}Y\right],$$

which is an IV–like estimand with $s(D, Z) \equiv (\Pi E[\widetilde{Z}\widetilde{X}'])^{-1}\Pi\widetilde{Z}$.

C Proofs for Section 5

Proof of Theorem 1. First, we show in Lemma D.1 that

$$T_n = \sup_{b^* \in \mathcal{D}} \sqrt{n} \left\{ \langle b^*, \hat{\beta} \rangle - \nu(b^*, \hat{\Gamma}(\mathcal{M})) \right\} + O_p(\delta_n^s + \delta_n^c), \tag{97}$$

uniformly over $P \in \mathbf{P}$. Next, recalling that $\mathcal{D}$ is the unit sphere in $\mathbf{B}^*$ and $\mathcal{D}_P(\kappa_n^u)$ is defined as in (49), we define the event

$$\Omega_n(P) \equiv \left[\sup_{b^* \in \mathcal{D} \setminus \mathcal{D}_P(\kappa_n^u)} \left\{ \langle b^*, \hat{\beta} \rangle - \nu(b^*, \hat{\Gamma}(\mathcal{M})) \right\} \leq \sup_{b^* \in \mathcal{D}_P(\kappa_n^u)} \left\{ \langle b^*, \hat{\beta} \rangle - \nu(b^*, \hat{\Gamma}(\mathcal{M})) \right\} \right]. \tag{98}$$

For any $b^* \in \mathbf{B}^*$, let

$$\Delta_n(b^*) \equiv \sqrt{n} \left| \langle b^*, \hat{\beta} - \beta_P \rangle - \left\{ \nu(b^*, \hat{\Gamma}(\mathcal{M})) - \nu(b^*, \Gamma_P(\mathcal{M})) \right\} \right|.$$

Observe that, since $\langle b^*, \beta_P \rangle - \nu(b^*, \Gamma_P(\mathcal{M})) \leq -\kappa_n^u$ for all $b^* \in \mathcal{D} \setminus \mathcal{D}_P(\kappa_n^u)$, we have

$$\begin{aligned} P\left(\Omega_n(P)^c\right) &\leq P\left(\sup_{b^* \in \mathcal{D}} 2\Delta_n(b^*) - \sqrt{n}\kappa_n^u > \sup_{b^* \in \mathcal{D}_P(\kappa_n^u)} \sqrt{n} \left\{ \langle b^*, \beta_P \rangle - \nu(b^*, \Gamma_P(\mathcal{M})) \right\} \right) \\ &\leq P\left(\sup_{b^* \in \mathcal{D}} 2\Delta_n(b^*) > \sqrt{n}\kappa_n^u \right) \end{aligned} \tag{99}$$

where $\Omega_n(P)^c$ denotes the complement of $\Omega_n(P)$ and in the final inequality we used $0 \in \mathcal{D}_P(\kappa_n^u)$. Hence, since $\sqrt{n}\kappa_n^u \uparrow \infty$, (99) and Lemma D.2 yield

$$\limsup_{n \to \infty} \sup_{P \in \mathbf{P}} P(\Omega_n(P)^c) = 0. \tag{100}$$

In addition, note that the definition of $\nu(b^*, \hat{\Gamma}(\mathcal{M}))$ and Assumption S.3 imply

$$\begin{aligned} &\sup_{b^* \in \mathcal{D}_P(\kappa_n^u)} \sqrt{n} \left\{ \langle b^*, \hat{\beta} \rangle - \nu(b^*, \hat{\Gamma}(\mathcal{M})) \right\} \\ &= \sup_{b^* \in \mathcal{D}_P(\kappa_n^u)} \inf_{m \in \mathcal{M}} \left\{ \langle b^*, \sqrt{n}\{\hat{\beta} - \beta_P\} - \sqrt{n}\{\hat{\Gamma}(m) - \Gamma_P(m)\} \rangle + \sqrt{n} \langle b^*, \beta_P - \Gamma_P(m) \rangle \right\} \\ &= \sup_{b^* \in \mathcal{D}_P(\kappa_n^u)} \inf_{m \in \mathcal{M}} \left\{ \langle b^*, \mathbb{G}_{P,\beta} - \mathbb{G}_{P,\Gamma}(m) \rangle + \sqrt{n} \langle b^*, \beta_P - \Gamma_P(m) \rangle \right\} + O_p(\delta_n^c) \end{aligned} \tag{101}$$

uniformly in $P \in \mathbf{P}$. The first claim of Theorem 1 now follows from (97), (100), and (101) together with $\delta_n^c \downarrow 0$ and $\delta_n^s \downarrow 0$ from Assumptions S.3 and S.5.

To establish the second claim, we note that if $P \in \mathbf{P}_0$, then the set $\{m \in \mathcal{M} :$

$\sqrt{n}\langle b^*, \beta_P - \Gamma_P(m)\rangle \le \delta_n^s\}$ cannot be empty for any $b^* \in \mathbf{B}^*$, because for such P there exists an $m_P \in \mathcal{M}$ such that $\beta_P = \Gamma_P(m_P)$. Hence, we conclude from (97), (100), and (101) that

$$\begin{aligned} T_n &= \sup_{b^* \in \mathcal{D}_P(\kappa_n^u)} \inf_{m \in \mathcal{M}} \left\{ \langle b^*, \mathbb{G}_{P,\beta} - \mathbb{G}_{P,\Gamma}(m)\rangle + \sqrt{n}\langle b^*, \beta_P - \Gamma_P(m)\rangle \right\} + O_p(\delta_n^c + \delta_n^c) \\ &\le \mathbb{U}_{P,n}(\kappa_n^u) + O_p(\delta_n^c + \delta_n^s) + \delta_n^s \qquad (102) \end{aligned}$$

uniformly in $P \in \mathbf{P}_0$, where the inequality follows by set inclusion and the definition of $\mathbb{U}_{P,n}(\kappa_n^u)$. This implies the second claim of Theorem 1. *Q.E.D.*

Proof of Lemma 5.1. Note that $\hat{\mathcal{D}}_n \subseteq \mathcal{D}$, $\hat{\mathcal{M}}_n \subseteq \mathcal{M}$, and $|\langle b^*, b\rangle| \le \|b^*\|_{\mathbf{B}^*}\|b\|_{\mathbf{B}} \le \|b\|_{\mathbf{B}}$ for any $b^* \in \mathcal{D}$. These observations with Assumption S.6 imply

$$\begin{aligned} &\sup_{b^* \in \hat{\mathcal{D}}_n} \sup_{m \in \hat{\mathcal{M}}_n} \left| \langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m)\rangle - \langle b^*, \mathbb{G}_{P,\beta}^{\mathrm{bs}} - \mathbb{G}_{P,\Gamma}^{\mathrm{bs}}(m)\rangle \right| \\ &\qquad \le \|\hat{\mathbb{G}}_\beta - \mathbb{G}_{P,\beta}^{\mathrm{bs}}\|_{\mathbf{B}} + \sup_{b^* \in \mathcal{D}} \sup_{m \in \mathcal{M}} \left| \langle b^*, \hat{\mathbb{G}}_\Gamma(m) - \mathbb{G}_{P,\Gamma}^{\mathrm{bs}}(m)\rangle \right| = O_p(\delta_n^c) \qquad (103) \end{aligned}$$

uniformly in $P \in \mathbf{P}$. Next, define the event $\Omega_n(P) \equiv \Omega_{1n}(P) \cap \Omega_{2n}(P)$ where

$$\begin{aligned} &\Omega_{1n}(P) \equiv \left[\mathcal{D}_P(\kappa_n) \subseteq \hat{\mathcal{D}}_n \right] \\ \text{and} \quad &\Omega_{2n}(P) \equiv \left[\Pi_n m_P \in \hat{\mathcal{M}}_n \text{ for all } m_P \in \mathcal{M} \text{ s.t. } \Gamma_P(m_P) = \beta_P \right]. \end{aligned}$$

Observe that Lemmas D.4 and D.5 imply

$$\limsup_{n\to\infty} \sup_{P \in \mathbf{P}_0} P(\Omega_n(P)^c) \le \limsup_{n\to\infty} \left\{ \sup_{P \in \mathbf{P}} P(\Omega_{1n}(P)^c) + \sup_{P \in \mathbf{P}_0} P(\Omega_{2n}(P)^c) \right\} = 0. \quad (104)$$

Moreover, $\{m \in \mathcal{M} : \Gamma_P(m) = \beta_P\} \ne \emptyset$ for every $P \in \mathbf{P}_0$. It follows that, for any $P \in \mathbf{P}_0$, if $\Omega_n(P)$ occurs, then it is also true that $\hat{\mathcal{M}}_n \ne \emptyset$. In this event, Lemma D.3 implies that for any $b^* \in \mathbf{B}^*$ there exists an $m \in \hat{\mathcal{M}}_n$ such that $\langle b^*, \Gamma_P(m)\rangle = \nu(b^*, \Gamma_P(\hat{\mathcal{M}}_n))$, so that $I_n(\Gamma_P)$ is indeed well defined. Thus, (103) and (104) yield

$$\begin{aligned} I_n(\Gamma_P) = \sup_{b^* \in \hat{\mathcal{D}}_n} \inf_{m \in \hat{\mathcal{M}}_n} &\Big\{ \langle b^*, \mathbb{G}_{P,\beta}^{\mathrm{bs}} - \mathbb{G}_{P,\Gamma}^{\mathrm{bs}}(m)\rangle \\ &\text{s.t. } \langle b^*, \Gamma_P(m)\rangle = \nu(b^*, \Gamma_P(\hat{\mathcal{M}}_n)) \Big\} + O_p(\delta_n^c) \qquad (105) \end{aligned}$$

uniformly in $P \in \mathbf{P}_0$.

Next, note that, for any $b^* \in \mathcal{D}$, if $\hat{m} \in \hat{\mathcal{M}}_n$ satisfies $\langle b^*, \Gamma_P(\hat{m})\rangle = \nu(b^*, \Gamma_P(\hat{\mathcal{M}}_n))$, then it also satisfies $\langle b^*, \Gamma_P(\hat{m} - m)\rangle \ge 0$ for all $m \in \hat{\mathcal{M}}_n$. Hence, when $\Omega_{2n}(P)$ is true, we obtain

$$\begin{aligned}\sqrt{n}\langle b^*, \beta_P - \Gamma_P(\hat{m})\rangle &\leq \sqrt{n}\langle b^*, \Gamma_P(m_P) - \Gamma_P(\Pi_n m_P)\rangle \\ &\leq \sqrt{n}\|\Gamma_P(m_P - \Pi_n m_P)\|_{\mathbf{B}} \leq \delta_n^s \end{aligned} \tag{106}$$

for any $b^* \in \mathcal{D}$, any $\hat{m} \in \mathcal{M}_n$ satisfying $\langle b^*, \Gamma_P(\hat{m})\rangle = \nu(b^*, \Gamma_P(\hat{\mathcal{M}}_n))$, and every $m_P \in \mathcal{M}$ such that $\Gamma_P(m_P) = \beta_P$. The second equality in (106) used $\langle b^*, b\rangle \leq \|b\|_{\mathbf{B}}$ for all $b^* \in \mathcal{D}$, while the third inequality followed from Assumption S.5. Since under $\Omega_{1n}(P)$ we have $\mathcal{D}_P(\kappa_n) \subseteq \hat{\mathcal{D}}_n$, (106) and $\Omega_n(P) \equiv \Omega_{1n}(P) \cap \Omega_{2n}(P)$ imply that whenever $\Omega_n(P)$ occurs

$$\mathbb{U}_{P,n}^{\text{bs}}(\kappa_n^u) \leq \sup_{b^* \in \hat{\mathcal{D}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \left\{\langle b^*, \mathbb{G}_{P,\beta}^{\text{bs}} - \mathbb{G}_{P,\Gamma}^{\text{bs}}(m)\rangle \text{ s.t. } \langle b^*, \Gamma_P(m)\rangle = \nu(b^*, \Gamma_P(\hat{\mathcal{M}}_n))\right\}. \tag{107}$$

The result now follows from (107) together with (104) and (105). *Q.E.D.*

Proof of Lemma 5.2. Since $\{m \in \mathcal{M} : \beta_P = \Gamma_P(m)\} \neq \emptyset$ for all $P \in \mathbf{P}_0$, Lemma D.5 implies that

$$\liminf_{n\to\infty} \inf_{P \in \mathbf{P}_0} P(\hat{\mathcal{M}}_n \neq \emptyset) = 1. \tag{108}$$

Given (108), we establish the claim by showing that (67) holds whenever $\hat{\mathcal{M}}_n \neq \emptyset$.

To this end, note that if $\hat{\mathcal{M}}_n \neq \emptyset$, then by definition of the support function, an $m \in \hat{\mathcal{M}}_n$ satisfies $\langle b^*, \Gamma_P(m)\rangle = \nu(b^*, \Gamma_P(\hat{\mathcal{M}}_n))$ if and only if it satisfies $\langle b^*, \Gamma_P(m - \tilde{m})\rangle \geq 0$ for all $\tilde{m} \in \hat{\mathcal{M}}_n$. Equivalently, since $\langle b^*, \Gamma_P(m)\rangle = \langle \Gamma_P^*(b^*), m\rangle$ for all $(b^*, m) \in \mathbf{B}^* \times \mathbf{M}$, we obtain by set inclusion that

$$\begin{aligned} I_n(\Gamma_P) &= \sup_{b^* \in \hat{\mathcal{D}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \left\{\langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m)\rangle \text{ s.t. } \langle \Gamma^*(b^*), m - \tilde{m}\rangle \geq 0 \text{ for all } \tilde{m} \in \hat{\mathcal{M}}_n\right\} \\ &\geq \sup_{b^* \in \hat{\mathcal{D}}_n} \sup_{\tilde{m} \in \hat{\mathcal{M}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \left\{\langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m)\rangle \text{ s.t. } \langle \Gamma^*(b^*), m - \tilde{m}\rangle \geq 0\right\}. \end{aligned} \tag{109}$$

For any $(b^*, \tilde{m}) \in \hat{\mathcal{D}}_n \times \hat{\mathcal{M}}_n$, define

$$C(b^*, \tilde{m}) \equiv \{m \in \hat{\mathcal{M}}_n : \langle \Gamma_P^*(b^*), m - \tilde{m}\rangle \geq 0\}.$$

Recall that $\Gamma_P^*(b^*) \in \mathbf{M}^*$ and that $\hat{\mathcal{M}}_n$ is compact in the weak topology by Lemma D.3. It follows that, provided $\hat{\mathcal{M}}_n \neq \emptyset$, there exists an $m_{b^*}^o \in \hat{\mathcal{M}}_n$ such that $\langle \Gamma_P^*(b^*), m_{b^*}^o\rangle = \sup_{m \in \hat{\mathcal{M}}_n} \langle \Gamma^*(b^*), m\rangle$. Such an $m_{b^*}^o$ satisfies $C(b^*, m_{b^*}^o) \subseteq C(b^*, \tilde{m})$ for all $\tilde{m} \in \hat{\mathcal{M}}_n$.

Using set inclusion and $\langle \Gamma^*_P(b^*), m^o_{b^*} - \tilde{m}\rangle \geq 0$ for all $\tilde{m} \in \hat{\mathcal{M}}_n$, we obtain

$$\begin{aligned}
&\sup_{b^* \in \hat{\mathcal{D}}_n} \sup_{\tilde{m} \in \hat{\mathcal{M}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \{\langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m)\rangle \text{ s.t. } \langle \Gamma^*(b^*), m - \tilde{m}\rangle \geq 0\} \\
&\quad = \sup_{b^* \in \hat{\mathcal{D}}_n} \inf_{m \in C(b^*, m^o_{b^*})} \langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m)\rangle \qquad (110)\\
&\quad = \sup_{b^* \in \hat{\mathcal{D}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \{\langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m)\rangle \text{ s.t. } \langle \Gamma^*_P(b^*), m - \tilde{m}\rangle \geq 0 \text{ for all } \tilde{m} \in \hat{\mathcal{M}}_n\}.
\end{aligned}$$

The claim now follows from (109) and (110). *Q.E.D.*

Proof of Theorem 2. For any $b^* \in \mathbf{B}^*$, define the set

$$\tilde{\mathcal{G}}_n(b^*) \equiv \{g \in \mathbf{M}^* : |\langle g - \hat{\Gamma}^*(b^*), v\rangle| \leq \hat{\kappa}^q_n \text{ for all } v \in \mathcal{V}_n\}.$$

and define

$$\tilde{T}^{\text{bs}}_n \equiv \sup_{b^* \in \hat{\mathcal{D}}_n} \sup_{(\tilde{g}, \tilde{m}) \in \tilde{\mathcal{G}}_n(b^*) \times \hat{\mathcal{M}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \{\langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m)\rangle \text{ s.t. } \langle \tilde{g}, m - \tilde{m}\rangle \geq 0\}. \qquad (111)$$

Define the event $\Omega_n(P) \equiv \Omega_{1n}(P) \cap \Omega_{2n}(P)$, where

$$\begin{aligned}
\Omega_{1n}(P) &\equiv \left[\Gamma^*_P(b^*) \in \tilde{\mathcal{G}}_n(b^*) \text{ for all } b^* \in \hat{\mathcal{D}}_n\right] \qquad (112)\\
\Omega_{2n}(P) &\equiv \left[T^{\text{bs}}_n = \tilde{T}^{\text{bs}}_n\right]. \qquad (113)
\end{aligned}$$

We show that $\Omega_n(P)$ occurs with probability approaching 1, uniformly over $P \in \mathbf{P}_0$.

Using the definition of $\tilde{\mathcal{G}}_n(b^*)$ and Assumption S.10 we have

$$\begin{aligned}
&\liminf_{n \to \infty} \inf_{P \in \mathbf{P}} P(\Omega_{1n}(P)) \\
&\quad \geq \liminf_{n \to \infty} \inf_{P \in \mathbf{P}} P\left(\sup_{b^* \in \hat{\mathcal{D}}_n} \sup_{v \in \mathcal{V}_n} |\langle \sqrt{n}\{\Gamma^*_P(b^*) - \hat{\Gamma}^*(b^*)\}, v\rangle| \leq \sqrt{n}\hat{\kappa}^g_n\right) \\
&\quad \geq \liminf_{C \uparrow \infty} \liminf_{n \to \infty} \inf_{P \in \mathbf{P}} P\left(\sup_{b^* \in \hat{\mathcal{D}}_n} \sup_{v \in \mathcal{V}_n} |\langle \sqrt{n}\{\Gamma^*_P(b^*) - \hat{\Gamma}^*(b^*)\}, v\rangle| \leq C\right). \qquad (114)
\end{aligned}$$

Recall that under Assumption S.11, $\mathcal{V}_n \subseteq \lambda(\mathcal{M}_n - m_c)$ for some $m_c \in \mathcal{M}_n$, while $\mathcal{M}_n \subseteq \mathcal{M}$ by Assumption S.5, and $\hat{\mathcal{D}}_n \subseteq \mathcal{D}$ by construction. Using these inclusions, we have

$$\begin{aligned}
&\sup_{b^*\in\hat{\mathcal{D}}_n}\sup_{v\in\mathcal{V}_n}\left|\langle\sqrt{n}\{\Gamma_P^*(b^*)-\hat{\Gamma}^*(b^*)\},v\rangle\right| \\
&\quad\le \sup_{b^*\in\mathcal{D}}\sup_{m_1,m_2\in\mathcal{M}}\left|\langle\sqrt{n}\{\Gamma_P^*(b^*)-\hat{\Gamma}^*(b^*)\},\lambda(m_1-m_2)\rangle\right| \\
&\quad\le \sup_{b^*\in\mathcal{D}}\sup_{m\in\mathcal{M}} 2|\lambda|\left|\langle b^*,\sqrt{n}\{\hat{\Gamma}-\Gamma_P\}(m)\rangle\right| \\
&\quad= \sup_{b^*\in\mathcal{D}}\sup_{m\in\mathcal{M}} 2|\lambda|\,|\langle b^*,\mathbb{G}_{P,\Gamma}(m)\rangle| + O_p(\delta_n^c) \qquad (115)
\end{aligned}$$

where the second inequality follows from $\hat{\Gamma}^*$ and Γ_P^* being the adjoints of $\hat{\Gamma}$ and Γ_P respectively, while the equality follows from Assumption S.3, and holds uniformly over $P\in\mathbf{P}$. Since $|\langle b^*,b\rangle|\le\|b\|_{\mathbf{B}}$ for all $b^*\in\mathcal{D}$ and $b\in\mathbf{B}$, (115) implies that

$$\begin{aligned}
&\liminf_{C\uparrow\infty}\liminf_{n\to\infty}\inf_{P\in\mathbf{P}} P\left(\sup_{b^*\in\hat{\mathcal{D}}_n}\sup_{v\in\mathcal{V}_n}|\langle\sqrt{n}\{\Gamma_P^*(b^*)-\hat{\Gamma}^*(b^*)\},v\rangle|\le C\right) \\
&\quad\ge \liminf_{C\uparrow\infty}\inf_{P\in\mathbf{P}} P\left(\sup_{m\in\mathcal{M}} 2|\lambda|\|\mathbb{G}_{P,\Gamma}(m)\|_{\mathbf{B}}\le\frac{C}{2}\right)=1 \qquad (116)
\end{aligned}$$

where we have applied Markov's inequality with Assumption S.4. From (114), (116), and Lemma D.6 we conclude that

$$\limsup_{n\to\infty}\sup_{P\in\mathbf{P}_0} P(\Omega_n(P)^c)\le\limsup_{n\to\infty}\left\{\sup_{P\in\mathbf{P}} P(\Omega_{1n}(P)^c)+\sup_{P\in\mathbf{P}_0} P(\Omega_{2n}(P)^c)\right\}=0. \qquad (117)$$

Next, we recall that Lemma 5.1 establishes that for any κ_n^u satisfying Assumption S.8 and $\sqrt{n}\kappa_n^u\uparrow\infty$, there exists a $\tilde{\xi}_n^{\text{bs}}\in\mathbf{R}$ satisfying $\tilde{\xi}_n^{\text{bs}}=O_p(\delta_n^c)$ uniformly in $P\in\mathbf{P}_0$ such that

$$\mathbb{U}_{P,n}^{\text{bs}}(\kappa_n^u)\le I_n(\Gamma_P)+\tilde{\xi}_n^{\text{bs}} \qquad (118)$$

for all $P\in\mathbf{P}_0$. From Lemma 5.2 we can conclude that $I_n(\Gamma_P)\le\tilde{T}_n^{\text{bs}}$ whenever the event $\Omega_{1n}(P)$ occurs, whereas by definition $T_n^{\text{bs}}=\tilde{T}_n^{\text{bs}}$ whenever the event $\Omega_{2n}(P)$ occurs. Therefore, the claim of the result follow from (117), (118), and setting $\xi_n^{\text{bs}}=\tilde{\xi}_n^{\text{bs}}+1\{\{Y_i,Z_i\}_{i=1}^n\in\Omega_n(P)^c\}(I_n(\Gamma_P)-T_n^{\text{bs}})$. *Q.E.D.*

Proof of Theorem 3. The proof of the first claim closely follows the arguments in Lemma D.6 of Chernozhukov et al. (2015).

First, note that Theorems 1 and 2 imply that

$$T_n\le\mathbb{U}_{P,n}(\kappa_n^u)+O_p(\delta_n^c+\delta_n^s)\quad\text{and}\quad\mathbb{U}_{P,n}^{\text{bs}}(\kappa_n^u)\le T_n^{\text{bs}}+O_p(\delta_n^c), \qquad (119)$$

both uniformly over $P \in \mathbf{P}$. Together with Markov's inequality, this implies that for any $\epsilon > 0$ and $C_n \uparrow \infty$ we have

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} P\Big(P\left(\mathbb{U}^{\text{bs}}_{P,n}(\kappa^u_n) > T^{\text{bs}}_n + C_n\delta^c_n|\{Y_i, D_i, Z_i\}^n_{i=1}\right) > \epsilon\Big)$$
$$\leq \limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} \frac{1}{\epsilon}P\Big(\mathbb{U}^{\text{bs}}_{P,n}(\kappa^u_n) > T^{\text{bs}}_n + C_n\delta^c_n\Big) = 0. \tag{120}$$

In particular, this implies that there is a sequence $\eta_n \downarrow 0$ (which depends on $C_n \uparrow \infty$) such that the event

$$\Omega_n(P) \equiv \Big[\{Y_i, D_i, Z_i\}^n_{i=1} : P\Big(\mathbb{U}^{\text{bs}}_{P,n}(\kappa^u_n) > T^{\text{bs}}_n + C_n\delta^c_n|\{Y_i, D_i, Z_i\}^n_{i=1}\Big) \leq \eta_n\Big] \tag{121}$$

satisfies $\sup_{P\in\mathbf{P}_0} P(\Omega_n(P)^c) = o(1)$. Furthermore, for any $t \in \mathbf{R}$,

$$P\left(T^{\text{bs}}_n \leq t|\{Y_i, D_i, Z_i\}^n_{i=1}\right) 1\left[\{Y_i, D_i, Z_i\}^n_{i=1} \in \Omega_n(P)\right]$$
$$\leq P\left(T^{\text{bs}}_n \leq t \text{ and } \mathbb{U}^{\text{bs}}_{P,n}(\kappa^u_n) \leq T^{\text{bs}}_n + C_n\delta^c_n|\{Y_i, D_i, Z_i\}^n_{i=1}\right) + \eta_n$$
$$\leq P\left(\mathbb{U}^{\text{bs}}_{P,n}(\kappa^u_n) \leq t + C_n\delta^c_n\right) + \eta_n, \tag{122}$$

where the final inequality used the independence of $\mathbb{U}^{\text{bs}}_{P,n}(\kappa^u_n)$ and $\{Y_i, D_i, Z_i\}^n_{i=1}$ under Assumption S.6.

By evaluating (122) at $t = \hat{c}_{1-\alpha}$, we obtain

$$P\Big(\hat{c}_{1-\alpha} + C_n\delta^c_n \geq c_{1-\alpha-\eta_n}(\mathbb{U}_{P,n}(\kappa^u_n))\Big) \geq P\big(\{Y_i, D_i, Z_i\}^n_{i=1} \in \Omega_n(P)\big) \tag{123}$$

where used the equality in distribution of $\mathbb{U}_{P,n}(\kappa^u_n)$ and $\mathbb{U}^{\text{bs}}_{P,n}(\kappa^u_n)$ under Assumption S.6. Theorem 1, (123), and $\sup_{P\in\mathbf{P}_0} P(\Omega_n(P)^c) = o(1)$ then yield

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} P(T_n > \hat{c}_{1-\alpha})$$
$$\leq \limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} P\big(\mathbb{U}_{P,n}(\kappa^u_n) + C_n(\delta^s_n + \delta^c_n) > \hat{c}_{1-\alpha}\big)$$
$$\leq \limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} P\big(\mathbb{U}_{P,n}(\kappa^u_n) > c_{1-\alpha-\eta_n}(\mathbb{U}_{P,n}(\kappa^u_n)) - 2C_n(\delta^c_n + \delta^s_n)\big). \tag{124}$$

Notice, however, that because ζ_n (defined in Assumption S.12) satisfies $\zeta_n(\delta^s_n + \delta^c_n) = o(1)$, we can select $C_n \uparrow \infty$ slowly enough so that $\zeta_n C_n(\delta^s_n + \delta^c_n) = o(1)$. Since $\eta_n \downarrow 0$, we conclude that for such a choice of C_n,

$$\begin{aligned}&\limsup_{n\to\infty}\sup_{P\in\mathbf{P}_0} P\Big(c_{1-\alpha-\eta_n}(\mathbb{U}_{P,n}(\kappa_n^u)) \geq \mathbb{U}_{P,n}(\kappa_n^u) > c_{1-\alpha-\eta_n}(\mathbb{U}_{P,n}(\kappa_n^u)) - 2C_n(\delta_n^c+\delta_n^s)\Big)\\ &\quad\leq \limsup_{n\to\infty}\zeta_n \times 2C_n(\delta_n^c+\delta_n^s) = 0. \end{aligned} \tag{125}$$

Combining (124) and (125) establishes the first claim, since

$$\limsup_{n\to\infty}\sup_{P\in\mathbf{P}_0} P(T_n > \hat{c}_{1-\alpha}) \leq \limsup_{n\to\infty}\sup_{P\in\mathbf{P}_0} P\Big(\mathbb{U}_{P,n}(\kappa_n^u) > c_{1-\alpha-\eta_n}(\mathbb{U}_{P,n}(\kappa_n^u))\Big) \leq \alpha. \tag{126}$$

To establish the second claim, we first note that for any $b^* \in \mathbf{B}^*$, and any sequence m_n that converges (in the weak topology) to an $m \in \mathbf{M}$, we have

$$\lim_{n\to\infty}\langle b^*, \Gamma_P(m_n) - \Gamma_P(m)\rangle = \lim_{n\to\infty}\langle \Gamma_P^*(b^*), m_n - m\rangle = 0, \tag{127}$$

which implies $\Gamma_P : \mathbf{M} \mapsto \mathbf{B}$ is continuous when both $\mathbf{M}$ and $\mathbf{B}$ are equipped with their respective weak topologies. Also, note that $m \mapsto \|\beta_P - \Gamma_P(m)\|_{\mathbf{B}}$ is lower semicontinuous (with respect to the weak topologies), which follows because $\Gamma_P : \mathbf{M} \mapsto \mathbf{B}$ is continuous, and because the norm functional $\|\cdot\|_{\mathbf{B}}$ is lower semicontinuous with respect to the weak topology in $\mathbf{B}$ (see Lemma 6.22 in Aliprantis and Border (2006)). Since $\mathcal{M}$ is compact in the weak topology by Assumption S.1, we conclude that

$$\Delta_0 \equiv \inf_{m\in\mathcal{M}}\|\beta_P - \Gamma_P(m)\|_{\mathbf{B}} = \min_{m\in\mathcal{M}}\|\beta_P - \Gamma_P(m)\|_{\mathbf{B}} > 0 \tag{128}$$

for any $P \in \mathbf{P}\setminus\mathbf{P}_0$.

Next, observe that Lemmas D.1 and D.2 imply that for any $P \in \mathbf{P}\setminus\mathbf{P}_0$

$$\begin{aligned} T_n &= \sup_{b^*\in\mathcal{D}}\sqrt{n}\Big\{\langle b^*, \hat{\beta}\rangle - \nu(b^*, \hat{\Gamma}(\mathcal{M}))\Big\} + o_p(1)\\ &= \sup_{b^*\in\mathcal{D}}\sqrt{n}\left\{\langle b^*, \beta_P\rangle - \nu(b^*, \Gamma_P(\mathcal{M}))\right\} + O_p(1)\\ &= \inf_{m\in\mathcal{M}}\sqrt{n}\|\beta_P - \Gamma_P(m)\|_{\mathbf{B}} + O_p(1) = \sqrt{n}\Delta_0 + O_p(1), \end{aligned} \tag{129}$$

where the third equality holds by Theorem 5.13.1 in Luenberger (1969), and the final equality in (129) follows from (128). On the other hand, for any $C > 0$ we have

$$P(\hat{c}_{1-\alpha} > C) \leq P\Big(P\left(T_n^{\text{bs}} > C|\{Y_i, D_i, Z_i\}_{i=1}^n\right) > \alpha\Big) \leq \frac{1}{\alpha}P(T_n^{\text{bs}} > C), \tag{130}$$

where the first inequality follows by definition of $\hat{c}_{1-\alpha}$ and the second by Markov's inequality. From $\hat{\mathcal{D}} \subseteq \mathcal{D}$, $\hat{\mathcal{M}}_n \subseteq \mathcal{M}$, and $\sup_{b^* \in \mathcal{D}} \langle b^*, b \rangle = \|b\|_{\mathbf{B}}$ for any $b \in \mathbf{B}$ (see Lemma 6.10 in Aliprantis and Border (2006)), and Assumption S.6, we obtain

$$\begin{aligned} T_n^{\text{bs}} &\leq \|\hat{\mathbb{G}}_\beta\|_{\mathbf{B}} + \sup_{b^* \in \mathcal{D}} \sup_{m \in \mathcal{M}} \langle b^*, -\hat{\mathbb{G}}_\Gamma(m) \rangle \\ &= \|\mathbb{G}_{P,\beta}^{\text{bs}}\|_{\mathbf{B}} + \sup_{b^* \in \mathcal{D}} \sup_{m \in \mathcal{M}} \langle b^*, -\mathbb{G}_{P,\Gamma}^{\text{bs}}(m) \rangle + o_p(1) \\ &= \|\mathbb{G}_{P,\beta}^{\text{bs}}\|_{\mathbf{B}} + \sup_{m \in \mathcal{M}} \|\mathbb{G}_{P,\Gamma}^{\text{bs}}(m)\|_{\mathbf{B}} + o_p(1). \end{aligned} \tag{131}$$

Assumptions S.4 and S.6 together with (131) then imply that $T_n^{\text{bs}} = O_p(1)$ under any $P \in \mathbf{P}$. We conclude from (130) that

$$\limsup_{C \uparrow \infty} \limsup_{n \to \infty} P(\hat{c}_{1-\alpha} > C) = 0. \tag{132}$$

The second claim now follows from (129) and (132). *Q.E.D.*

D Proofs of Auxiliary Results

Lemma D.1. *If Assumptions S.1, S.2, S.3, and S.5 hold, then*

$$\limsup_{C \uparrow \infty} \limsup_{n \to \infty} \sup_{P \in \mathbf{P}} P\left(\left| T_n - \sup_{b^* \in \mathcal{D}} \sqrt{n} \left\{ \langle b^*, \hat{\beta} \rangle - \nu(b^*, \hat{\Gamma}(\mathcal{M})) \right\} \right| > C(\delta_n^c + \delta_n^s) \right) = 0.$$

Proof of Lemma D.1. Assumption S.5 implies that

$$\begin{aligned} \inf_{m \in \mathcal{M}_n} \sqrt{n} \|\hat{\beta} - \hat{\Gamma}(m)\|_{\mathbf{B}} &\leq \inf_{m \in \mathcal{M}} \sqrt{n} \|\hat{\beta} - \hat{\Gamma}(\Pi_n m)\|_{\mathbf{B}} \\ &\leq \inf_{m \in \mathcal{M}} \sqrt{n} \|\hat{\beta} - \hat{\Gamma}(m)\|_{\mathbf{B}} + \sup_{m \in \mathcal{M}} \sqrt{n} \|\hat{\Gamma}(m) - \hat{\Gamma}(\Pi_n m)\|_{\mathbf{B}}, \end{aligned} \tag{133}$$

where the first inequality follows from $\Pi_n m \in \mathcal{M}_n$ for all $m \in \mathcal{M}$, and the second inequality applies the triangle inequality. Note that

$$\|b\|_{\mathbf{B}} = \sup_{b^* \in \mathcal{D}} \langle b^*, b \rangle \tag{134}$$

for any $b \in \mathbf{B}$; see, e.g., Lemma 6.10 in Aliprantis and Border (2006). Assumption S.3

with (134) imply

$$\begin{aligned}
&\sup_{m\in\mathcal{M}} \sqrt{n}\|\hat{\Gamma}(m-\Pi_n m)-\Gamma_P(m-\Pi_n m)\|_{\mathbf{B}} \\
&\qquad = \sup_{m\in\mathcal{M}}\sup_{b^*\in\mathcal{D}}\langle b^*,\sqrt{n}\{\hat{\Gamma}-\Gamma_P\}(m-\Pi_n m)\rangle \\
&\qquad = \sup_{m\in\mathcal{M}}\sup_{b^*\in\mathcal{D}}\langle b^*,\mathbb{G}_{P,\Gamma}(m)-\mathbb{G}_{P,\Gamma}(\Pi_n m)\rangle + O_p(\delta_n^c) \qquad (135)
\end{aligned}$$

uniformly in $P\in\mathbf{P}$. Similarly, Assumption S.5 with (134) imply

$$\begin{aligned}
&\sup_{P\in\mathbf{P}} E\left[\sup_{m\in\mathcal{M}}\sup_{b^*\in\mathcal{D}}\langle b^*,\mathbb{G}_{P,\Gamma}(m)-\mathbb{G}_{P,\Gamma}(\Pi_n m)\rangle\right] \\
&\qquad = \sup_{P\in\mathbf{P}} E\left[\sup_{m\in\mathcal{M}}\|\mathbb{G}_{P,\Gamma}(m)-\mathbb{G}_{P,\Gamma}(\Pi_n(m))\|_{\mathbf{B}}\right] = O(\delta_n^s). \qquad (136)
\end{aligned}$$

Combining (133), (135), and (136) with Assumption S.5 and applying Markov's inequality, we obtain

$$\begin{aligned}
\inf_{m\in\mathcal{M}_n}\sqrt{n}\|\hat{\beta}-\hat{\Gamma}(m)\|_{\mathbf{B}} &\le \inf_{m\in\mathcal{M}}\sqrt{n}\|\hat{\beta}-\hat{\Gamma}(m)\|_{\mathbf{B}} + O_p(\delta_n^s+\delta_n^c) \\
&\le \inf_{m\in\mathcal{M}_n}\sqrt{n}\|\hat{\beta}-\hat{\Gamma}(m)\|_{\mathbf{B}} + O_p(\delta_n^s+\delta_n^c) \qquad (137)
\end{aligned}$$

uniformly in $P\in\mathbf{P}$, where the second inequality in (137) used $\mathcal{M}_n\subseteq\mathcal{M}$.

Since $\hat{\Gamma}:\mathbf{M}\mapsto\mathbf{B}$ is continuous under Assumption S.2, we can define its adjoint $\hat{\Gamma}^*:\mathbf{B}^*\mapsto\mathbf{M}^*$, so that $\langle b^*,\hat{\Gamma}(m)\rangle=\langle\hat{\Gamma}^*(b^*),m\rangle$. Note that $\hat{\Gamma}^*(b^*)\in\mathbf{M}^*$, and therefore $\langle b^*,\hat{\Gamma}(m)\rangle=\langle\hat{\Gamma}^*(b^*),m\rangle$ implies $m\mapsto\langle b^*,\hat{\Gamma}(m)\rangle$ is continuous in the weak topology. Then, since $\mathcal{M}$ is compact in the weak topology under Assumption S.1, we obtain

$$\begin{aligned}
T_n &= \inf_{m\in\mathcal{M}}\sup_{b^*\in\mathcal{D}}\sqrt{n}\langle b^*,\hat{\beta}-\hat{\Gamma}(m)\rangle + O_p(\delta_n^c+\delta_n^s) \\
&= \sup_{b^*\in\mathcal{D}}\inf_{m\in\mathcal{M}}\sqrt{n}\langle b^*,\hat{\beta}-\hat{\Gamma}(m)\rangle + O_p(\delta_n^c+\delta_n^s) \\
&= \sup_{b^*\in\mathcal{D}}\sqrt{n}\left\{\langle b^*,\hat{\beta}\rangle-\nu(b^*,\hat{\Gamma}(\mathcal{M}))\right\} + O_p(\delta_n^c+\delta_n^s), \qquad (138)
\end{aligned}$$

uniformly in $P\in\mathbf{P}$, where the first equality follows from results (134) and (137), the second equality results from Theorem 4.2 in Sion (1958), and the final equality is implied by the definition of $\nu(b^*,\hat{\Gamma}(\mathcal{M}))$. The claim now follows from (138). *Q.E.D.*

Lemma D.2. *If Assumptions S.3 and S.4 hold, then*

$$\lim_{C\uparrow\infty} \limsup_{n\to\infty} \sup_{P\in\mathbf{P}} \\ P\left(\sup_{b^*\in\mathcal{D}} \sup_{\mathcal{C}\subseteq\mathcal{M}} \left|\langle b^*, \hat{\beta}-\beta_P\rangle - \left\{\nu(b^*, \hat{\Gamma}(\mathcal{C})) - \nu(b^*, \Gamma_P(\mathcal{C}))\right\}\right| > \frac{C}{\sqrt{n}}\right) = 0.$$

Proof of Lemma D.2. Throughout the proof, we use (134) from Lemma D.1, see e.g. Lemma 6.10 in Aliprantis and Border (2006). Together with Assumption S.3, this implies that

$$\sup_{b^*\in\mathcal{D}} \sqrt{n}|\langle b^*, \hat{\beta}-\beta_P\rangle| = \|\sqrt{n}\{\hat{\beta}-\beta_P\}\|_{\mathbf{B}} = \|\mathbb{G}_{P,\beta}\|_{\mathbf{B}} + o_p(1) \tag{139}$$

uniformly in $P \in \mathbf{P}$. Similarly,

$$\begin{aligned} &\sup_{b^*\in\mathcal{D}} \sup_{\mathcal{C}\subseteq\mathcal{M}} \sqrt{n}|\nu(b^*, \hat{\Gamma}(\mathcal{C})) - \nu(b^*, \Gamma_P(\mathcal{C}))| \\ &\quad = \sup_{b^*\in\mathcal{D}} \sup_{\mathcal{C}\subseteq\mathcal{M}} \sqrt{n}\left|\sup_{m\in\mathcal{C}}\langle b^*, \hat{\Gamma}(m)\rangle - \sup_{m\in\mathcal{C}}\langle b^*, \Gamma_P(m)\rangle\right| \\ &\quad \leq \sup_{b^*\in\mathcal{D}} \sup_{m\in\mathcal{M}} |\langle b^*, \sqrt{n}\{\hat{\Gamma}-\Gamma_P\}(m)\rangle| = \sup_{m\in\mathcal{M}} \|\mathbb{G}_{P,\Gamma}(m)\|_{\mathbf{B}} + o_p(1), \end{aligned} \tag{140}$$

uniformly in $P \in \mathbf{P}$, where the inequality follows from $\mathcal{C} \subseteq \mathcal{M}$, and the final equality follows from Assumption S.3. From (139) and (140), we conclude that for any $C > 0$,

$$\begin{aligned} &\limsup_{n\to\infty} \sup_{P\in\mathbf{P}} P\left(\sup_{b^*\in\mathcal{D}} \sup_{\mathcal{C}\subseteq\mathcal{M}} \left|\langle b^*, \hat{\beta}-\beta_P\rangle - \{\nu(b^*, \hat{\Gamma}(\mathcal{C})) - \nu(b^*, \Gamma_P(\mathcal{C}))\}\right| > \frac{C}{\sqrt{n}}\right) \\ &\quad \leq \sup_{P\in\mathbf{P}} P\left(\|\mathbb{G}_{P,\beta}\|_{\mathbf{B}} + \sup_{m\in\mathcal{M}} \|\mathbb{G}_{P,\Gamma}(m)\|_{\mathbf{B}} > \frac{C}{2}\right) \end{aligned}$$

so that the result follows from Markov's inequality with Assumption S.4. *Q.E.D.*

Lemma D.3. *If Assumptions S.1, S.2, and S.7 hold, then $\hat{\mathcal{M}}_n$ is compact in the weak topology. Moreover, if $\hat{\mathcal{M}}_n \neq \emptyset$, then for all $b^* \in \mathbf{B}^*$*

$$\left\{m \in \hat{\mathcal{M}}_n : \langle b^*, \Gamma_P(m)\rangle = \nu(b^*, \Gamma_P(\hat{\mathcal{M}}_n))\right\} \neq \emptyset.$$

Proof of Lemma D.3. We note that $\{m \in \mathbf{M} : \|\hat{\beta} - \hat{\Gamma}(m)\|_{\mathbf{B}} \leq \hat{\kappa}_n^m\}$ is closed in the weak topology of $\mathbf{M}$. This follows by Lemma 6.22 in Aliprantis and Border (2006), which implies that $\{b \in \mathbf{B} : \|\hat{\beta} - b\|_{\mathbf{B}} \leq \hat{\kappa}_n^m\}$ is closed in the weak topology of $\mathbf{B}$, and because $\hat{\Gamma} : \mathbf{M} \mapsto \mathbf{B}$ is continuous and linear under Assumption S.2. Hence,

$\{m \in \mathbf{M} : \|\hat{\beta} - \hat{\Gamma}(m)\|_{\mathbf{B}} \leq \hat{\kappa}_n^m\}$ being closed, together with Assumption S.7 and

$$\hat{\mathcal{M}}_n = \mathcal{M}_n \cap \{m \in \mathbf{M} : \|\hat{\beta} - \hat{\Gamma}(m)\|_{\mathbf{B}} \leq \hat{\kappa}_n^m\}, \tag{141}$$

shows that $\hat{\mathcal{M}}_n$ is closed in the weak topology, and therefore compact in the weak topology, since $\hat{\mathcal{M}}_n \subseteq \mathcal{M}$ and $\mathcal{M}$ is compact under Assumption S.1. Provided that $\hat{\mathcal{M}}_n \neq \emptyset$, it follows that for any $b^* \in \mathbf{B}^*$

$$\nu(b^*, \Gamma_P(\hat{\mathcal{M}}_n)) \equiv \sup_{m \in \hat{\mathcal{M}}_n} \langle b^*, \Gamma_P(m) \rangle = \max_{m \in \hat{\mathcal{M}}_n} \langle \Gamma_P^*(b^*), m \rangle, \tag{142}$$

where attainment in the final equality is guaranteed by the compactness of $\hat{\mathcal{M}}_n$ in the weak topology and the continuity of $m \mapsto \langle \Gamma_P^*(b^*), m \rangle$ in the weak topology. *Q.E.D.*

Lemma D.4. *Suppose that Assumptions S.3 and S.4 hold. Then*

$$\liminf_{n \to \infty} \inf_{P \in \mathbf{P}} P(\mathcal{D}_P(\kappa_n^u) \subseteq \hat{\mathcal{D}}_n) = 1$$

for any non-random sequence κ_n^u that satisfies Assumption S.8 and $\sqrt{n}\kappa_n^u \uparrow \infty$.

Proof of Lemma D.4. For any $b^* \in \mathbf{B}^*$ and $\mathcal{M}_n \subseteq \mathcal{M}$, define

$$\Delta_n(b^*) \equiv \sqrt{n} \left| \langle b^*, \hat{\beta} - \beta_P \rangle - \left\{ \nu(b^*, \hat{\Gamma}(\mathcal{M}_n)) - \nu(b^*, \Gamma_P(\mathcal{M}_n)) \right\} \right|. \tag{143}$$

Using the definition of $\mathcal{D}_P(\kappa_n^u)$, and noting that $\mathcal{M}_n \subseteq \mathcal{M}$ implies $\nu(b^*, \Gamma_P(\mathcal{M}_n)) \leq \nu(b^*, \Gamma_P(\mathcal{M}))$ for all $b^* \in \mathbf{B}^*$, we have

$$\begin{aligned} \inf_{b^* \in \mathcal{D}_P(\kappa_n^u)} \left\{ \langle b^*, \hat{\beta} \rangle - \nu(b^*, \hat{\Gamma}(\mathcal{M}_n)) \right\} &\geq \inf_{b^* \in \mathcal{D}_P(\kappa_n^u)} \left\{ \langle b^*, \beta_P \rangle - \nu(b^*, \Gamma_P(\mathcal{M}_n)) - \frac{\Delta_n(b^*)}{\sqrt{n}} \right\} \\ &\geq - \sup_{b^* \in \mathcal{D}_P(\kappa_n^u)} \left\{ \frac{\Delta_n(b^*)}{\sqrt{n}} + \kappa_n^u \right\}. \end{aligned} \tag{144}$$

The definition of $\hat{\mathcal{D}}_n$, together with (144) and Lemma D.2, then implies that

$$\begin{aligned} \liminf_{n \to \infty} \inf_{P \in \mathbf{P}_0} P(\mathcal{D}_P(\kappa_n^u) \subseteq \hat{\mathcal{D}}_n) &\geq \liminf_{n \to \infty} \inf_{P \in \mathbf{P}_0} P\left(\sup_{b^* \in \mathcal{D}_P(\kappa_n^u)} \left\{ \frac{\Delta_n(b^*)}{\sqrt{n}} + \kappa_n^u \right\} \leq \hat{\kappa}_n^u \right) \\ &\geq \liminf_{n \to \infty} \inf_{P \in \mathbf{P}_0} P((1+\delta)\kappa_n^u \leq \hat{\kappa}_n^u) = 1, \end{aligned} \tag{145}$$

where the final equality follows by Assumption S.8 for some $\delta > 0$. *Q.E.D.*

Lemma D.5. *Suppose that Assumptions S.3–S.5 and S.9 hold. Then*

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}_0} P(\Pi_n m_P \in \hat{\mathcal{M}}_n \textit{ for all } m_P \in \mathcal{M} \textit{ s.t. } \Gamma_P(m_P) = \beta_P) = 1.$$

Proof of Lemma D.5. Note that if $\beta_P = \Gamma_P(m_P)$, then by the triangle inequality

$$\begin{aligned} &\|\hat{\beta} - \hat{\Gamma}(\Pi_n m_P)\|_{\mathbf{B}} \\ &\quad \le \|\hat{\beta} - \beta_P\|_{\mathbf{B}} + \|\Gamma_P(m_P) - \Gamma_P(\Pi_n m_P)\|_{\mathbf{B}} + \|\Gamma_P(\Pi_n m_P) - \hat{\Gamma}(\Pi_n m_P)\|_{\mathbf{B}}. \end{aligned} \tag{146}$$

Using (134) from Lemma D.1 together with Assumption S.3, we have

$$\sup_{m\in\mathcal{M}} \|\sqrt{n}\{\hat{\Gamma} - \Gamma_P\}(m) - \mathbb{G}_{P,\Gamma}(m)\|_{\mathbf{B}} = o_p(1), \tag{147}$$

uniformly over $P \in \mathbf{P}$. Assumptions S.3 and S.5, together with (146) and (147), imply

$$\begin{aligned} &\sup_{m_P\in\mathcal{M}:\Gamma_P(m_P)=\beta_P} \sqrt{n}\|\hat{\beta} - \hat{\Gamma}(\Pi_n m_P)\|_{\mathbf{B}} \\ &\qquad \le \sup_{m_P\in\mathcal{M}:\Gamma_P(m_P)=\beta_P} \{\|\mathbb{G}_{P,\beta}\|_{\mathbf{B}} + \|\mathbb{G}_{P,\Gamma}(\Pi_n m_P)\|_{\mathbf{B}}\} + o_p(1), \end{aligned} \tag{148}$$

uniformly over $P \in \mathbf{P}$. Then by Assumption S.9, (148), and the definition of $\hat{\mathcal{M}}_n$,

$$\begin{aligned} &\liminf_{n\to\infty} \inf_{P\in\mathbf{P}_0} P(\Pi_n m_P \in \hat{\mathcal{M}}_n \text{ for all } m_P \in \mathcal{M} \text{ s.t. } \Gamma_P(m_P) = \beta_P) \qquad (149) \\ &\quad \ge \liminf_{n\to\infty} \inf_{P\in\mathbf{P}_0} P\left(\sup_{m_P\in\mathcal{M}:\Gamma_P(m_P)=\beta_P} \{\|\mathbb{G}_{P,\beta}\|_{\mathbf{B}} + \|\mathbb{G}_{P,\Gamma}(\Pi_n m_P)\|_{\mathbf{B}}\} \le \frac{\sqrt{n}\hat{\kappa}_n^m}{2}\right). \end{aligned}$$

Moreover, by Assumptions S.4, S.9, and Markov's inequality,

$$\begin{aligned} &\liminf_{n\to\infty} \inf_{P\in\mathbf{P}_0} P\left(\sup_{m\in\mathcal{M}} \{\|\mathbb{G}_{P,\beta}\|_{\mathbf{B}} + \|\mathbb{G}_{P,\Gamma}(m)\|_{\mathbf{B}}\} \le \frac{\sqrt{n}\hat{\kappa}_n^m}{2}\right) \\ &\qquad \ge \liminf_{C\uparrow\infty} \inf_{P\in\mathbf{P}_0} P\left(\sup_{m\in\mathcal{M}} \{\|\mathbb{G}_{P,\beta}\|_{\mathbf{B}} + \|\mathbb{G}_{P,\Gamma}(m)\|_{\mathbf{B}}\} \le C\right) = 1. \end{aligned} \tag{150}$$

The result follows from (149) and (150). *Q.E.D.*

Lemma D.6. *Suppose that Assumptions S.2–S.5, S.9, and S.11 hold. Define the set*

$$\tilde{\mathcal{G}}_n(b^*) \equiv \{g \in \mathbf{M}^* : |\langle g - \hat{\Gamma}^*(b^*), v\rangle| \le \hat{\kappa}_n^q \textit{ for all } v \in \mathcal{V}_n\}$$

and the statistic

$$\tilde{T}_n^{bs} \equiv \sup_{b^* \in \hat{\mathcal{D}}_n} \sup_{(\tilde{g},\tilde{m}) \in \tilde{\mathcal{G}}_n(b^*) \times \hat{\mathcal{M}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \Big\{ \langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m) \rangle \ s.t. \ \langle \tilde{g}, m - \tilde{m} \rangle \geq 0 \Big\} .$$

Then $T_n^{bs} = \tilde{T}_n^{bs}$ with probability tending to one, uniformly in $P \in \mathbf{P}_0$.

Proof of Lemma D.6. Since $\{m \in \mathcal{M} : \beta_P = \Gamma_P(m)\} \neq \emptyset$ for all $P \in \mathbf{P}_0$, Lemma D.5 implies that

$$\liminf_{n\to\infty} \inf_{P \in \mathbf{P}_0} P(\hat{\mathcal{M}}_n \neq \emptyset) = 1. \tag{151}$$

Hence, the claim follows provided that $T_n^{\text{bs}} = \tilde{T}_n^{\text{bs}}$ whenever $\hat{\mathcal{M}}_n \neq \emptyset$.

To show this, note that because $\mathbf{M}_n \subseteq \mathbf{M}$, the restriction of any $g \in \mathbf{M}^*$ to $\mathbf{M}_n$ is a continuous linear functional on $\mathbf{M}_n$. That is, for any $g \in \mathbf{M}^*$ there exists a $g_n \in \mathbf{M}_n^*$ such that $\langle g - g_n, m \rangle = 0$ for all $m \in \mathbf{M}_n$. Since $\mathcal{V}_n \subseteq \mathbf{M}_n$ by Assumption S.11, and because $\emptyset \neq \hat{\mathcal{M}}_n \subseteq \mathcal{M}_n \subseteq \mathbf{M}_n$, it follows that for any $\tilde{g} \in \tilde{\mathcal{G}}_n(b^*)$ there exists a $g \in \hat{\mathcal{G}}_n(b^*)$ such that $\langle g - \tilde{g}, m \rangle = 0$ for all $m \in \hat{\mathcal{M}}_n$. As a consequence,

$$\tilde{T}_n^{\text{bs}} \leq \sup_{b^* \in \hat{\mathcal{D}}_n} \sup_{(g,\tilde{m}) \in \hat{\mathcal{G}}_n(b^*) \times \hat{\mathcal{M}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \Big\{ \langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m) \rangle \text{ s.t. } \langle g, m - \tilde{m} \rangle \geq 0 \Big\} . \tag{152}$$

Conversely, note that any $g_n \in \mathbf{M}_n^*$ is a continuous linear functional defined on $\mathbf{M}_n$, which is a subspace of $\mathbf{M}$. By the Hahn-Banach Theorem (see e.g. Theorem 5.4.1 of Luenberger (1969)), there exists an extension $g \in \mathbf{M}^*$ such that $\langle g_n - g, m \rangle = 0$ for all $m \in \mathbf{M}_n$. As before, since $\mathcal{V}_n \subseteq \mathbf{M}_n$ under Assumption S.11, and because $\emptyset \neq \hat{\mathcal{M}}_n \subseteq \mathbf{M}_n$, we can conclude that for any $g \in \hat{\mathcal{G}}_n(b^*)$ there exists a $\tilde{g} \in \tilde{\mathcal{G}}_n(b^*)$ such that $\langle g - \tilde{g}, m \rangle = 0$ for all $m \in \hat{\mathcal{M}}_n$. Hence, we conclude that

$$\tilde{T}_n^{\text{bs}} \geq \sup_{b^* \in \hat{\mathcal{D}}_n} \sup_{(g,\tilde{m}) \in \hat{\mathcal{G}}_n(b^*) \times \hat{\mathcal{M}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \Big\{ \langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m) \rangle \text{ s.t. } \langle g, m - \tilde{m} \rangle \geq 0 \Big\} . \tag{153}$$

The result now follows from (152), (153), and the definition of T_n^{bs}. *Q.E.D.*

E Computation for Infinite B

Whenever **B** is infinite dimensional, i.e. we are employing an infinite number of IV-like specifications $\mathcal{S}$, the dual space $\mathbf{B}^*$ is infinite dimensional as well. Computing the bootstrap statistic T_n^{bs} may be challenging in this situation. In this appendix, we provide a supplemental result that establishes appropriate conditions under which T_n^{bs} can be approximated by a finite dimensional optimization problem when $\mathbf{B}^*$ is infinite dimensional.

Let $\mathcal{D}_n \subseteq \mathcal{D}$ be a finite dimensional subset of $\mathcal{D}$ and, in analogy to $\hat{\mathcal{D}}_n$, define

$$\tilde{\mathcal{D}}_n \equiv \{b^* \in \mathcal{D}_n : \langle b^*, \hat{\beta}\rangle - \nu(b^*, \hat{\Gamma}(\mathcal{M}_n)) \geq -\hat{\kappa}_n^u\}. \tag{154}$$

Heuristically, $\tilde{\mathcal{D}}_n$ consists of the set of "directions" b^* in the sieve $\mathcal{D}_n$ that are "close" to binding. Notice the contrast to $\hat{\mathcal{D}}_n$, which contains all such directions in $\mathcal{D}$ and hence is potentially infinite dimensional. Define the bootstrap statistic

$$C_n^{\text{bs}} \equiv \sup_{b^* \in \tilde{\mathcal{D}}_n} \sup_{(g,\tilde{m}) \in \hat{\mathcal{G}}_n(b^*) \times \hat{\mathcal{M}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \{\langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m)\rangle \text{ s.t. } \langle g, m - \tilde{m}\rangle \geq 0\}, \tag{155}$$

where $\hat{\mathcal{M}}_n$ and $\hat{\mathcal{G}}_n(b^*)$ remain as defined in (61) and (68), respectively. That is, C_n^{bs} is identical to T_n^{bs}, with the exception that $\hat{\mathcal{D}}_n$ has been replaced by $\tilde{\mathcal{D}}_n$.

In the following, we show that a version of Theorem 2 continues to hold with C_n^{bs} in place of T_n^{bs}. This implies that C_n^{bs} is also a valid bootstrap statistic. To do this, we impose the following conditions on the sieve $\mathcal{D}_n$ for $\mathcal{D}$:

Assumption U

For every $b^* \in \mathcal{D}$ there exists a $\Pi_n b^* \in \mathcal{D}_n \subseteq \mathcal{D}$ such that

$$\sup_{P \in \mathbf{P}} \sup_{b^* \in \mathcal{D}} \sqrt{n}|\langle b^* - \Pi_n b^*, \beta_P\rangle| \leq \delta_n^b,$$

$$\sup_{P \in \mathbf{P}} E\left[\sup_{(b^*,m) \in \mathcal{D} \times \mathcal{M}} |\langle b^* - \Pi_n b^*, \mathbb{G}_{P,\beta} - \mathbb{G}_{P,\Gamma}(m)\rangle|\right] \leq \delta_n^b,$$

$$\text{and} \quad \sup_{P \in \mathbf{P}} \sup_{(b^*,m) \in \mathcal{D} \times \mathcal{M}} \sqrt{n}|\langle b^* - \Pi_n b^*, \Gamma_P(m)\rangle| \leq \delta_n^b,$$

for some $\delta_n^b \downarrow 0$.

Assumption U places conditions on the sieve $\mathcal{D}_n$ that are analogous to those required of $\mathcal{M}_n$ by Assumption S.5. We note that Assumption U imposes no constraints on the rate of growth of $\mathcal{D}_n$. Instead, Assumption U demands that the rate of growth of $\mathcal{D}_n$ be *sufficiently fast* so that the approximation error introduced from optimizing over $\mathcal{D}_n$ in place of $\mathcal{D}$ is asymptotically negligible. This is likely to be the case when computation is the primary reason for using $\mathcal{D}_n$ over $\mathcal{D}$. Our next result provides an analog to Theorem 2 for situations when a sieve $\mathcal{D}_n$ is used in place of $\mathcal{D}$.

Lemma E.1. *Suppose that Assumptions S.1–S.7, S.9–S.11, and U hold. Then, for any sequence κ_n^u that satisfies Assumption S.8, as well as $\sqrt{n}\kappa_n^u \uparrow \infty$, there exists a* *sequence $\xi_n^{bs} \in \mathbf{R}$, with $\xi_n^{bs} = O_p(\delta_n^c + \delta_n^b)$ uniformly in $P \in \mathbf{P}_0$, and such that*

$$\mathbb{U}_{P,n}^{bs}(\kappa_n^u) \leq C_n^{bs} + \xi_n^{bs} \tag{156}$$

for any $P \in \mathbf{P}_0$.

Proof of Lemma E.1. We begin by defining an auxiliary lower bound C_n^L. Note that since $\hat{\kappa}_n^u$ satisfies Assumption S.8, there exists an $\eta \in (0,1)$ such that

$$\liminf_{n\to\infty} \inf_{P\in\mathbf{P}} P\left(\frac{\eta\hat{\kappa}_n^u}{\kappa_n^u} > (1+\delta^L)\right) = 1 \tag{157}$$

for some $\delta^L > 0$. Define

$$\begin{aligned} \hat{\mathcal{D}}_n^L &\equiv \{b^* \in \mathcal{D} : \langle b^*, \hat{\beta}\rangle - \nu(b^*, \hat{\Gamma}(\mathcal{M}_n)) \geq -\eta\hat{\kappa}_n^u\} \\ \text{and} \quad \hat{\mathcal{G}}_n^L(b^*) &\equiv \{g \in \mathbf{M}_n^* : |\langle g, v\rangle - \langle b^*, \hat{\Gamma}(v)\rangle| \leq \eta\hat{\kappa}_n^g \text{ for all } v \in \mathcal{V}_n\}, \end{aligned}$$

then let

$$C_n^L \equiv \sup_{b^*\in\hat{\mathcal{D}}_n^L} \sup_{(g,\tilde{m})\in\hat{\mathcal{G}}_n^L(b^*)\times\hat{\mathcal{M}}_n} \inf_{m\in\hat{\mathcal{M}}_n} \left\{\langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m)\rangle \text{ s.t. } \langle g, m-\tilde{m}\rangle \geq 0\right\}. \tag{158}$$

Note that Assumption S.10 is satisfied with $\eta\hat{\kappa}_n^g$ replacing $\hat{\kappa}_n^g$, and that (157) implies that Assumption S.8 is also satisfied with $\eta\hat{\kappa}_n^u$ in place of $\hat{\kappa}_n^u$. Hence, from Theorem 2, we can conclude that

$$\mathbb{U}_{P,n}^{\text{bs}}(\kappa_n^u) \leq C_n^L + \xi_n^L, \tag{159}$$

uniformly over $P \in \mathbf{P}_0$, with $\xi_n^L = O_p(\delta_n^c)$.

Next, note that because $\mathcal{D}_n \subseteq \mathcal{D}$ by Assumption U, and since $|\langle b^*, b\rangle| \leq \|b\|_{\mathbf{B}}$ for any $b^* \in \mathcal{D}$, we have

$$\begin{aligned} &\sup_{b^*\in\mathcal{D}} \sup_{m\in\mathcal{M}} |\langle b^* - \Pi_n b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m)\rangle - \langle b^* - \Pi_n b^*, \mathbb{G}_{P,\beta}^{\text{bs}} - \mathbb{G}_{P,\Gamma}^{\text{bs}}(m)\rangle| \\ &\qquad \leq 2\|\hat{\mathbb{G}}_\beta - \mathbb{G}_{P,\beta}^{\text{bs}}\|_{\mathbf{B}} + \sup_{b^*\in\mathcal{D}} \sup_{m\in\mathcal{M}} 2|\langle b^*, \hat{\mathbb{G}}_\Gamma(m) - \mathbb{G}_{P,\Gamma}^{\text{bs}}(m)\rangle| = O_p(\delta_n^c), \end{aligned} \tag{160}$$

uniformly over $P \in \mathbf{P}$, where the final equality is due to Assumption S.6. Moreover,

$$\sup_{b^*\in\mathcal{D}} \sup_{m\in\mathcal{M}} |\langle b^* - \Pi_n b^*, \mathbb{G}_{P,\beta}^{\text{bs}} - \mathbb{G}_{P,\Gamma}^{\text{bs}}(m)\rangle| = O_p(\delta_n^b) \tag{161}$$

uniformly in $P \in \mathbf{P}$ by Assumptions S.6, U, and Markov's inequality. Together, (160),

(161) imply that

$$C_n^L = \sup_{b^* \in \hat{\mathcal{D}}_n^L} \sup_{(g,\tilde{m}) \in \hat{\mathcal{G}}_n^L(b^*) \times \hat{\mathcal{M}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \Big\{ \langle \Pi_n b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m) \rangle \text{ s.t. } \langle g, m - \tilde{m} \rangle \geq 0 \Big\} \\ + O_p(\delta_n^c + \delta_n^b) \tag{162}$$

uniformly in $P \in \mathbf{P}$.

Define the event $\Omega_n(P) \equiv \Omega_{1n}(P) \cap \Omega_{2n}(P)$, where

$$\Omega_{1n}(P) \equiv \Big[\Pi_n b^* \in \tilde{\mathcal{D}}_n \text{ for all } b^* \in \hat{\mathcal{D}}_n^L \Big] \tag{163}$$

$$\Omega_{2n}(P) \equiv \Big[\hat{\mathcal{G}}_n^L(b^*) \subseteq \hat{\mathcal{G}}_n(\Pi_n b^*) \text{ for all } b^* \in \mathcal{D} \Big] . \tag{164}$$

Then note that by definition of $\tilde{\mathcal{D}}_n$ and $\hat{\mathcal{D}}_n^L$, with $\eta \in (0,1)$, we obtain

$$P(\Omega_{1n}(P)) \\ \geq P\left(\inf_{b^* \in \mathcal{D}} \langle \Pi_n b^* - b^*, \hat{\beta} \rangle - \Big\{ \nu(\Pi_n b^*, \hat{\Gamma}(\mathcal{M}_n)) - \nu(b^*, \hat{\Gamma}(\mathcal{M}_n)) \Big\} \geq -(1-\eta) \hat{\kappa}_n^u \right) \\ \geq P\left(\sup_{(b^*,m) \in \mathcal{D} \times \mathcal{M}} |\langle \Pi_n b^* - b^*, \hat{\beta} - \hat{\Gamma}(m) \rangle| \leq (1-\eta) \hat{\kappa}_n^u \right) . \tag{165}$$

By Assumptions S.3, U, and Markov's inequality we obtain

$$\sup_{(b^*,m) \in \mathcal{D} \times \mathcal{M}} |\langle \Pi_n b^* - b^*, \hat{\beta} - \hat{\Gamma}(m) \rangle| \tag{166}$$
$$\leq \sup_{(b^*,m) \in \mathcal{D} \times \mathcal{M}} \left| \frac{1}{\sqrt{n}} \langle \Pi_n b^* - b^*, \mathbb{G}_{P,\beta} - \mathbb{G}_{P,\Gamma}(m) \rangle \right| + O_p\left(\frac{\delta_n^c + \delta_n^b}{\sqrt{n}} \right) = O_p\left(\frac{\delta_n^c + \delta_n^b}{\sqrt{n}} \right)$$

uniformly in $P \in \mathbf{P}$. Since $\delta_n^c \downarrow 0$ by Assumption S.3, $\delta_n^b \downarrow 0$ by Assumption U, and $\hat{\kappa}_n^u$ satisfies Assumption S.8 with $\sqrt{n}\kappa_n^u \uparrow \infty$, we conclude from (165) and (166) that

$$\liminf_{n \to \infty} \inf_{P \in \mathbf{P}} P(\Omega_{1n}(P)) = 1. \tag{167}$$

Similarly, the definitions of $\hat{\mathcal{G}}_n^L(b^*)$ and $\hat{\mathcal{G}}_n(\Pi_n b^*)$, $\eta \in (0,1)$, and $\mathcal{V}_n \subseteq \lambda(\mathcal{M} - \mathcal{M})$ yield

$$P(\Omega_{2n}(P)) \geq P\left(\sup_{b^* \in \mathcal{D}} |\langle \Pi_n b^* - b^*, \hat{\Gamma}(v) \rangle| \leq (1-\eta) \hat{\kappa}_n^g \text{ for all } v \in \mathcal{V}_n \right) \\ \geq P\left(\sup_{(b^*,m) \in \mathcal{D} \times \mathcal{M}} |\langle \Pi_n b^* - b^*, \hat{\Gamma}(m) \rangle| \leq \frac{(1-\eta)}{2\lambda} \hat{\kappa}_n^g \right) . \tag{168}$$

After arguments analogous to those in (166) (evaluated with $\hat{\beta} = 0$), (168) implies

$$\liminf_{n \to \infty} \inf_{P \in \mathbf{P}} P(\Omega_{2n}(P)) = 1. \tag{169}$$

Since $\Omega_n(P) \equiv \Omega_{1n}(P) \cap \Omega_{2n}(P)$, results (167) and (169) establish that $\Omega_n(P)$ occurs with probability tending to one, uniformly over $P \in \mathbf{P}$.

To conclude, observe that when $\Omega_n(P)$ occurs, we have

$$\begin{aligned} &\sup_{b^* \in \hat{\mathcal{D}}_n^L} \sup_{(g,\tilde{m}) \in \hat{\mathcal{G}}_n^L(b^*) \times \hat{\mathcal{M}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \left\{ \langle \Pi_n b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m) \rangle \text{ s.t. } \langle g, m - \tilde{m} \rangle \geq 0 \right\} \\ &\quad \leq \sup_{b^* \in \tilde{\mathcal{D}}_n} \sup_{(g,\tilde{m}) \in \hat{\mathcal{G}}_n(b^*) \times \hat{\mathcal{M}}_n} \inf_{m \in \hat{\mathcal{M}}_n} \left\{ \langle b^*, \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m) \rangle \text{ s.t. } \langle g, m - \tilde{m} \rangle \geq 0 \right\}. \end{aligned} \tag{170}$$

The claim now follows from the definition of C_n^{bs} by combining (159), (162), (169), and (170). *Q.E.D.*

F Bernstein Polynomials

The kth Bernstein basis polynomial of degree K is defined as

$$b_k^K : [0,1] \mapsto \mathbf{R} : b_k^K(u) \equiv \binom{K}{k} u^k (1-u)^{K-k}$$

for $k = 0, 1, \ldots, K$. A degree K Bernstein polynomial B is a linear combination of these $K + 1$ basis polynomials:

$$B(u) : [0,1] \mapsto \mathbf{R} : B(u) \equiv \sum_{k=0}^{K} \theta_k b_k^K(u),$$

for some constants $\theta_0, \theta_1, \ldots, \theta_K$. As is well-known, any continuous function on $[0,1]$ can be uniformly well approximated by a Bernstein polynomial of sufficiently high order.

The shape of B can be constrained by imposing linear restrictions on $\theta_0, \theta_1, \ldots, \theta_K$. This computationally appealing property of the Bernstein polynomials has been noted elsewhere by Chak, Madras, and Smith (2005), Chang, Chien, Hsiung, Wen, and Wu (2007) and McKay Curtis and Ghosh (2011), among others. The following constraints are especially useful in the current application. Derivations of these properties can be found in Chang et al. (2007) and McKay Curtis and Ghosh (2011).

Shape Constraints

S.1 *Bounded below by* 0*:* $\theta_k \geq 0$ *for all* k*.*

S.2 *Bounded above by* 1*:* $\theta_k \leq 1$ *for all* k*.*

S.3 *Monotonically increasing:* $\theta_0 \leq \theta_1 \leq \cdots \leq \theta_K$*.*

S.4 *Concave:* $\theta_k - 2\theta_{k+1} + \theta_{k+2} < 0$ *for* $k = 0, \ldots, K-2$*.*

Each Bernstein basis polynomial is itself an ordinary degree K polynomial. The coefficients on this ordinary polynomial representation (i.e. the power basis representation) can be computed by applying the binomial theorem:

$$b_k^K(u) = \sum_{i=k}^{K} (-1)^{i-k} \binom{K}{i} \binom{i}{k} u^i. \tag{171}$$

Representation (171) is useful for computing the terms $\Gamma_d^\star(b_{dk})$ and $\Gamma_{ds}(b_{dk})$ that appear in the finite-dimensional program (25). To see this note for example that with $d = 1$,

$$\Gamma_{1s}(b_{1k}) \equiv E\left[\int_0^1 b_{1k}(u, Z)\omega_{1s}(u, Z)\, du\right] = E\left[s(1, Z)\int_0^{p(Z)} b_{1k}(u, Z)\, du\right]$$

If $b_{1k}(u, Z) = b_{1k}(u)$ is a Bernstein basis polynomial, then $\int_0^{p(Z)} b_{1k}(u)\, du$ can be computed analytically through elementary calculus using (171). The result of this integral is a known function of $p(Z)$. The coefficient $\Gamma_{1s}(b_{1k})$ is then simply the population average of the product of this known function with $s(0, Z)$, which is also known or identified. Thus, no numerical integration is needed to compute or estimate the γ_{dks} terms. This conclusion depends on the form of the weights, and may not hold for all target weights $\omega_{dk}^\star$, although it holds for all of the parameters listed in Table 1. When it does not, one dimensional numerical integration can be used instead.

G Implementation and Computation

In this appendix, we discuss computation for the sample analog bounds, the test statistic, the bootstrap statistic, and our data-driven choices of tuning parameters for the bootstrap statistic.

G.1 Estimating Bounds

Consider the finite dimensional problem (25). We assume throughout Appendix G that Θ is polyhedral, so that it can be represented as

$$\Theta \equiv \{\theta \in \mathbf{R}^{d_\theta} : R\theta \leq q\} \tag{172}$$

for a known vector $q \in \mathbf{R}^{d_q}$ and a known $d_q \times (K_0 + K_1)$ dimensional matrix R, where $\theta \equiv (\theta_0, \theta_1)$, and $d_\theta \equiv K_0 + K_1$. We also assume throughout Appendix G that $\Theta \subset \mathbf{R}^{d_\theta}$ is a bounded set.

Let $\hat{\Gamma}^\star_d(b_{dk})$ and $\hat{\Gamma}_{ds}(b_{dk})$ denote estimators of $\Gamma^\star_d(b_{dk})$ and $\Gamma_{ds}(b_{dk})$, respectively. For the target parameter, these can be constructed as, e.g.

$$\hat{\Gamma}^\star_d(b_{dk}) \equiv \frac{1}{n} \sum_{i=1}^{n} \int_0^1 b_{dk}(u, X_i) \hat{\omega}^\star_d(u, Z_i) \, d\mu^\star(u), \tag{173}$$

where $\hat{\omega}^\star_d$ is an estimator of the known or identified weighting function, $\hat{\omega}^\star_d$. Depending on the choices of basis and target parameter, the integral can often be evaluated analytically. An estimator analogous to (173) can also be constructed for each $\Gamma_{ds}(b_{dk})$. These require an estimator $\hat{\omega}_{ds}$, which in turn requires an estimator for the propensity score, $p(z)$, and possibly the functions $s(d, z)$ that define the IV–like estimand. Letting $\hat{s}(d, z)$ be the latter estimator, we then define

$$\hat{\beta}_s \equiv \frac{1}{n} \sum_{i=1}^{n} \hat{s}(D_i, Z_i) Y_i$$

as an estimator of β_s.

Given these estimators, the bound $\overline{\beta}^\star_{\text{fd}}$ can be estimated by solving the linear program

$$\begin{aligned} \hat{\overline{\beta}}^\star_{\text{fd}} \equiv \max_{\theta \equiv (\theta_0, \theta_1)} & \sum_{k=1}^{K_0} \theta_{0k} \hat{\Gamma}^\star_0(b_{0k}) + \sum_{k=1}^{K_1} \theta_{1k} \hat{\Gamma}^\star_1(b_{1k}) \\ \text{s.t.} \quad & \sum_{k=1}^{K_0} \theta_{0k} \hat{\Gamma}_{0s}(b_{0k}) + \sum_{k=1}^{K_1} \theta_{1k} \hat{\Gamma}_{1s}(b_{1k}) = \hat{\beta}_s \text{ for all } s \in \mathcal{S} \\ \text{and} \quad & R\theta \leq q. \end{aligned} \tag{174}$$

Solving the analogous minimization problem yields an estimator of the lower bound, $\hat{\underline{\beta}}^\star_{\text{fd}}$. Both of these problems are linear programs with $d_\theta \equiv K_0 + K_1$ variables and $|\mathcal{S}| + d_q$ constraints.

It is possible that no solution to (174) exists. This could be an indication that no solution to the population problem (25) exists either, which would imply that the model is misspecified. However, it could also be that (25) is feasible, but that (174) is infeasible due to statistical error in the estimation of the $\Gamma_{ds}(b_{dk})$ and β_s terms. Our results in Section 5 provide a formal statistical test of the null hypothesis that the model is not misspecified. Those results also provide a procedure for building a confidence region for $\beta^\star$. It is possible for this confidence region to be nonempty even when (174) is infeasible.

G.2 Reformulation of the Test Statistic

We continue to assume that m has been parameterized by some finite dimensional $\theta \in \Theta$, where Θ is polyhedral and characterized by the linear constraints $R\theta \leq q$, as in (172). We also assume that β is a finite dimensional parameter so that $\mathbf{B} = \mathbf{R}^{d_\beta}$.

The test statistic, T_n, is defined by the optimization problem (35). The choice of norm on $\mathbf{B}$ affects the nature of the objective in (35), and will affect both the objective and constraints for the bootstrap statistic problem discussed in the next section. For computational reasons, it turns out to be convenient to choose a norm $\|\cdot\|_{\mathbf{B}}$ for which the unit ball is polyhedral. This suggests taking $\|\cdot\|_{\mathbf{B}}$ to be either the 1–norm or the max–norm on $\mathbf{R}^{d_\beta}$. For concreteness, we will take $\|\cdot\|_{\mathbf{B}}$ to be the 1–norm, so that $\|\hat{\beta}\|_{\mathbf{B}} = \|\hat{\beta}\|_1 \equiv \sum_{l=1}^{d_\beta} |\hat{\beta}_l|$.

Given these choices, we can rewrite (35) as

$$T_n = \min_{\theta} \sqrt{n}\|\hat{\beta} - \hat{\Gamma}\theta\|_1 \quad \text{s.t.} \quad R\theta \leq q. \tag{175}$$

This problem can be reformulated as a linear program by introducing non-negative slack variables, $w^+, w^- \in \mathbf{R}^{d_\beta}$. Specifically, it can be shown that (175) is equivalent to

$$\begin{aligned} T_n = \min\nolimits_{\theta, w^+, w^-} \quad & \sqrt{n}\left(\textstyle\sum_{l=1}^{d_\beta} w_l^+ + w_l^-\right) \\ \text{s.t.} \quad & R\theta \leq q \\ & w^+ - w^- = \hat{\beta} - \hat{\Gamma}\theta \\ & w^-, w^+ \geq 0. \end{aligned} \tag{176}$$

For a discussion of this reformulation see e.g. Boyd and Vandenberghe (2004, pg. 294).

G.3 Reformulation of the Bootstrap Statistic

In this section, we show that the optimization problem that defines the bootstrap statistic, i.e. (70), can be reformulated as a bilinear program. Throughout this section, we assume that the researcher has selected values of the tuning parameters $\hat{\kappa}_n^m, \hat{\kappa}_n^u$, and $\hat{\kappa}_n^g$. In the next section, we discuss the computational aspects of our data-driven recommendations for choosing these parameters.

Consider the inner minimization problem in (70). We rewrite this problem here in terms of $\theta \in \hat{\Theta}$ (replacing $m \in \hat{\mathcal{M}}_n$) as

$$\begin{aligned} T^{\text{bs}}_{n,\text{inner}}(b^*, g, \tilde{\theta}) \equiv \min\nolimits_{\theta} \quad & (\hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma \theta)' b^* \\ \text{s.t.} \quad & g'(\theta - \tilde{\theta}) \geq 0 \\ & R\theta \leq q \\ & \|\hat{\beta} - \hat{\Gamma}\theta\|_1 \leq \hat{\kappa}^m_n, \end{aligned} \tag{177}$$

where we continue to use the notation $\hat{\kappa}^m_n$ despite having exchanged m's for θ's elsewhere. We write this inner problem as a function of $(b^*, g, \tilde{\theta})$ to emphasize that these variables are fixed from the outer maximization problem in (70). Using a similar idea as in the reformulation (176) for the test statistic, we rewrite (177) as the following linear program:

$$\begin{aligned} T^{\text{bs}}_{n,\text{inner}}(b^*, g, \tilde{\theta}) = \min\nolimits_{\theta, w^+, w^-} \quad & (\hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma \theta)' b^* \\ \text{s.t.} \quad & g'(\theta - \tilde{\theta}) \geq 0 \\ & R\theta \leq q \\ & w^+ - w^- = \hat{\beta} - \hat{\Gamma}\theta \\ & \textstyle\sum_{l=1}^{d_\beta} w^+_l + w^-_l \leq \hat{\kappa}^m_n \\ & w^-, w^+ \geq 0. \end{aligned} \tag{178}$$

As long as $\hat{\Theta}_n \equiv \{\theta \in \Theta : \|\hat{\beta} - \hat{\Gamma}\theta\|_1 \leq \hat{\kappa}^m_n\}$ is nonempty, a feasible solution to (178) can always be achieved by taking $\theta = \tilde{\theta}$ since $\tilde{\theta} \in \hat{\Theta}_n$ from the outer optimization constraint in (70). We assume that $\hat{\kappa}^m_n$ has been chosen to be sufficiently large to ensure this is the case, which simply requires $\hat{\kappa}^m_n \geq T_n/\sqrt{n}$. Together with our assumption that Θ is bounded, we can conclude that strong duality holds; see, e.g. Corollary 5.3.7 in Borwein and Lewis (2010). The dual of (178) can be shown to be

$$\begin{aligned} T^{\text{bs}}_{n,\text{inner}}(b^*, g, \tilde{\theta}) = \max\nolimits_{\sigma} \quad & \hat{\mathbb{G}}'_\beta b^* + q'\sigma_1 + \hat{\kappa}^m_n \sigma_2 - g'\tilde{\theta}\sigma_3 + \hat{\beta}'\sigma_4 \\ \text{s.t.} \quad & \sigma_1, \sigma_2, \sigma_3 \leq 0 \\ & R'\sigma_1 + \hat{\Gamma}'\sigma_4 - g\sigma_3 = -\hat{\mathbb{G}}'_\Gamma b^* \\ & \sigma_2 \leq \sigma_{4,l} \leq -\sigma_2 \quad \text{for } l = 1, \ldots, d_\beta, \end{aligned} \tag{179}$$

where $\sigma = (\sigma'_1, \sigma_2, \sigma_3, \sigma'_4)'$ with $\sigma_1 \in \mathbf{R}^{d_q}$, $\sigma_2 \in \mathbf{R}$, $\sigma_3 \in \mathbf{R}$, and $\sigma_4 \in \mathbf{R}^{d_\beta}$. Given (179), we can now write (70) as a single maximization problem:

$$\begin{aligned} T_n^{\mathrm{bs}} = \max_{b^*,g,\tilde{\theta},\sigma} \quad & \hat{\mathbb{G}}_\beta' b^* + q'\sigma_1 + \hat{\kappa}_n^m \sigma_2 - g'\tilde{\theta}\sigma_3 + \hat{\beta}'\sigma_4 \\ \text{s.t.} \quad & \sigma_1, \sigma_2, \sigma_3 \leq 0 \\ & R'\sigma_1 + \hat{\Gamma}'\sigma_4 - g\sigma_3 = -\hat{\mathbb{G}}_\Gamma' b^* \\ & \sigma_2 \leq \sigma_{4,l} \leq -\sigma_2 \quad \text{for } l = 1, \ldots, d_\beta. \\ & b^* \in \hat{\mathcal{D}}_n \quad g \in \hat{\mathcal{G}}_n(b^*) \\ & R\tilde{\theta} \leq q \\ & \|\hat{\beta} - \hat{\Gamma}\tilde{\theta}\|_1 \leq \hat{\kappa}_n^m. \end{aligned} \tag{180}$$

Our next task is to reformulate (180) into a bilinear maximization problem. This involves several steps.

First, we reformulate the constraint $b^* \in \hat{\mathcal{D}}_n$. To this end, note that $\mathbf{B}^* = \mathbf{R}^{d_\beta}$ but equipped with the max–norm $\|\cdot\|_\infty$, defined as $\|b^*\|_\infty \equiv \max_l |b_l^*|$. Hence, by definition of $\hat{\mathcal{D}}_n$ (see (58)) the constraint $b^* \in \hat{\mathcal{D}}_n$ is equivalent to

$$\|b^*\|_\infty \leq 1 \quad \text{and} \quad \hat{\beta}' b^* - \nu(b^*, \hat{\Gamma}(\Theta)) \geq -\hat{\kappa}_n^u. \tag{181}$$

The first constraint in (181) can be rewritten as the set of linear constraints $-1 \leq b_l^* \leq 1$ for $l = 1, \ldots, d_\beta$. To reformulate the second constraint in (181), we rephrase the definition of a support function (see (38)) as a minimization problem by applying strong duality; see, e.g., Corollary 5.3.7 in Borwein and Lewis (2010). That is,

$$\nu(b^*, \hat{\Gamma}(\Theta)) = \begin{pmatrix} \max_\theta & ((b^*)'\hat{\Gamma})\theta \\ \text{s.t.} & R\theta \leq q \end{pmatrix} = \begin{pmatrix} \min_{\sigma_5 \geq 0} & q'\sigma_5 \\ \text{s.t.} & R'\sigma_5 = \hat{\Gamma}' b^* \end{pmatrix}.$$

Using the minimization form of the support function, we conclude that the second constraint in (181) is equivalent to the existence of a $\sigma_5 \in \mathbf{R}^{d_q}$ such that

$$\sigma_5 \geq 0 \quad \text{and} \quad R'\sigma_5 = \hat{\Gamma}' b^* \quad \text{and} \quad q'\sigma_5 \leq \hat{\beta}' b^* + \hat{\kappa}_n^u. \tag{182}$$

Notice that the constraints in (182) are linear in the variables of optimization, which now include the dual variables, σ_5.

Second, we reformulate the constraint that $g \in \mathcal{G}_n(b^*)$. To implement this constraint, we take $\mathcal{V}_n = \{\pm e_j\}_{j=1}^{d_\theta}$, where e_j is the jth unit basis vector in $\mathbf{R}^{d_\theta}$. Then, from (69), $g \in \mathcal{G}_n(b^*)$ is equivalent to $\|g - (b^*)'\hat{\Gamma}\|_\infty \leq \hat{\kappa}_n^g$. As noted above, this max norm constraint can also be written as a set of linear inequalities.

Third, we observe that the transpose of the equality constraint in (180) implies that

$$\sigma_3 g' = \sigma_1' R + \sigma_4' \hat{\Gamma} + (b^*)' \hat{\mathbb{G}}_{\Gamma}.$$

We substitute this expression into the term $g'\tilde{\theta}\sigma_3 = \sigma_3 g'\tilde{\theta}$ in the objective of (180). This allows us to rewrite the objective as

$$\hat{\mathbb{G}}_{\beta}' b^* + q'\sigma_1 + \hat{\kappa}_n^m \sigma_2 - \sigma_1' R\tilde{\theta} - \sigma_4' \hat{\Gamma}\tilde{\theta} - (b^*)'\hat{\mathbb{G}}_{\Gamma}\tilde{\theta} + \hat{\beta}'\sigma_4. \tag{183}$$

The benefit of this substitution is that, whereas the term $\sigma_3\gamma'\tilde{\theta}$ in the objective of (180) is the product of three variables of optimization, every term in (183) is the product of at most two variables of optimization.

Fourth, we recall the reformulation that we used on the inner problem to move from (177) to (178). Here, it will be applied to the constraint $\|\hat{\beta} - \hat{\Gamma}\tilde{\theta}\|_1 \leq \hat{\kappa}_n^m$.

Incorporating these four observations into (180), we reformulate the program as

$$\begin{array}{rl}
T_n^{\text{bs}} = \max & \hat{\mathbb{G}}_{\beta}' b^* + q'\sigma_1 + \hat{\kappa}_n^m \sigma_2 - \sigma_1' R\tilde{\theta} - \sigma_4' \hat{\Gamma}\tilde{\theta} - (b^*)'\hat{\mathbb{G}}_{\Gamma}\tilde{\theta} + \hat{\beta}'\sigma_4 \\
\text{as a function of} & b^*, g, \tilde{\theta}, \sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \tilde{w}^+, \tilde{w}^- \\
\text{s.t.} & \sigma_1, \sigma_2, \sigma_3 \leq 0, \quad \sigma_5, \tilde{w}^+, \tilde{w}^- \geq 0 \\
& R'\sigma_1 + \hat{\Gamma}'\sigma_4 - g\sigma_3 = -\hat{\mathbb{G}}_{\Gamma}' b^* \\
& \sigma_2 \leq \sigma_{4,l} \leq -\sigma_2 \quad \text{for } l = 1, \ldots, d_{\beta}. \\
& -1 \leq b_l^* \leq 1 \quad \text{for } l = 1, \ldots, d_{\beta} \\
& R'\sigma_5 = \hat{\Gamma}' b^* \\
& q'\sigma_5 \leq \hat{\beta}' b^* + \hat{\kappa}_n^u \\
& -\hat{\kappa}_n^g \leq e_j'(g - (b^*)'\hat{\Gamma}) \leq \hat{\kappa}_n^g \quad \text{for } j = 1, \ldots, d_{\theta} \\
& R\tilde{\theta} \leq q \\
& \tilde{w}^+ - \tilde{w}^- = \hat{\beta} - \hat{\Gamma}\tilde{\theta} \\
& \sum_{l=1}^{d_{\beta}} \tilde{w}_l^+ + \tilde{w}_l^- \leq \hat{\kappa}_n^m.
\end{array} \tag{184}$$

This program is almost linear in both its objective and constraints. However, it does contain the following terms that are bilinear in the sense of being the product of two different variables of optimization: $\sigma_1 R\tilde{\theta}$, $\sigma_4'\hat{\Gamma}\tilde{\theta}$, $(b^*)'\hat{\mathbb{G}}_{\Gamma}\tilde{\theta}$, and $g\sigma_3$. As a result, (184) is a bilinear programming problem. Despite being non-convex, bilinear programs like these can be reliably solved to global optimality, see e.g. Tawarmalani and Sahinidis (2005) and the references cited therein.

G.4 Reformulation of the Tuning Parameter Selection Problems

We use (84) to choose $\hat{\kappa}_n^u$, as well as $\hat{\kappa}_n^m$, which we take as $\hat{\kappa}_n^m = \hat{\kappa}_n^u + T_n/\sqrt{n}$. To do this, we compute $\sup_{m \in \mathcal{M}_n} \|\hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\Gamma(m)\|_{\mathbf{B}}$. In terms of the finite dimensional framework used throughout this section, this problem can be written as

$$\max_\theta \|\hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\beta \theta\|_1 \quad \text{s.t.} \quad R\theta \leq q. \tag{185}$$

This problem looks superficially similar to the reformulated test statistic problem, (175), with the important difference that it is a maximization problem, rather than a minimization problem. Since $\|\cdot\|_1$ is a convex function, this means that (185) is a non-convex optimization problem.

However, (185) can be reformulated as a mixed integer linear program (MILP) by applying a fairly standard argument. The argument is based on the observation that the nonlinearity of the objective comes from the absolute value function, i.e.:

$$\|\hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\beta \theta\|_1 \equiv \sum_{l=1}^{d_\beta} \left| e_l' \left(\hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\beta \theta \right) \right|,$$

where e_l are unit vectors in $\mathbf{R}^{d_\beta}$. As a result, the objective can be linearized by introducing d_β binary variables that indicate whether each absolute value is obtained for a positive or negative number, together with d_β slack variables to stand in for the magnitude of the absolute value itself. Specifically, (185) is equivalent to

$$\begin{aligned} \max\nolimits_{\theta,\zeta,\sigma,w} \quad & \textstyle\sum_{l=1}^{d_\beta} \sigma_l \\ \text{s.t.} \quad & R\theta \leq q \\ & \zeta_l \in \{0,1\} \quad \text{for } l = 1,\ldots,d_\beta \\ & w = \hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\beta \theta \\ & w_l \leq \sigma_l \leq w_l + (1-\zeta_l)\text{BigM}_l \quad \text{for } l = 1,\ldots,d_\beta \\ & -w_l \leq \sigma_l \leq -w_l + \zeta_l \text{BigM}_l \quad \text{for } l = 1,\ldots,d_\beta \end{aligned} \tag{186}$$

In (186), ζ are binary variables, w is a definitional variable that helps with notation, and BigM_l are large numbers, referred to as "big M" parameters in the operations research literature (e.g. pp. 136-137 of Schrijver (1998)). The big M parameters are chosen by the researcher in such a way as to ensure that constraints in which they enter are never binding when BigM_l is multiplied by 1. It is important to note that these parameters are not tuning parameters in the usual statistical sense. In particular, while the choice of the big M parameters can impact the speed at which (186) is solved, these parameters can (and should) always be chosen so as not to impact the optimal value of (186).

To see how (186) reformulates (185), first note that we have set $w_l = e_l'(\hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\beta \theta)$ in (186) as a (notational) constraint. Next, observe that if $w_l \geq 0$, then $w_l \leq \sigma_l \leq -w_l$ is a contradiction, so the only feasible solution in this case is to take $\zeta_l = 1$. However, with $\zeta_l = 1$, the constraints enforce $\sigma_l = w_l$. Similarly, if $w_l \leq 0$, then the binary constraints enforce any feasible solution to have $\zeta_l = 0$ and hence $\sigma_l = -w_l$. In both cases, $\sigma_l = |e_l'(\hat{\mathbb{G}}_\beta - \hat{\mathbb{G}}_\beta \theta)|$, so that the objective in (186) is always identical to that in (185).

Using (86) to select $\hat{\kappa}_n^g$ involves solving a problem similar to (185), which can also be reformulated as a MILP. In finite dimensions, and with the choice of $\mathcal{V}_n$ discussed in the previous section, this problem can be written as

$$\max_{b^*} \|(b^*)'\hat{\mathbb{G}}_\Gamma\|_\infty \quad \text{s.t.} \quad \|b^*\|_\infty \leq 1 \quad \text{and} \quad \nu(b^*, \hat{\Gamma}(\Theta)) \leq \hat{\beta}' b^* + \hat{\kappa}_n^u. \tag{187}$$

We have already shown how to reformulate the constraints in this problem as linear constraints; recall (182). However, as with (185), the objective in (187) is a nonlinear, convex function, so maximizing it subject to linear constraints is a non-convex problem.

We reformulate (187) as the following MILP:

$$\begin{aligned}
\max \quad & \sigma_1 \\
\text{as a function of} \quad & b^*, \sigma_1, \sigma_2, w, \pi, \zeta_1, \zeta_2 \\
\text{s.t.} \quad & -1 \leq b_l^* \leq 1 \quad \text{for } l = 1, \ldots, d_\beta \\
& \sigma_2 \geq 0 \\
& q'\sigma_2 \leq \hat{\beta}' b^* + \hat{\kappa}_n^u \\
& R'\sigma_2 = \hat{\Gamma}' b^* \\
& \zeta_{1,j}, \zeta_{2,j} \in \{0, 1\} \quad \text{for } j = 1, \ldots, d_\theta \\
& w = (b^*)'\hat{\mathbb{G}}_\Gamma \\
& w_j \leq \pi_j \leq w_j + (1 - \zeta_{1,j})\text{BigM}_{1,j} \quad \text{for } j = 1, \ldots, d_\theta \\
& -w_j \leq \pi_j \leq -w_j + \zeta_{1,j}\text{BigM}_{1,j} \quad \text{for } j = 1, \ldots, d_\theta \\
& \pi_j \leq \sigma_1 \leq \pi_j + (1 - \zeta_{2,j})\text{BigM}_{2,j} \quad \text{for } j = 1, \ldots, d_\theta \\
& \textstyle\sum_{j=1}^{d_\theta} \zeta_{2,j} = 1
\end{aligned} \tag{188}$$

The justification of this reformulation is similar to the one discussed for (186). The constraints involving the ζ_1 binary variables ensure that π_j is always the absolute value of w_j, which is constrained (defined) as the jth element of $(b^*)'\hat{\mathbb{G}}_\Gamma$. The additional constraints involving the ζ_2 binary variables then ensure that σ_1 is always the maximum of π_j, since $\sum_{j=1}^{d_\theta} \zeta_{2,j} = 1$ can be satisfied if and only if a single $\zeta_{2,j}$ is equal to 1.

G.5 Some Notes on Computation

We have shown that the main optimization problems in our methodology can all be reformulated as problems with well-understood properties for which there exist globally optimal algorithms. Admittedly, this has taken a substantial amount of work. However, once the theoretical reformulation work has been done once, it does not need to be performed again by a practitioner. The reformulated problems can be implemented directly and solved using appropriate software. We are in the process of developing a software package that processes the necessary optimization problems in the background without requiring additional input from the user.

We summarize the computational steps involved in statistical inference. First, construct consistent estimators of identified population quantities, as discussed in Section G.1. The exact definition of the terms involved here will depend on the null hypothesis of interest. Second, compute the test statistic, T_n, by solving (176). Third, solve the MILP (186) one time each for a large number of bootstrap draws. Given a choice of quantile α_n, this provides a data-driven choice of $\hat{\kappa}_n^u$ through (84), as well as a data-driven choice of $\hat{\kappa}_n^m = \hat{\kappa}_n^u + T_n$. Fourth, solve the MILP (188) one time each for a large number of bootstrap draws. This yields a data-driven choice of $\hat{\kappa}_n^g$. Fifth, with all tuning parameters selected, solve the bilinear maximization problem (184) one time each for a large number of bootstrap draws. This provides the critical value $\hat{c}_{1-\alpha}$ defined in (79). The null hypothesis is then rejected at level α if $T_n > \hat{c}_{1-\alpha}$.

In practice, we have found that the part of this procedure that takes the longest is by far the bilinear program (184). The MILPs used to select the tuning parameters are relatively small. Even though these programs must be solved a large number of times, we have found that they can be solved extremely quickly using modern algorithms like Gurobi (Gurobi Optimization, 2015), which is the software we use for this step in our empirical application.

References

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117. 18

Aliprantis, C. D. and K. C. Border (2006): *Infinite Dimensional Analysis – A Hitchhiker's Guide*, Berlin: Springer-Verlag. 35, 68, 69, 71

Andrews, D. W. and G. Soares (2010): "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, 78, 119–157. 38

Angrist, J. D. and I. Fernandez-Val (2013): "ExtrapoLATE-ing: External Validity and," in *Advances in Economics and Econometrics: Volume 3, Econometrics: Tenth World Congress*, Cambridge University Press, vol. 51, 401–. 6

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, 499–527. 2

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, 431–442. 2

Ashraf, N., J. Berry, and J. Shapiro (2010): "Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia," *American Economic Review*, 100, 2383–2413. 50, 51

Balke, A. and J. Pearl (1997): "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176. 5, 6, 22

Beresteanu, A. and F. Molinari (2008): "Asymptotic properties for a class of partially identified models," *Econometrica*, 76, 763–814. 7

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil (2012): "Treatment effect bounds: An application to SwanGanz catheterization," *Journal of Econometrics*, 168, 223–243. 5

Bierens, H. J. (1990): "A consistent conditional moment test of functional form," *Econometrica: Journal of the Econometric Society*, 1443–1458. 15

Bontemps, C., T. Magnac, and E. Maurin (2012): "Set identified linear models," *Econometrica*, 80, 1129–1155. 7

Borwein, J. and A. S. Lewis (2010): *Convex analysis and nonlinear optimization: theory and examples*, Springer Science & Business Media. 82, 83

Boyd, S. and L. Vandenberghe (2004): *Convex optimization*, Cambridge university press. 81

Brinch, C. N., M. Mogstad, and M. Wiswall (2015): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, forthcoming. 5, 6, 15, 16, 54

Bugni, F., I. Canay, and X. Shi (2015): "Inference for functions of partially identified parameters in moment inequality models," Tech. rep., cemmap working paper, Centre for Microdata Methods and Practice. 6

Byrd, R. H., J. Nocedal, and R. A. Waltz (2006): "KNITRO: An integrated package for nonlinear optimization," in *Large-scale nonlinear optimization*, Springer, 35–59. 43

Canay, I. and A. Shaikh (2016): "Practical and theoretical advances in inference for partially identified models," Tech. rep., cemmap working paper, Centre for Microdata Methods and Practice. 36

Carneiro, P., J. J. Heckman, and E. Vytlacil (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica*, 78, 377–394. 5, 16, 20

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101, 2754–81. 5, 11, 16, 20

Chak, P. M., N. Madras, and B. Smith (2005): "Semi-nonparametric estimation with Bernstein polynomials," *Economics Letters*, 89, 153–156. 78

Chamberlain, G. (2011): "Bayesian aspects of treatment choice," *The Oxford Handbook of Bayesian Econometrics*, 11–39. 6

Chang, I.-S., L.-C. Chien, C. A. Hsiung, C.-C. Wen, and Y.-J. Wu (2007): "Shape restricted regression with random Bernstein polynomials," in *Lecture Notes–Monograph Series*, ed. by R. Liu, W. Strawderman, and C.-H. Zhang, Beachwood, Ohio, USA: Institute of Mathematical Statistics, vol. Volume 54, 187–202. 78

Chen, X. (2007): "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 5549–5632. 38

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection bounds: estimation and inference," *Econometrica*, 81, 667–737. 47

Chernozhukov, V., W. K. Newey, and A. Santos (2015): "Constrained conditional moment restriction models," *arXiv preprint arXiv:1509.06311.* 6, 36, 66

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441. 2

Cohen, J. and P. Dupas (2010): "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment," *The Quarterly Journal of Economics*, 125, 1–45. 50, 51

Dupas, P., H. V. K. M. and A. P. Zwane (2016): "Targeting health subsidies through a non-price mechanism: A randomized controlled trial in Kenya," *Science*, 353, 889–895. 50

Dupas, P. (2014): "ShortRun Subsidies and LongRun Adoption of New Health Products: Evidence From a Field Experiment," *Econometrica*, 82, 197–228. 5, 50, 52

Fang, Z. and A. Santos (2014): "Inference on directionally differentiable functions," *arXiv preprint arXiv:1404.3763.* 39

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76, 1191–1206. 2

Gurobi Optimization, I. (2015): "Gurobi Optimizer Reference Manual," . 25, 87

Heckman, J., J. L. Tobias, and E. Vytlacil (2003): "Simple Estimators for Treatment Parameters in a Latent-Variable Framework," *Review of Economics and Statistics*, 85, 748–755. 6

Heckman, J. J. and R. J. Robb (1985): "Alternative methods for evaluating the impact of interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman and B. Singer, Cambridge University Press. 6

Heckman, J. J. and S. Urzua (2010): "Comparing IV with structural models: What simple IV can and cannot identify," *Journal of Econometrics*, 156, 27–37. 2

Heckman, J. J., S. Urzua, and E. Vytlacil (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88, 389–432. 2

Heckman, J. J. and E. Vytlacil (2001a): "Policy-Relevant Treatment Effects," *The American Economic Review*, 91, 107–111. 3, 5, 20

——— (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738. 2, 3, 5, 9, 11, 12, 16, 19, 21

Heckman, J. J. and E. J. Vytlacil (1999): "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences of the United States of America*, 96, 4730–4734. 2, 3, 9, 16, 19

——— (2001b): "Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effect," in *Econometric Evaluations of Active Labor Market Policies in Europe*, ed. by M. Lechner and F. Pfeiffer, Heidelberg and Berlin: Physica. 3, 5

——— (2001c): "Local Instrumental Variables," in *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, ed. by K. M. C Hsiao and J. Powell, Cambridge University Press. 3, 16

——— (2007a): "Chapter 70 Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 4779–4874. 3

——— (2007b): "Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 4875–5143. 2, 3

Huber, M. and G. Mellace (2014): "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints," *Review of Economics and Statistics*, 97, 398–411. 6

Imbens, G. W. (2010): "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, 48, 399–423. 21

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. 2, 3, 8, 21, 24, 29

Imbens, G. W. and C. F. Manski (2004): "Confidence intervals for partially identified parameters," *Econometrica*, 72, 1845–1857. 4, 30

Imbens, G. W. and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512. 2

Imbens, G. W. and D. B. Rubin (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *The Review of Economic Studies*, 64, 555–574. 6, 22

Kaido, H., F. Molinari, and J. Stoye (2016): "Confidence intervals for projections of partially identified parameters," *arXiv preprint arXiv:1601.00934*. 7

Kaido, H. and A. Santos (2014): "Asymptotically efficient estimation of models defined by convex moment inequalities," *Econometrica*, 82, 387–413. 7

Kirkeboen, L., E. Leuven, and M. Mogstad (2016): "Field of Study, Earnings and Self-Selection," *The Quarterly Journal of Economics*, 131, 1057–1111. 2

Kitagawa, T. (2009): "Identification Region of the Potential Outcome Distributions under Instrument Independence," *Cemmap working paper*. 5

——— (2015): "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063. 6, 22

Koltchinskii, V. I. (1994): "Komlos-Major-Tusnady approximation for the general empirical process and Haar expansions of classes of functions," *Journal of Theoretical Probability*, 7, 73–118. 36

Kowalski, A. (2016): "Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments," *NBER Working paper 22363*. 16

Lee, S. and B. Salanié (2016): "Identifying Effects of Multivalued Treatments," *Working paper*. 2

Luenberger, D. G. (1969): *Optimization by vector space methods*, John Wiley & Sons. 33, 68, 74

Machado, C., A. M. Shaikh, and E. J. Vytlacil (2013): "Instrumental Variables and the Sign of the Average Treatment Effect," *Working paper*. 6

Maestas, N., K. J. Mullen, and A. Strand (2013): "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt," *The American Economic Review*, 103, 1797–1829. 11

Manski, C. (1994): "The selection problem," in *Advances in Econometrics, Sixth World Congress*, vol. 1, 143–70. 5

Manski, C. F. (1989): "Anatomy of the Selection Problem," *The Journal of Human Resources*, 24, 343–360. 5

——— (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80, 319–323. 5

——— (1997): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334. 5, 16

——— (2003): *Partial identification of probability distributions*, Springer. 5

Manski, C. F. and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010. 5, 16

——— (2009): "More on monotone instrumental variables," *Econometrics Journal*, 12, S200–S216. 5

Masten, M. A. (2015): "Random Coefficients on Endogenous Variables in Simultaneous Equations Models," *cemmap working paper 25/15.* 2

Masten, M. A. and A. Torgovitsky (2016): "Identification of Instrumental Variable Correlated Random Coefficients Models," *The Review of Economics and Statistics*, forthcoming. 2

McCormick, G. P. (1976): "Computability of global solutions to factorable nonconvex programs: Part IConvex underestimating problems," *Mathematical programming*, 10, 147–175. 43

McKay Curtis, S. and S. K. Ghosh (2011): "A variable selection approach to monotonic regression with Bernstein polynomials," *Journal of Applied Statistics*, 38, 961–976. 78

Mourifié, I. (2015): "Sharp bounds on treatment effects in a binary triangular system," *Journal of Econometrics*, 187, 74–81. 5

Mourifié, I. and Y. Wan (2016): "Testing Local Average Treatment Effect Assumptions," *The Review of Economics and Statistics*, 99, 305–313. 6

Rio, E. (1994): "Local invariance principles and their application to density estimation," *Probability Theory and Related Fields*, 98, 21–45. 36

Romano, J. P. and A. M. Shaikh (2008): "Inference for identifiable parameters in partially identified econometric models," *Journal of Statistical Planning and Inference*, 138, 2786–2807. 6

——— (2012): "On the uniform asymptotic validity of subsampling and the bootstrap," *The Annals of Statistics*, 40, 2798–2822. 47

Romano, J. P., A. M. Shaikh, and M. Wolf (2014): "A Practical Two-Step Method for Testing Moment Inequalities," *Econometrica*, 82, 1979–2002. 48

Schrijver, A. (1998): *Theory of linear and integer programming*, John Wiley & Sons. 85

Shaikh, A. M. and E. J. Vytlacil (2011): "Partial Identification in Triangular Systems of Equations With Binary Dependent Variables," *Econometrica*, 79, 949–955. 5

Sion, M. (1958): "On general minimax theorems," *Pacific J. Math*, 8, 171–176. 70

Stinchcombe, M. B. and H. White (1998): "Consistent specification testing with nuisance parameters present only under the alternative," *Econometric theory*, 14, 295–325. 15, 34

Tawarmalani, M. and N. V. Sahinidis (2005): "A polyhedral branch-and-cut approach to global optimization," *Mathematical Programming*, 103, 225–249. 43, 84

Torgovitsky, A. (2015): "Identification of Nonseparable Models Using Instruments With Small Support," *Econometrica*, 83, 1185–1197. 2

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341. 3, 8